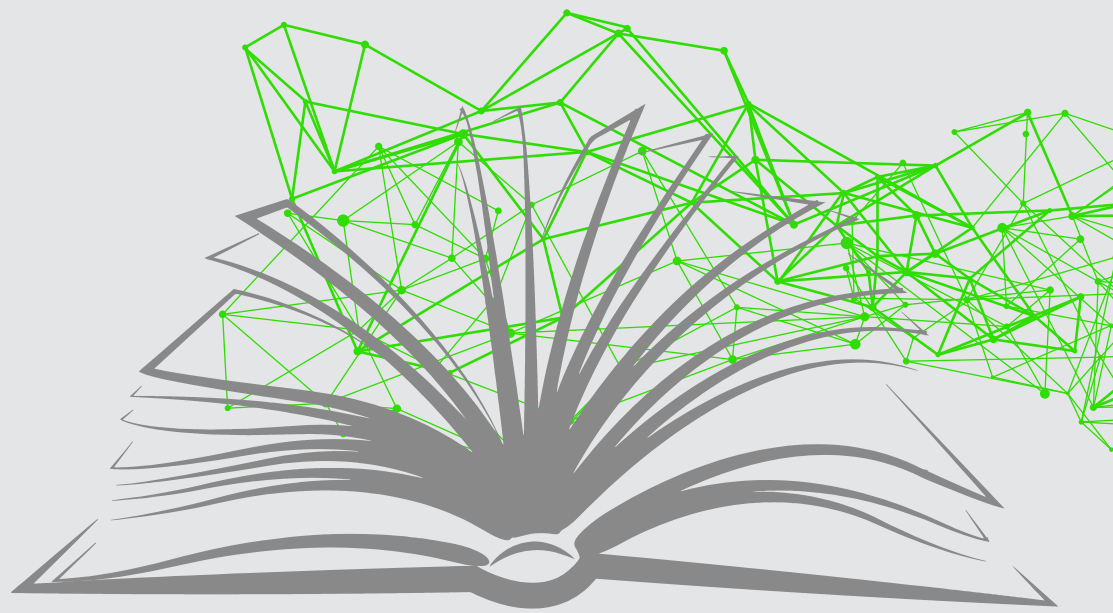


Kulturen des
digitalen
Gedächtnisses

DHd2022 – Potsdam

KONFERENZABSTRACTS



DHd2022 Potsdam
Kulturen des
digitalen
Gedächtnisses
07.–11.03.2022

8. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V.

DHd2022: Kulturen des digitalen Gedächtnisses

Konferenzabstracts

Universität Potsdam & Fachhochschule Potsdam
07. bis 11. März 2022

Partner

MUSEUM BARBERINI
POTSDAM



CnFdI
nationale
Forschungsdaten
Infrastruktur
for CULTURE



Archiv
THEODOR FONTANE

Sponsoren

GERDA HENKEL STIFTUNG

DFG Deutsche
Forschungsgemeinschaft



Die Abstracts wurden von den Autorinnen und Autoren in einem Template erstellt und mittels des von Marco Petris, Universität Hamburg, entwickelten DHConvalidators in eine TEI-konforme XML-Datei konvertiert.

Repo des DHd-Verbands mit den XML-Daten:

<https://github.com/DHd-Verband/DHd-Abstracts-2022>

Herausgeberin: Michaela Geierhos

Redaktion und Korrektur der Auszeichnungen: Sabine Seifert, Kristina Genzel, Peer Trilcke, Anna Busch, Melanie Seltmann

Konvertierung TEI nach PDF: Ingo Börner

https://github.com/ingoboerner/dhd2022_boa

Historie der Autorinnen und Autoren sowie Versionen der Konversionsskripte:

Nina Seemann (2020) – <https://github.com/NinaSeemann/DHd2020-BoA>

Attila Klett (2019) – <https://github.com/texttechnologylab/DHd2019BoA>

Claes Neufeind (2018) – <https://github.com/GVogeler/DHd2018>

Aramís Concepción Durán (2016) – <https://github.com/aramiscd/dhd2016-boa.git>

Karin Dalziel (2013) – <https://github.com/karindalziel/TEI-to-PDF>

Konferenz-Logo: cosmoblond Berlin

Gestaltung des Covers: Stefanie Zeise (ZIM – Zentrum für Informationstechnologie und Medienmanagement, Universität Potsdam)

Potsdam 2022

DOI: [10.5281/zenodo.6304590](https://doi.org/10.5281/zenodo.6304590)

Vorwort

Dieses Jahr steht die 8. Jahrestagung des Verbands „Digital Humanities im deutschsprachigen Raum“ unter dem Motto „Kulturen des digitalen Gedächtnisses“ und wird gemeinsam von der Universität Potsdam und der Fachhochschule Potsdam ausgerichtet. Zahlreiche Tagungsbeiträge greifen die Leitfragen der DHd2022 auf, wovon an dieser Stelle exemplarisch die Folgenden genannt seien: Welche Folgen hat die Digitalisierung für die grundlegenden Kulturpraktiken des Erinnerns und des Vergessens oder des Speicherns und Wiederholens? Und welche digitalen – wissenschaftlichen wie kulturellen – Praktiken entstehen in den neuen digitalen Infrastrukturen des kulturellen Gedächtnisses?

Das Programmkomitee der DHd2022 stand vor der Herausforderung, aus rund 220 Einreichungen eine Auswahl zu treffen, was bei der hohen Qualität der Beiträge nicht einfach war. Ohne die Unterstützung unserer knapp 200 Gutachterinnen und Gutachter, die insgesamt 760 Gutachten erstellt haben, wäre das nicht möglich gewesen. Mit dem nun vorliegenden Tagungsprogramm und diesem *Book of Abstracts* hoffen wir, Ihnen im Jahr 2022 wieder facettenreiche Einblicke in die aktuelle Forschung der digitalen Geisteswissenschaften bieten zu können. Auch in diesem Jahr haben wir wieder Schritte unternommen, um die im *Book of Abstracts* publizierten Versionen der Vorträge zu vollwertigen, wissenschaftlichen Publikationen auszubauen, weshalb sie im Vergleich zur letzten Jahrestagung noch einmal im Umfang auf bis zu 2.500 Wörter gewachsen sind.

Zur wissenschaftlichen Qualitätssicherung wurde der zweistufige Begutachtungsprozess beibehalten, allerdings wurde erstmalig nach dem sogenannten Open-Peer-Review-Verfahren begutachtet, bei dem die Namen der Gutachterinnen und Gutachter den Autorinnen und Autoren offengelegt werden. Neben der fachlichen Zuordnung bei der Begutachtung achtete das Programmkomitee auch darauf, dass das akademische Alter der Einreichenden und Begutachtenden berücksichtigt wurde. Ein weiteres Novum war die Einführung einer Rückmeldephase, in welcher die Gutachten von den Einreichenden kommentiert wurden und im Anschluss die Begutachtenden darauf reagieren konnten. Dieser Schritt sollte den wissenschaftlichen Diskurs in den Begutachtungsprozess zurückbringen.

Darüber hinaus wurden die Arbeiten des Programmkomitees erstmals von der vom Verband eingesetzten Task Force „Optimizing Peer Review“ flankiert, die mit uns die „Handreichung zum Begutachtungsprozess“ entwickelt hat, um den Gutachterinnen und Gutachtern beim Verfassen ihrer Stellungnahmen bessere Orientierung zu geben. Zudem verfolgt diese Task Force den Entwicklungsprozess des Begutachtungsverfahrens der DHd-Jahrestagungen weiter und begleitet ihn durch entsprechende empirische Forschung. Die ersten Ergebnisse werden auf der DHd2022 im Panel „Offen für alle(s)? Open Identities im Reviewprozess der DHd-Konferenz“ präsentiert und diskutiert.

Wissenschaft lebt vom Austausch und von Vernetzung. Deshalb danken wir allen Autorinnen und Autoren, die Beiträge für die DHd2022 eingereicht haben. Sie alle ermöglichen es uns, bei der Jahrestagung die Vielfalt und Qualität der Forschung in den digitalen Geisteswissenschaften aufzuzeigen. Dank gebührt natürlich auch unseren unermüdlichen Gutachterinnen und Gutachtern. Auch wenn das Programmkomitee der DHd2022 schlussendlich das letzte Wort bei der Entscheidung über die Annahme und Ablehnung von Beiträgen hat, braucht es Sie alle als wichtigen Bestandteil der wissenschaftlichen Gemeinschaft, um die Jahrestagung in dieser Form erst möglich zu machen.

Zu den Mitgliedern des Programmkomitees zählen diesmal Anna Busch, Alexander Czmiel, Lisa Dieckmann, Michaela Geierhos (Vorsitz), Evelyn Gius, Katrin Glinka, Andreas Henrich, Andreas Münzmay, Patrick Sahle (stellvertretender Vorsitz), Stefan Schmunk, Peer Trilcke (Vertreter des lokalen Organisationsteams), Lars Wieneke und Heike Zinsmeister. Allen Kolleginnen und Kollegen danke ich ganz herzlich für ihr Engagement. Mein persönlicher Dank gilt insbesondere den Potsdamern: Die Gruppe um Anna Busch, Peer Trilcke und Ulrike Wuttke hat Unglaubliches geleistet, um sowohl die Arbeit des Programmkomitees zu unterstützen als auch die DHd2022 unvergesslich zu machen.

München, im Februar 2022

Michaela Geierhos
für das Programmkomitee der DHd2022

Reviewer:innen der DHd2022

Review Award 2022

Liedtke, Clemens

Nominierte für den Review Award 2022

Blumtritt, Jonathan
Dängeli, Peter
Deicke, Aline
Gasser, Sonja
Gradl, Tobias
Hinzmann, Maria
Krautter, Benjamin
Mathiak, Brigitte
Puppe, Frank
Reiter, Nils
Roller, Ramona
Schmunk, Stefan
Scholger, Walter
Schommer, Christoph
Steyer, Timo
Trilcke, Peer
Wübbena, Thorsten

Reviewer:innen

Acquavella-Rauch, Stefanie
Ahmed, Sajawel
Akkermann, Miriam
Andresen, Melanie
Andrews, Tara
Arnold, Eckhart
Baillot, Anne
Barabucci, Gioele
Barzen, Johanna
Bäumer, Frederik
Bernhart, Toni
Blaschitz, Edith
Bläsi, Christoph
Blumtritt, Jonathan
Börner, Ingo
Brinkmann, Hanna
Brodhun, Maximilian
Bunout, Estelle
Burch, Thomas
Burekhardt, Daniel
Bürgermeister, Martina
Burr, Elisabeth
Busch, Hannah
Busch, Anna
Capelle, Irmlind
Casties, Robert
Clados, Christiane
Cremer, Fabian
Czmiel, Alexander
Dängeli, Peter
Declerck, Thierry
Deicke, Aline
Dieckmann, Lisa
Dogunke, Swantje
Dörk, Marian
Draxler, Christoph
Dröge, Martin
Du, Keli
Dunst, Alexander
Düring, Marten
Eggert, Lisa
Eide, Øyvind
Elwert, Frederik
Ernst, Thomas
Fechner, Martin
Fischer, Frank
Franken, Lina
Freyberg, Linda
Fritze, Christiane
Gasser, Sonja
Geierhos, Michaela
Geiger, Jonathan

Gius, Evelyn
Glawion, Anastasia
Glinka, Katrin
Grabsch, Sascha
Gradl, Tobias
Guhr, Svenja
Gutiérrez De la Torre, Silvia
Eunice
Hahn, Udo
Hall, Mark
Hedeland, Hanna
Heftberger, Adelheid
Hegel, Philipp
Heinisch, Barbara
Helling, Patrick
Henny-Krahmer, Ulrike
Henrich, Andreas
Henzel, Katrin
Hermes, Jürgen
Herrmann, J. Berenike
Hertling, Anke
Heßbrüggen-Walter, Stefan
Heyer, Gerhard
High-Steskal, Nicole
Hinrichs, Erhard
Hinzmann, Maria
Hodel, Tobias
Hoenen, Armin
Hohmann, Georg
Homburg, Timo
Horstmann, Jan
Howanitz, Gernot
Illmayer, Klaus
Jannidis, Fotis
Janz, Nina
Jäschke, Robert
Jeller, Daniel
Jung, Kerstin
Kampkaspar, Dario
Karcher, Stefan
Keck, Jana
Kepper, Johannes
Klaffki, Lisa
Kleymann, Rabea
Klinke, Harald
Koch, Walter
Kocher, Ursula
Konle, Leonard
Krause, Thomas
Krause, Thomas
Krautter, Benjamin
Kröger, Bärbel
Krug, Markus

Kuczera, Andreas	Schmidt, Thomas
Kurz, Stephan	Schmunk, Stefan
Lang, Sarah	Schneider, Stefanie
Langner, Martin	Schöch, Christof
Lassner, David	Scholger, Walter
Leinen, Peter	Scholz, Martin
Liedtke, Clemens	Schommer, Christoph
Liem, Johannes	Schubert, Zoe
Lüdeling, Anke	Schumacher, Mareike
Lüschow, Andreas	Schwandt, Silke
Mandl, Thomas	Seifert, Sabine
Mathiak, Brigitte	Seltmann, Melanie Elisabeth-H.
Matzner, Tobias	Söring, Sibylle
Mayr, Eva	Stadler, Peter
Meier-Vieracker, Simon	Staecker, Thomas
Meister, Jan Christoph	Stange, Jan-Erik
Menzel, Wolfgang	Stede, Manfred
Mertgens, Andreas	Steyer, Timo
Messemer, Heike	Teich, Elke
Meyer, Holger J	Thomas, Christian
Mischke, Dennis	Trilcke, Peer
Molitor, Paul	Trippel, Thorsten
Münzmay, Andreas	Veit, Joachim
Nantke, Julia	Viehhauser, Gabriel
Nerbonne, John	Vogeler, Georg
Neuber, Frederike	Voges, Ramon
Neuefeind, Claes	von Vlahovits, Frederic
Neuroth, Heike	Wagner, Andreas
Nicka, Isabella	Wettlaufer, Jörg
Niebling, Florian	Wieneke, Lars
Niekler, Andreas	Windhager, Florian
Nunn, Christopher	Wörner, Kai
Odebrecht, Carolin	Wübbena, Thorsten
Offert, Fabian	Wuttke, Ulrike
Pado, Sebastian	Zaagsma, Gerben
Pagel, Janis	Zeppezauer-Wachauer,
Pfeffer, Magnus	Katharina
Pichler, Axel	Zirker, Angelika
Pielström, Steffen	
Proisl, Thomas	
Puppe, Frank	
Raspe, Martin	
Rehm, Georg	
Reiners, Stefan	
Reiter, Nils	
Resch, Claudia	
Richts-Matthaei, Kristina	
Rißler-Pipka, Nanette	
Ritter, Jörg	
Roeder, Torsten	
Roller, Ramona	
Rosenthaler, Lukas	
Rüdiger, Jan Oliver	
Ruiz Fabo, Pablo	
Sahle, Patrick	
Scheuermann, Leif	
Schlögl, Matthias	

Inhaltsverzeichnis

Panels

Daten im Raum – Visualisierungen und Physikalisisierungen im Medium Ausstellung	
Bentz, Isabelle; Gfrereis, Heike; Hildenbrandt, Vera; Mayr, Eva; Offenberger, Eva; Tropper, Eva; Windhager, Florian	9
Digitale Archive für Literatur	
Busch, Anna; Fetz, Bernhard; Lepper, Marcel; Wirtz Eybl, Irmgard; Richter, Sandra; Trilcke, Peer	11
Erinnern durch Vernetzen – Digitale Sammlungsforschung	
Alschnner, Stefan; Baumgarten, Marcus; Horstmann, Jan; Müller, Christiane; Nantke, Julia; Weis, Joëlle; Wübbena, Thorsten	14
Kinetik und Methodik – Film als dynamische und multimodale Herausforderung für die DH	
Bateman, John; Diecke, Josephine; Ewerth, Ralph; Heftberger, Adelheid; Howanitz, Gernot; Spiegel, Simon; Loertscher, Miriam	17
Kultur – Daten – Kuratierung – Was speichern wir und wozu?	
Altenhöner, Reinhard; Dieckmann, Lisa; Münzmay, Andreas; Pratschke, Margarete; Primavesi, Patrick; Richts-Matthaei, Kristina; Röwenstrunk, Daniel; Schulz, Christoph; Stellmacher, Martha	19
Offen für alle(s)? – Open Identities im Reviewprozess der DHd-Konferenz	
Burghardt, Manuel; Czmiel, Alexander; Dieckmann, Lisa; Guhr, Svenja; Jacke, Janina; Reiter, Nils; Scholger, Walter; Wuttke, Ulrike	21
Protokolle – Modellierung einer administrativen Textsorte	
Arndt, Nadine; Baddack, Cornelia; Fischer-Nebmaier, Wladimir; Gleixner, Sebastian; von Hindenburg, Barbara; Jüngerkes, Sven; Kurz, Stephan; Schrott, Maximilian	24

Vorträge

AdA Annotation Explorer – Ein Framework für zeitbasierte Linked Open Data-Annotationen zur Analyse audiovisueller Korpora	
Agt-Rickauer, Henning; Scherer, Thomas; Stratil, Jasper	29
Adapting Coreference Algorithms to German Fairy Tales	
Schmidt, David; Krug, Markus; Puppe, Frank	34
Anzeigen als Daten – Dynamisches Tagging und iterative Auswertung eines frühneuzeitlichen Intelligenzblattes	
Serif, Ina; Reimann, Anna; Engel, Alexander	37
Auf den Spuren einer altnordischen Saga-Ästhetik – Poetologische Aussagen in den Erzählerbemerkungen der Isländersagas	
Göggelmann, Michael; Heiniger, Anna Katharina; Reiter, Nils; Zirker, Angelika	40
Aufführungsinformationen in der Mixed Music – Systematische Herausforderungen als Indikatoren musikpraktischer Tendenzen	
Akkermann, Miriam	43
Automatisierte Extraktion und Klassifikation von Variantenschreibungen historischer Berufsbezeichnungen in seriellen Quellen des 16. bis 20. Jahrhunderts	
Moeller, Katrin; Goldberg, Jan Michael	47
Back 'em up – Computerspiele als Objekte kulturellen Erbes	
Schneider, Sophie	50
Best of Both Worlds – Zur Kombination algorithmischer und manueller Verfahren bei der Erschließung großer Handschriftenkorpora	
Nantke, Julia; Bläß, Sandra; Flueh, Marie; Maus, David	54
Data Cleaning als digitale Quellenkritik – VD17 und das Genre der katholischen Dissertation im Alten Reich	
Heßbrüggen-Walter, Stefan	57
Der CLARIAH-DE Tutorial Finder – Eine Suchumgebung für Lehr- und Schulungsmaterialien in den Digital Humanities	
Werthmann, Antonina; Gradl, Tobias	60
Der Einsatz von Computer Vision-Methoden für Filme – Eine Fallanalyse für die Kriminalfilm-Reihe Tatort	
Schmidt, Thomas; Kurek, Sarah	65

Der SSH Open Marketplace – Kontextualisiertes Praxiswissen für die Digital Humanities	
Zarei, Alireza; Seung-Bin, Yim; Đurčo, Matej; Illmayer, Klaus; Barbot, Laure; Fischer, Frank; Gray, Edward	72
Die Aktualität des Unzeitgemäßen –	
Krewet, Michael; Ernst, Felix; Götzmann, Germaine; Hegel, Philipp; Schenk, Torsten; Söring, Sibylle; Tonne, Danah	75
Digitale Kontextualisierung und Visualisierung der Quellen-Trias Bild-Text-Realia zu historischer Kleidung, ihrer Ausformung, Zeichenhaftigkeit und Dreidimensionalität	
de Günther, Sabine; Freyberg, Linda	78
Digital Environmental Humanities – Zum Potential von „Computational and Literary Biodiversity Studies“ (CoLiBiS)	
Langer, Lars; Burghardt, Manuel; Borgards, Roland; Köhring, Esther; Wirth, Christian	82
Dokument, Transkription, Forschungsdatum – Technische und kulturelle Überlegungen für interdisziplinäre Transkriptionspraxis	
Baierer, Konstantin; Boenig, Matthias; Engl, Elisabeth; Geestmann, Mareen; Hinrichsen, Lena; Neudecker, Clemens; Pestov, Paul; Weidling, Michelle	86
Dramatische Metadaten – Die Datenbank deutschsprachiger Einakter 1740–1850	
Çakir, Dilan Canan; Fischer, Frank	89
Emotionen im kulturellen Gedächtnis bewahren	
Dennerlein, Katrin; Schmidt, Thomas; Wolff, Christian	93
Empirische Aufmerksamkeitseffekte multimodaler Kohäsion im Film	
Laubrock, Jochen; Tseng, Chiao-I	98
Erweiterungen der Digital Humanities durch kulturwissenschaftliche Perspektiven	
Franken, Lina	101
Evaluating Hyperparameter Alpha of LDA Topic Modeling	
Du, Keli	104
Evaluation computergestützter Verfahren der Emotionsklassifikation für deutschsprachige Dramen um 1800	
Schmidt, Thomas; Dennerlein, Katrin; Wolff, Christian	107
Executable Papers in den Computational Humanities – Von technischen Herausforderungen und erkenntnistheoretischen Mehrwerten	
Walkowski, Niels-Oliver; Burghardt, Manuel	113
Fluch und Segen der Visualisierung – Unterschiedliche Zielfunktionen im Forschungsprozess der historischen Netzwerkanalyse	
Balck, Sandra; Menzel, Sina; Petras, Vivien; Schnaitter, Hannes; Zinck, Josefine	117
Forschendes Lernen digital	
Bläß, Sandra; Flüh, Marie; Gerstorfer, Dominik; Gius, Evelyn; Meister, Malte; Nantke, Julia; Schumacher, Mareike	120
Gedächtnis digitaler Kulturen und digitaler Geisteswissenschaften – Plädoyer für eine Wissenschaftsgeschichte der DH	
Bernhart, Toni	124
Genitivmetaphern in der Lyrik des Realismus und der frühen Moderne	
Kröncke, Merten; Konle, Leonard; Jannidis, Fotis; Winko, Simone	126
Hackathons als kollektiv-kreative Bildungsereignisse – Ein Konzept zur Gestaltung offener Lehrveranstaltungen in den Digital Humanities	
Mischke, Dennis; Trilcke, Peer; Sluyter-Gäthje, Henny	131
Handwritten Text Recognition und Word Mover’s Distance als Grundlagen der digitalen Edition “Die Kindheit Jesu Konrads von Fußesbrunnen”	
Tomasek, Stefan; Reul, Christian; Wehner, Maximilian	134
Hemisphären des digitalen Gedächtnisses – Analyse von TEI-kodierten Bibelreferenzen mit XQuery im Rahmen der »Bibliothek der Neologie«	
Stallmann, Marco; Sikora, Uwe; Kreß, Hannah; Lemitz, Bastian; Pietsch, Andreas; Wünsch, Lukas	138
iART – Eine Suchmaschine zur Unterstützung von bildorientierten Forschungsprozessen	
Schneider, Stefanie; Springstein, Matthias; Rahnama, Javad; Kohle, Hubertus; Ewerth, Ralph; Hüllermeier, Eyke	142
Inter Annotator Agreement und Intersubjektivität – Ein Vorschlag zur Messbarkeit der Qualität literaturwissenschaftlicher Annotationen	
Gius, Evelyn; Vauth, Michael	147
Japanese Visual Media Graph – Bündelung des Wissens von Fan-Gemeinschaften in einem domänenspezifischen Knowledge Graph	
Pfeffer, Magnus; Kacsuk, Zoltan; Roth, Martin	151
Jung, wild, emotional? – Rollen und Emotionen Jugendlicher in zeitgenössischer Fantasy-Literatur	
Flüh, Marie; Schumacher, Mareike	155
Kategorientheoretische Ontologieentwicklung und Wissensmodellierung für die Digital Humanities	
Gerstorfer, Dominik	160

Lesen, was wirklich wichtig ist – Die Identifikation von Schlüsselstellen durch ein neues Instrument zur Zitatanalyse	
Arnold, Frederik; Fiechter, Benjamin	162
Lieblingsgegenden, Fenster und Mauern – Zur emotionalen Enkodierung von Raum in Deutschschweizer Prosa zwischen 1850 und 1930	
Herrmann, J. Berenike; Grisot, Giulia	166
Literaturgeschichtsschreibung datenbasiert und wikifiziert? – Automatische Extraktion thematischer Statements aus französischen Primärtexten mithilfe von Topic Modeling, RDF und eines kontrollierten Vokabulars in LOD	
Röttgermann, Julia; Klee, Anne; Hinzmann, Maria; Schöch, Christof	170
Mithilfe von Machine Reasoning alchemische Decknamen entschlüsseln	
Lang, Sarah	175
Multimodale KI zur Unterstützung geschichtswissenschaftlicher Quellenkritik – Ein Forschungsaufriß	
Muenster, Sander; Bruschke, Jonas; Kroeber, Cindy; Hoppe, Stephan; Maiwald, Ferdinand; Niebling, Florian; Pattee, Aaron; Utescher, Ronja; Zarriess, Sina	179
Nachhaltige Softwareentwicklung – Von der Inhouse-Lösung zur Open Source-Community am Beispiel von MerMEId	
Henny-Krahmer, Ulrike; Stadler, Peter	183
Nathan nicht ihr Vater? – Wissensvermittlungen im Drama annotieren	
Andresen, Melanie; Krautter, Benjamin; Pagel, Janis; Reiter, Nils	186
Poesie als Fehler – Ein ‘Tool Misuse’-Experiment zur Prozessierung von Lyrik	
Sluyter-Gäthje, Henny; Trilcke, Peer	190
Pragmatisches Forschungsdatenmanagement – Qualitative und quantitative Analyse der Bedarfslandschaft in den Computational Literary Studies	
Helling, Patrick; Jung, Kerstin; Pielström, Steffen	193
Praktiken der digitalen Erinnerung an den 2. Weltkrieg – Netzwerkmodellierungen des „Axis History Forum“	
Glawion, Anastasia	199
Softwarezitation als Technik der Wissenskulturskultur – Vom Umgang mit Forschungssoftware in den Digital Humanities	
Henny-Krahmer, Ulrike; Jettka, Daniel	203
The Digital Archive and the Politics of Digitization	
Zaagsma, Gerben	207
Verwendung von Wissensgraphen zur inhaltlichen Ergänzung kleinerer Textkorpora –	
Hagen, Thora	209
Vom gedruckten Gazetteer zum digitalen Ortsverzeichnis – Das Geschichtliche Ortsverzeichnis (GOV)	
Purschwitz, Anne; Zedlitz, Jesper	212
Von der Wolke zum Pfad – Visuelle und assoziative Exploration zweier kultureller Sammlungen	
Brüggemann, Viktoria; Bludau, Mark-Jan; Pietsch, Christopher; Dörk, Marian	216
Was sehe ich? – Visualisierungsstrategien für Datentransparenz in der Historischen Netzwerkanalyse	
Bludau, Mark-Jan; Halling, Thorsten; Holly, Eva Maria; Wieloch, Jasmin; Schnaitter, Hannes; Balck, Sandra; Plakidis, Melina; Rehm, Georg; Fangerau, Heiner; Dörk, Marian	220
„Wertlose“ Taggings und ihr Nutzen für die Kunstgeschichte	
Poetis, Panoria; Radmacher, Emilia; Smiatek, Katharina; Schneider, Stefanie	225
What's in a name? – Die Rolle der Sprache zur Kultivierung von inklusiven Zugängen zu Kulturerbe	
High-Steskal, Nicole	229
“Wie Wölkchen im Morgenlicht” – Zur automatisierten Metaphern-Erkennung und der Datenbank literarischer Raummetaphern laRa	
Schumacher, Mareike	232

Doctoral Consortium

Adnominal Possession in einem Bibel-Parallelkorpus	
Fleischmann, Florian	238
Das mediale und politische Framing von Extremismusformen im Zeitraum der Jahre 1999 – 2021	
Feldmüller, Tim	239
Digitale Methodenkritik – Die Integration computergestützter Textanalyseverfahren in den Werkzeugkasten der Historiker:innen	
Althage, Melanie	241

Kontextwissen zu historischen Quellen im Semantic Web – Die computergestützte Analyse heraldischer Wand- und Deckenmalereien mit Hilfe von Background Knowledge	
Schneider, Philipp	242
Relating the Unread – Modellierungen der Literaturgeschichte	
Brottrager, Judith	244
Selektion und Nutzer*innen-Position in traditionellen und Internet-Informationsintermediären	
Leyrer, Katharina	245
The Remembered – A Global Study of Literature Dissertations' Bibliography	
Gutiérrez De la Torre, Silvia Eunice	247
Transformation der Geschichtsschreibung? – Der Einsatz von digitalen Medien und Forschungsmethoden in der historischen Praxis und dessen Folgen	
Siebold, Anna	248

Posterpräsentationen

Aktualität und Gedächtnis – Zur korpusanalytischen Untersuchung von Gegenwartsliteratur auf Twitter	
Meier-Vieracker, Simon; Kreuzmair, Elias	251
Analyse der Rezeption von Telenovelas und Serien über lateinamerikanische Geschichte durch Algorithmen	
Meding, Holle Ameriga; Contreras Saiz, Mónica; Muessemann, Hannah	252
Anpassungen von LERA zum Vergleich hebräischer Textzeugen des kabbalistischen Traktats Keter Shem Tov	
Pöckelmann, Marcus; Rebiger, Bill	253
„Arbeitskulturen“ im Wandel – Erfahrungen und Entwicklungen in 20 Jahren DH-Praxis	
Czmiel, Alexander; Neuber, Frederike	256
Aufbau eines Referenzkorpus „Erste Sätze in der deutschsprachigen Literatur“	
Busch, Anna; Roeder, Torsten	259
Berlin's Australian Archive – Eine virtuelle Forschungsumgebung für naturkundliche Sammlungen aus den australischen Kolonien	
Bischoff, Eva; Schwarz, Anja	262
Beyond Budweiser – Creating a Digital Archive of Popular German-American Newspaper Literature	
Keck, Jana; Blessing, Andre	264
Beyond the render silo – Semantically annotating 3D data within an integrated knowledge graph and 3D-rendering toolchain	
Rossenova, Lozana; Schubert, Zoe; Vock, Richard; Blümel, Ina	265
Brücken bauen für Buddha – Das Projekt „Digitalisierung Gandharischer Artefakte“ (DiGA) und die Pelagios Working Group „Linked Data Methodologies in Gandharan Buddhist Art and Texts“	
Elwert, Frederik; Pons, Jessie	267
Building and Improving an OCR Classifier for Republican Chinese Newspaper Text	
Arnold, Matthias; Henke, Konstantin	268
Computational Literary Studies Data Landscape Review	
Börner, Ingo; Charvat, Vera Maria; Đurčo, Matej; Mrugalski, Michał; Odebrecht, Carolin	272
Corpus Nummorum – Eine digitale Forschungsinfrastruktur für antike Münzen	
Köster, Jan; Franke, Claus; Peter, Ulrike	273
Das DFG Schwerpunktprogramm Computational Literary Studies –	
Pielström, Steffen; Jung, Kerstin	274
Das optimale Datenmodell – Eine Spurensuche im Möglichkeitsfeld der Kodierung	
Saric, Sanja; Steiner, Elisabeth; Vogeltanz, Maximilian	275
„Das Puzzle zusammensetzen“ – Von analogen Dokumentensammlungen zu datenbankbasierten Biografien sowjetischer Kriegsgefangener des Zweiten Weltkriegs	
Kindler, Sebastian; Wolf, Katrin	277
Das zoroastrische Mittelpersische – Digitales Corpus und Wörterbuch (MPCD)	
Neuefeind, Claes; Mondaca, Francisco; Eide, Øyvind; Colditz, Iris; Jügel, Thomas; Rezanian, Kianoosh; Zeini, Arash; Cantera, Alberto; Emanuel, Chagai; Shaked, Shaul	278
Datenbiographik im Literaturarchiv – Konzept und Umsetzung digitaler Dienste am Theodor-Fontane-Archiv	
Seifert, Sabine; Busch, Anna; Trilcke, Peer; Genzel, Kristina; Heilmann, Juliane; Möller, Klaus-Peter	280
Datenschutz in der wissenschaftlichen Praxis – Der DARIAH-EU ELDAH Consent Form Wizard	
Scholger, Walter; Hanneschläger, Vanessa; Kamocki, Pawel; Kuzman-Šlogar, Koraljka	282

Der DHd Data Steward – Maßnahmen zur Entwicklung einer nachhaltigen Datenstrategie für die Digital Humanities im deutschsprachigen Raum	
Borges, Rebekka; Debbeler, Anke; Helling, Patrick	283
Der Dienstekatalog der AG Datenzentren – Ein digitales Verzeichnis für Forschungsdatenmanagement-Services in den Geisteswissenschaften	
Rau, Felix; Helling, Patrick	286
DH2go – Lehr- und Lernumgebung für die Digital Humanities	
Heckelen, Malte; Schlesinger, Claus-Michael; Burkhard, Fabienne	288
DiaCollo für GEI-Digital – Ein experimentelles Projekt zur weiteren Erschließung digitalisierter historischer Schulbuchbestände	
Nieländer, Maret; Jurish, Bryan; Scheel, Christian	289
Die Preußische Monarchie visualisieren – Ein Bericht aus dem Werkzeugkasten	
Wierzoch, Jan; Klappenbach, Lou	291
Die Vermessung der (musikalischen) Welt –	
Rettinghaus, Klaus	293
Digitale Texte vom Religionsfrieden bis hin zum Liebesbrief – Das Zentrum für digitale Editionen in Darmstadt stellt sich vor	
Kalmer, Silke; Kampkaspar, Dario; Müller, Sophie; Seltmann, Melanie E.-H.; Stegmeier, Jörn; Wunsch, Kevin	294
Digitalisierte Ego-Dokumente als Quellen für die historische Forschung	
Adamczak, Katarzyna; Štanzel, Arnošt	296
Doing (Digital) History – Kollaborative Formen der Erforschung von Geschichte in sozialen Medien im Projekt #SocialMediaHistory	
Berg, Mia; Lorenz, Andrea	297
Ein Thesaurus für die digitale Edition der Ästhetikvorlesungen von Friedrich Schleiermacher	
Kelm, Holden; Klappenbach, Lou	298
Entdeckung der Korrespondenz Alexander von Humboldts durch Such- und Visualisierungsfunktionen	
Lecroq, Axelle	300
FDM-Awareness in Zeiten von Corona – Sammelkarten zum Forschungsdatenmanagement „Daten & Datteln digital“	
Mollenhauer, Elisabeth; Rau, Felix	303
Fidus Writer als Alternative zum DH ConValidator? – Ein Prototyp	
Gebhard, Henning	306
GitMA-Poster – CATMA-Daten via Git abrufen und mittels Python-Bibliothek weiterverarbeiten	
Meister, Malte; Vauth, Michael; Gerstorfer, Dominik	307
Grenzüberschreitendes Textmining von Historischen Zeitungen – Das impresso-Projekt zwischen Text- und Bildverarbeitung, Design und Geschichtswissenschaft	
Ehrmann, Maud; Bunout, Estelle; Clematide, Simon; Düring, Marten; Fickers, Andreas; Guido, Daniele; Kalyakin, Roman; Kaplan, Frederic; Romanello, Matteo; Schroeder, Paul; Ströbel, Philip; van Beek, Thijs; Volk, Martin; Wieneke, Lars	308
Informationstechnologische Gedächtnisarbeit in der Rezensionszeitschrift RIDE	
Henny-Krahmer, Ulrike; Neuber, Frederike; Scholger, Martina	313
„Ja, jetzt ist das langweilig. Aber in zwanzig Jahren!“ – Bereitstellung, Zugang und Analyse literarischer Blogs am Beispiel des Techniktagebuchs	
Blessing, André; Hess, Jan; Jung, Kerstin	314
Kontrastive Textanalyse mit pydistinto – Ein Python-Paket zur Nutzung unterschiedlicher Distinktivitätsmaße	
Du, Keli; Dudar, Julia; Rok, Cora; Schöch, Christof	316
Leben, Werke und Datensilos – Zur Verknüpfung und Visualisierung von im/materiellen Komponenten des kulturellen Erbes	
Mayr, Eva; Liem, Johannes; High-Steskal, Nicole; Grebe, Anja; Windhager, Florian	317
Linked Open Data für die Literaturgeschichtsschreibung – Das Projekt "Mining and Modeling Text"	
Hinzmann, Maria; Schöch, Christof; Dietz, Katharina; Klee, Anne; Erler-Fridgen, Katharina; Röttgermann, Julia; Steffes, Moritz	320
Linked Open Tafsir – Rekonstruktion der Entstehungsdynamik(en) des Korans mithilfe der Netzwerkmodellierung früher islamischer Überlieferungen	
Ahmed, Sajawel; Rehman, Misbahur; Tischlik, Joshua; Kruse, Carl; Mahmutovic, Edin; Özsoy, Ömer	323
Mediatheken der Darstellenden Kunst digital vernetzen	
Illmayer, Klaus; Tiefenbacher, Sara; Voß, Franziska; Beck, Julia; Henninger, Christine; Wittenbecher, Maxim	325
Mein liebster Schatz! – Das Citizen Science-Projekt Gruß & Kuss stellt sich vor	
Rapp, Andrea; Büdenbender, Stefan; Dietz, Nadine; Dunkelmann, Lena; Gnau-Franké, Birte; Liesenfeld, Nina; Schmunk, Stefan; Seltmann, Melanie E.-H.; Stäcker, Thomas; Werner, Stephanie; Wyss, Eva L.	326

Möglichkeiten und Grenzen eines digitalen barocken Gedächtnisses – Ein DFG-Projekt in der Rückschau	
Müller, Melissa	327
MUSE4Anything – Ontologiebasierte Generierung von Werkzeugen zur strukturierten Erfassung von Daten	
Bühler, Fabian; Barzen, Johanna; Leymann, Frank; Standl, Bernhard; Schlomske-Bodenstein, Nadine	329
Muster von “you” und “thou” – Modellierung der Anrede im englischen Sonett	
Rath, Brigitte	331
NERDPool – Datenpool für Named Entity Recognition	
Andorfer, Peter; Schlögl, Matthias; Bleier, Roman	333
Ontological modelling of the Greek Intangible Cultural Heritage for complex geo-semantic querying	
Baglatzi, Alkyoni; Velissaropoulos, Georgios	334
Projektpräsentation "Early Medieval Glosses And The Question Of Their Genesis – A Case Study On The Vienna Bede" (Gloss-ViBe)	
Bauer, Bernhard	335
Prosopographische Interoperabilität (IPIF) – Stand der Entwicklungen	
Vogeler, Georg; Hadden, Richard; Schlögl, Matthias; Vasold, Gunter	337
Referenzierung des digitalen kulturellen (Text-)Erbes – Digitale Quellenkritik und Modellierung von Metadaten	
Althage, Melanie; Dreyer, Malte; Guescini, Rolf; Hiltmann, Torsten; Lüdeling, Anke; Odebrecht, Carolin	338
Reflexive Passagen und ihre Attribution	
Varachkina, Hanna; Barth, Florian; Gödeke, Luisa; Hofmann, Anna Mareike; Dönicke, Tillmann	340
Repositorien als digitale Gedächtnisträger zwischen Evolution und Langzeitplanung	
Steiner, Elisabeth; Vasold, Gunter; Scholger, Martina	341
Semantische Suche mit Word Embeddings für ein mehrsprachiges Wörterbuchportal	
Tu, Ngoc Duyen Tanja; Meyer, Peter	342
Semi-automatische Erschließung von Rechnungsbüchern am Beispiel des Stadtarchivs Leuven	
Bigalke, Jan; Drach, Sviatoslav; Neuefeind, Claes	344
Spotlights – Wie das OpenMethods-Metablog Digital Humanities-Methoden, -Tools und -Toolmaker ins Scheinwerferlicht rückt	
Wutke, Ulrike; Tóth-Czifra, Erzsébet; Testori, Marinella; Horvath, Aliz; Spence, Paul; Katsiadakis, Helen	345
Strukturen und Impulse zur Weiterentwicklung der DHd-Abstracts	
Busch, Anna; Cremer, Fabian; Lordick, Harald; Mischke, Dennis; Steyer, Timo	347
Studying the ephemeral, cultures of digital oblivion – Identifying patterns in Instagram Stories.	
Achmann, Michael; Hampel, Lisa; Asabidi, Ruslan; Wolff, Christian	349
Szenario-basierte Planung eines semantischen Digitalisierungsvorhabens in der digitalen Geschichtswissenschaft	
Scheltjens, Werner; Schlieder, Christoph	350
Text-Bild-Gefüge – Digital Humanities und der Diskurs der Moderne	
Klemstein, Franziska	351
The Digitized minutes of the Habsburg Governments, 1848-1918	
Fischer-Nebmaier, Wladimir; Kurz, Stephan; Lein, Richard	353
Towards a Computational Study of German Book Reviews – A Comparison between Emotion Dictionaries and Transfer Learning in Sentiment Analysis	
Rebora, Simone; Messerli, Thomas; Herrmann, J. Berenike	354
Training the Archive – Von der maschinellen Exploration musealer Sammlungsdaten zur Curator's Machine	
Bönisch, Dominik	356
Volltexterkennung für historische Sammlungen mit OCR4all-libraries iterativ und partizipativ gestalten	
Klaes, Jan Sebastian; Korwisi, Kristof; Krüger, Katharina; Reul, Christian; Towara, Nadine	358
Webservice correspSearch – subVersion 2	
Dumont, Stefan; Grabsch, Sascha; Müller-Laackman, Jonas	359
What was Theoretical Biology? – A Topic-Modelling Analysis of a Multilingual Corpus of Monographs and Journals, 1914-1945	
Böhm, Alexander; Reiners-Selbach, Stefan; Baedke, Jan; Fábregas Tejeda, Alejandro; Nicholson, Daniel J.	360
Who CAREs, really? – Vom schwierigen Umgang mit digitalisierten Kulturgütern aus kolonialen Kontexten	
Lange, Felix; Kuper, Heinz-Günter; Müller, Anja; Amrhein, Kilian; Klindt, Marco; Nowicki, Anna-Lena	362
Zeitgeschichte untersuchen – Topic Modeling von #blackouttuesday-Inhalten auf Instagram	
Knierim, Aenne; Achmann, Michael; Wolff, Christian	363

Workshops

Annotorious – Eine JavaScript-Bibliothek für die Entwicklung maßgeschneiderter Bildannotationstools Rainer, Simon; Radisch, Erik	367
Barcamp "Headlines & Highlights" der AG Zeitungen & Zeitschriften Rißler-Pipka, Nanette; Roeder, Torsten	369
Die Sprache der Erinnerung – analysieren und verstehen – Korpuslinguistische Zugänge zu Oral-History-Daten Gerstenberg, Annette; Leh, Almut; Möbus, Dennis; Pagenstecher, Cord	371
Einführung in DraCor – Programmable Corpora für die digitale Dramenanalyse Börner, Ingo; Fischer, Frank; Milling, Carsten; Sluyter-Gäthje, Henny	373
Ethisch - transparent - offen – Die CARE-Prinzipien und ihre Implikationen für geisteswissenschaftliche FDM-Services Moeller, Katrin; Söring, Sibylle; Imeri, Sabine; Lemaire, Marina; Reichert, Nils	376
FAIRes Datenmanagement mit dem DARIAH-DE Repository Jander, Melina; Weimer, Lukas	378
Flexibles Arbeiten mit OCR4all – Massenvolltextdigitalisierung von Drucken mithilfe von OCR-D und hochqualitative Transkription von Handschriften Langhanki, Florian; Wehner, Maximilian; Baierer, Konstantin; Hinrichsen, Lena; Reul, Christian	381
GitMA oder CATMA für Fortgeschrittene – Projektdaten via Git abrufen und mittels Python-Bibliothek weiterverarbeiten Schumacher, Mareike; Vauth, Michael; Gerstorfer, Dominik; Meister, Malte	384
HISB vorgestellt: Eine virtuelle Arbeitsumgebung für die akademische Forschung wie auch die Digitalisierung von strukturierten Informationen aus Archivalien – Eine Anwendung der virtuellen Forschungsplattform Geovistory Knecht, David; Beretta, Francesco; Hotz, Gerhard	387
Introduction to Docker Lampert, Marcus	389
Manifest für digitale Editionen Fritze, Christiane	391
Optimiertes Peer Reviewing in den Digital Humanities Guhr, Svenja; Steyer, Timo; Scholger, Walter; Burghardt, Manuel; Dieckmann, Lisa; Reiter, Nils; Wuttke, Ulrike	393
Parser bauen für domänenspezifische Notationen Arnold, Eckhart	396
Peer-to-Peer-Workshop zum Projekt Management in den Digital Humanities Cremer, Fabian; Dogunke, Swantje; Neubert, Anna; Wübbena, Thorsten	397
Projektmanagement für die Digital Humanities Frank, Markus	400
Repräsentativität in digitalen Archiven Dziudzia, Corinna; Hall, Mark	402
Textexplorationen in der digitalen Literaturwissenschaft – Eine kritische und angewandte Auseinandersetzung mit Repräsentations- und Interpretationsansätzen von Text Brandl, Stephanie; Lassner, David; Krömer, Cora; Baillot, Anne	405
Vom Begriff über das Phänomen zur Analyse – Ein CRETA-Workshop zur Operationalisierung in den DH Andresen, Melanie; Krautter, Benjamin; Pagel, Janis; Pichler, Axel	408
Wahrnehmungsstrukturen und User Experience des digitalen Kulturerbes – Ein Blick auf museale Online Sammlungen Kienbaum, Janna; Kreiseler, Sarah; Heidmann, Frank	411

Anhang

Index der Autorinnen und Autoren	415
--	-----

Panels

Daten im Raum

Visualisierungen und

Physikalisierungen im Medium

Ausstellung

Bentz, Isabelle

isabelle.bentz@hslu.ch
Hochschule Luzern

Gfrereis, Heike

heike.gfrereis@dla-marbach.de
Deutsches Literaturarchiv Marbach

Hildenbrandt, Vera

vera.hildenbrandt@dla-marbach.de
Deutsches Literaturarchiv Marbach

Mayr, Eva

eva.mayr@donau-uni.ac.at
Universität für Weiterbildung, Krems

Offenberg, Eva

eva.offenberg@artcom.de
ART+COM AG

Tropper, Eva

eva.tropper@museum-joanneum.at
Museumsakademie Joanneum, Graz

Windhager, Florian

florian.windhager@donau-uni.ac.at
Universität für Weiterbildung, Krems

Hintergrund

Visuelle Repräsentationen von abstrakten Daten und komplexen Themen spielen im Museums- und Ausstellungsraum seit langem eine wichtige Rolle (Vossoughian, 2003) und ihre GestalterInnen gewannen im Zuge der digitalen Transformation substantielle neue Möglichkeiten an die Hand. Ob es um komplexe gesellschaftliche, politische, historische oder geisteswissenschaftliche Themen geht (Caraban et al., 2018; Ehmelt et al., 2021; Latour & Weibel 2020; Reiner & Patkowski, 2017) oder um distanzierte Blicke auf Sammlungen und Ausstellungen selbst (Windhager et al., 2018) – Bildstatistiken, Zeitstrahlen, Karten, Graphen und andere Formate ermöglichen die Darstellung komplexer Sachverhalte, die über Objekte allein nicht erzählt werden können. Konkrete Visualisierungsformate in diesem Feld reichen vom punktuellen Einsatz von Informationsgrafiken über interaktive Touchscreens (Hinrichs et al., 2008) und mobile Applikationen (Rogers et al., 2014) bis hin zu raumgreifenden Datenphysikalisierungen (Alexander et al., 2019; Dragicevic et al., 2021) und Ausstellungen, die

sich gelegentlich von Originalobjekten ganz verabschieden, um Inhalte nur mehr über Informationsdesign zu erzählen.¹

Die durch COVID-19 erzwungenen Schließungen physischer Ausstellungsräume beschleunigten zwar vielerorts die Verschiebung von musealen Inhalten in digitale Schauräume (Hoffman, 2020; Markopoulos et al., 2021), machten aber auch die Unersetzbarkeit von Originalobjekten und die spezifische Räumlichkeit von Ausstellungserfahrungen für viele umso deutlicher spürbar (Amorim, J. P., & Teixeira, 2021). Zudem können räumlich situierte Visualisierungen und Physikalisierungen in Ausstellungsräumen in direkten kontextuellen Relationen zu räumlichen Objekten des “close readings” kinetisch und körperlich erfahren werden, die über web-basierte Darbietungen weit hinausgehen (Rogers et al., 2014; Alexander et al., 2019). Aus Digitalisierungs- und Visualisierungsperspektive resultiert daraus ein Bedarf nach einer intensivierten Diskussion von Visualisierungen und Physikalisierungen in hybriden und realräumlichen Formaten. Während sich die methodologische und theoretische Reflexion von web-basierten Visualisierungen im Kontext von kulturellen Sammlungen bereits konsolidiert (Dörk & Glinka, 2018; Windhager et al., 2018), so steht die Reflexion von Visualisierungen und Physikalisierungen im Ausstellungsraum bislang noch aus.²

Leitfragen

Vor diesem Hintergrund will das in Kooperation mit der Museumsakademie Joanneum konzipierte Panel zu “Daten im Raum” hybride Praktiken von digital-materialen Vermittlungs- und Gedächtnistechnologien sondieren und relevante Positionen der Visualisierung und Physikalisierung im Ausstellungsraum mit Blick auf vier verknüpfte Themencluster kartieren.

- **Visualisierung & Ausstellungsdesign:** Was sind etablierte Strategien und aktuelle Trends der visuellen Repräsentation im Rahmen der Ausstellungsszenografie? Wie lässt sich der Problematik der quasi-objektiven Autorisierung von Sachverhalten (Stichwort “Erzeugung von Evidenz”) begegnen? Gibt es Modelle, die hier eine Ebene kritischer Reflexion einführen? Welche Rolle spielen Visualisierungen und Physikalisierungen im Kontext barrierefreier Informationsaufbereitung? Wie können auch kleine Museen mit wenig Budget zu ansprechenden Datenvisualisierungen kommen?
- **Kuratorische Praxis:** Inwiefern verändern sich kuratorische Zugänge unter dem Eindruck einer neuen Verfügbarkeit von Daten? Welche neuen Formen der Zusammenarbeit zwischen DH-Forschung, Informationsdesign und kuratorischer Praxis sind möglich? Welche Relationen der Kontextuierung oder Hierarchisierung lassen sich zwischen Objekten und Visualisierungen im Ausstellungsraum denken? Welche Ergänzungs- oder Konkurrenz-Beziehungen gibt es zwischen physischen und digitalen Ausstellungsräumen?
- **Besucher*innen:** Wie viel visuelle Diversität und Komplexität im Ausstellungsraum ist angesichts allgemeiner “Visual Literacy” möglich? Was sind Strategien der Validierung von Physikalisierungen und Visualisierungen im Ausstellungsraum? Wie sehr können bestehende Methoden der Evaluation für diesen Zweck übernommen werden, wo gibt es Bedarf für Weiterentwicklungen, z.B. mit Blick auf multiple “casual user”-Kategorien?
- **Digitale Geisteswissenschaften:** Welche Themen der digitalen Geisteswissenschaften eignen sich für visuelle und

physische Remediatisierung? Welche neuen Formen der Zusammenarbeit zwischen DH-Forschung, Informationsdesign und kuratorischer Praxis sind möglich? Wie können relevante Communities of Practice über DH-nahe Infrastrukturen besser vernetzt und der Diskurs zu "Daten im Raum" weiter entfaltet und fortgesetzt werden?

Impulsvorträge

Daten im Raum – physische und virtuelle Wissensorte für Austausch, Diskussion & Handlung

Isabelle Bentz (Data Design & Art BA, Hochschule Luzern)

Heute entscheiden wir über unsere Zukunft und die Art und Weise, wie wir als Individuum, als Gesellschaft und als Teil unserer Umwelt zusammen weiterleben werden. Daten sind dabei eine wichtige Grundlage von Wissen. Ihren Wert entfalten sie aber erst dann, wenn sie sorgfältig interpretiert, verständlich visualisiert, transparent vermittelt und kritisch diskutiert werden. Doch wie kann komplexe Information so dargestellt werden, dass sie nicht nur gelesen, sondern auch mit allen Sinnen erlebt, verstanden und erinnert werden kann? Wie können informierte Orte konzipiert werden, die durch Interaktion und Entdeckung die Wissensvermittlung so anregen, dass sie Menschen zum Anhalten, zum Austausch und zur Diskussion antreiben? Der Impulsvortrag gibt Einblicke in die Entwicklung, Vermittlung und Rezeption von Dateninszenierungen im physischen und digitalen Raum anhand von Beispielen aus der eigenen Praxis und Arbeiten von Studierenden.

Wie Literatur durch Daten im Raum sichtbar werden kann

Heike Gfrereis / Vera Hildenbrandt (DLA Marbach)

Beim Nachdenken über Literaturausstellungen scheint ein Satz unausweichlich: Literatur könne man nicht ausstellen, da sie – anders etwa als Werke der bildenden Kunst – nicht fürs Zeigen gemacht sei. Literarische Texte entzögen sich durch ihren Umfang, ihre Gebundenheit an das Medium Buch und ihre sprachliche Disposition, die der Erfindung und Imagination zum Beispiel einer Geschichte diene, dem Ausstellen. Ihr Ziel sei eine bestimmte literarische Erfahrung: das identifikatorische Lesen von der ersten bis zur letzten Seite, der ‚Flow‘. Diese Begründung geht vom Roman als literarische Leitgattung aus, ignoriert aber viele andere Erscheinungsformen von Literatur, ebenso wie andere Erfahrungsmöglichkeiten von Literatur. In unserem Beitrag möchten wir am Beispiel von zwei Ausstellungen („Hölderlin, Celan und die Sprachen der Poesie“ im Literaturmuseum der Moderne 2020/21 sowie der neuen, für Anfang 2023 geplanten Dauerausstellung im Schiller-Nationalmuseum) skizzieren, welche Möglichkeiten Close und Distant Reading, empirische Leserforschung und das Visualisieren von Daten im Raum dem Ausstellen von Literatur eröffnen.

Daten im physischen Raum – Vermittlungsformen

Eva Offenberg (ART+COM)

Der Kontext Raum (z. B. in Form einer Ausstellung) bietet Darstellungsmöglichkeiten für Daten, die über die Möglichkeiten von klassischen Print- oder Digitalmedien weit hinausreichen. So können bei der Vermittlung zusätzliche physische Größendimensionen, Licht, haptische Interaktionen oder auch alle Formen von großformatigen digitalen Medienflächen zum Einsatz kommen. Mit diesen Mitteln können zum einen Inhalte und Zusammenhänge auf unerwartete Weise betrachtet und neuartig vermittelt werden. Zum anderen können emotionale und ikonische Momente für den Betrachter*innen geschaffen werden, die einen hohen Erinnerungswert an die Daten und Inhalte haben. Aus der Erfahrung von 15 Jahren Ausstellungs- und Exponatentwicklung in verschiedensten Themenfeldern werden Beispiele solcher Darstellungsmöglichkeiten aufgezeigt, und deren Möglichkeiten, Potentiale aber auch Limitierungen bzw. Grenzen betrachtet. Die Ergebnisse erfordern oft einen engen Dialog zwischen Kurator*innen, Gestalter*innen und Programmierer*innen, um sowohl die Aussage und Lesbarkeit der Daten, aber auch Machbarkeit, bezogen auf Faktoren wie Haltbarkeit oder Realisierungsbudgets, zu einem überzeugenden Gesamtergebnis zu vereinen.

Zur Visualisierung und Physikalisierung kultureller Objekt- & Biographiedaten

Eva Mayr (Universität für Weiterbildung, Krems)

Das H2020-Projekt "In/Tangible European Heritage - Visual Analysis, Curation & Communication" (<https://intavia.eu>) bringt erstmalig die Datenbestände von europäischen Objekt- und Biographiedatenbanken zusammen, um neue Verknüpfungen und Kontextualisierungen zu analysieren, zu kuratieren und visuell zu repräsentieren. Hierbei spielt auch die dreidimensionale Visualisierung von biographischen Trajektorien in geographischer und diagrammatischer Raumzeit eine zentrale Rolle. Der Impulsvortrag geht in diesem Kontext auf diverse Optionen der kontextuellen Visualisierung und Physikalisierung ein, reflektiert Möglichkeiten der Evaluation (inkl. Stärken und Schwächen für diverse Zielgruppen), und verortet die entsprechenden Überlegungen im größeren Feld der Forschung zur Visualisierung kultureller Sammlungen (<http://collectionVis.org>).

Teilnehmende

Isabelle Bentz ist Leiterin des Bachelorstudiengangs Data Design & Art an der Hochschule Luzern. Nach ihrem Architekturdiplom an der ETH Zürich entwickelte sie als Architektin mehrfach ausgezeichnete, interaktive Dateninstallationen und interaktive Rechercheplattformen, die Mensch und Information im Raum zusammenbringen.

Heike Gfrereis studierte Germanistik und Kunstgeschichte in Stuttgart, Marburg und Tübingen und promovierte mit einer Arbeit über Heinrich von Kleist und literarische Systemimmanenz. Seit 2001 ist sie Leiterin der Museumsabteilung im Deutschen Literaturarchiv Marbach. Sie hält eine Honorarprofessur an der Universität Stuttgart und arbeitete als freie Kuratorin in München, Stuttgart und Berlin.

Vera Hildenbrandt studierte Germanistik und Französischen Philologie und promovierte zum Thema Europa in Alfred Döblins Amazonas-Trilogie. Sie arbeitete als wissenschaftliche Mitarbeiterin am Trier Center for Digital Humanities und war dessen Geschäftsführerin. Seit einer DH-Vertretungsprofessur an der

Universität Trier ist sie als wissenschaftliche Mitarbeiterin am Deutschen Literaturarchiv Marbach tätig.

Eva Offenberg ist Gestalterin mit Schwerpunkt Medien und Kommunikationsdesign. Nach einem Studium an der HfG Schwäbisch Gmünd, und Gastsemestern an der UdK Berlin und der ESDI in Rio de Janeiro, entwickelt sie seit 2008 bei ART+COM mehrfach ausgezeichnete mediale Exponate und Ausstellungen an der Schnittstelle von visueller Gestaltung, Technologie und Architektur.

Eva Mayr ist wissenschaftliche Mitarbeiterin am Department für Kunst- und Kulturwissenschaften der Universität für Weiterbildung, Krems. Nach einem Psychologiestudium beschäftigte sich Ihre Dissertation mit der Rolle neuer Medien für informelles Lernen im Museum. Ihre aktuelle Forschung fokussiert auf Visualisierung kultureller Sammlungen, kognitive Prozesse und informelles Lernen.

Eva Tropper (Moderation) ist Teil des Leitungsteams der Museumsakademie Joanneum. Sie hat Geschichte und Medienkultur studiert und war in verschiedenen Projekten mit Schwerpunkt Visuelle Kultur an der Schnittstelle von Forschung und Museum tätig. Sie kuratierte für das GrazMuseum, das Photoinstitut Bonartes und das Pavelhaus und lehrte an den Universitäten Graz, Krems und Klagenfurt.

Florian Windhager (Moderation) ist wissenschaftlicher Mitarbeiter an der Universität für Weiterbildung Krems und lehrt an den Universitäten Passau und Wien. Nach einem Studium der Philosophie promovierte er zum Thema der synoptischen Visualisierung von Objekt- und Biographiedaten. Forschungsschwerpunkte im Feld von Visualisierung, digitale Geisteswissenschaften und Sammlungen des kulturellen Erbes.

Fußnoten

1. Siehe z.B. die Ausstellung "Fontane 200" (Neuruppin, 2019) oder "Level Green" (Autostadt Wolfsburg, 2009).
2. Einen Anstoß der Debatte stellte der Workshop "Daten im Raum. Visualisierungen als Formen des Argumentierens in Ausstellungen" der Museumsakademie Joanneum (2019, online) dar.

Bibliographie

Abel, G. J., & Sander, N. (2014). "Quantifying global international migration flows", in: *Science*, 343 (6178), 1520-1522.

Alexander, J., Isenberg, P., Jansen, Y., Rogowitz, B., & Moere, A. V. (2019). "Data Physicalization", in: *Dagstuhl Reports*, 8 (10), 127-147.

Amorim, J. P., & Teixeira, L. M. L. (2021). "Art in the Digital during and after Covid: Aura and Apparatus of Online Exhibitions", in: *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 12 (5), 1-8.

Caraban, A., Paulino, T., Pereira, R., Spence, R., & Campos, P. (2018, July). "The monarch room: an interactive system for visualization of global migration data", in: *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32* (pp. 1-5).

Dörk, M., & Glinka, K. (2018). "Der Sammlung gerecht werden: Kritisch-generative Methoden zur Konzeption experimenteller Visualisierungen", in: *Book of Abstracts, DHd 2018, Köln*.

Dragicevic, P., Jansen, Y., & Moere, A. V. (2021). *Data Physicalization*. Jean Vanderdonckt (Ed.). Springer Handbook of Human Computer Interaction, New York: Springer.

Ehmel, F., Brüggemann, V., & Dörk, M. (2021). "Topography of Violence: Considerations for Ethical and Collaborative Visualization Design", in: *Computer Graphics Forum* (Vol. 40, No. 3, pp. 13-24).

Hoffman, S. K. (2020). "Online Exhibitions during the COVID-19 Pandemic", in: *Museum Worlds*, 8 (1), 210-215.

Latour, B., & Weibel, P. (Eds.) (2020). *Critical zones: observatories for earthly politics*. MIT Press.

Markopoulos, E., Ye, C., Markopoulos, P., & Luimula, M. (2021). "Digital Museum Transformation Strategy Against the Covid-19 Pandemic Crisis", in: *International Conference on Applied Human Factors and Ergonomics* (pp. 225-234). Springer, Cham.

Reiner, N. E., & Patkowski, J. *Inventing Abstraction, Re-inventing Our Selves: The Museum of Modern Art's Artist Network Diagram and the Culture of Capitalism*.

Rogers, K., Hinrichs, U., & Quigley, A. (2014). *It doesn't compare to being there: In-situ vs. remote exploration of museum collections*.

Vossoughian, N. (2003). "The Language of the World Museum: Otto Neurath, Paul Otlet, Le Corbusier", in: *Transnational Associations*, 1 (2), 82-93.

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E. (2018). "Visualization of cultural heritage collection data: State of the art and future challenges", in: *IEEE transactions on visualization and computer graphics*, 25 (6), 2311-2330.

Digitale Archive für Literatur

Busch, Anna

annabus@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Fetz, Bernhard

bernhard.fetz@onb.ac.at
Literaturarchiv und Literaturmuseum Österreichische Nationalbibliothek

Lepper, Marcel

Marcel.Lepper@klassik-stiftung.de
Goethe- und Schiller-Archiv Weimar

Wirtz Eybl, Irmgard

Irmgard.Wirtz@nb.admin.ch
Schweizerisches Literaturarchiv

Richter, Sandra

Sandra.Richter@dla-marbach.de
Deutsches Literaturarchiv Marbach

Trilcke, Peer

trilcke@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Thematischer Rahmen

Die Reflexion der institutionellen Transformation von Literaturarchiven angesichts der Digitalisierung überschreitet die Einzelinstitutionen notwendig, gerade dort, wo es gemeinsame Praktiken, Routinen, Standards und Infrastrukturen zu entwickeln gilt. Das Panel greift diesen Bedarf durch seinen internationalen und interinstitutionellen Ansatz auf.

Literaturarchive stehen durch die Digitalisierung vor einer Vielzahl an Herausforderungen: Begriff, Praxis und Materialität des Literaturarchivs befinden sich in einem Transformationsprozess, den die Institution 'Literaturarchiv' in dieser Grundsätzlichkeit seit ihrer konzeptionellen Erfindung im 19. Jahrhundert (Goethe 1823, Dilthey 1970 [1889], Thaller 2011) nicht durchlaufen hat. Die Anforderungen nach Partizipation (Theimer 2018), die Digitalisierung der Bestände und die Umsetzung von Open Access- und Data-Strategien (Szekely 2017), die Adressierung der Fragen, die born-digitals mit sich bringen, gehen in vielen Fällen einher mit einem Umbau von Routinen und Handlungsprogrammen wie mit einer Befragung und Neuerfindung der eigenen Identität als Institutionen (Cook 2013).

Archivmitarbeiterinnen und -mitarbeiter werden mit immer größeren informatischen Herausforderungen und Aufgabenspektren betraut, immer öfter übernehmen Informatikerinnen und Informatiker entscheidende Rollen beim Sammlungszugang und bei der Überlieferungspräsentation. Digital- und Datenkompetenzen werden zum unverzichtbaren Handwerkszeug für moderne Literaturarchive, die sich zu Datendienstleistern wandeln – ein Prozess, mit dem sich Bibliotheken bereits seit längerer Zeit beschäftigen (exemplarisch Stäcker 2019).

Es gilt folglich, über die Aufgaben und Herausforderungen, die die fortschreitende Digitalisierung des kulturellen Gedächtnisses speziell für Literaturarchive mit sich bringt, nachzudenken und Lösungsansätze zu entwickeln, wie ihnen zukünftig begegnet werden kann. Mit dem Panel soll ein in der Community der Literaturarchive – etwa im Netzwerk "KOOP-LITERA", in der Schriftenreihe *Literatur und Archiv* (Dallinger / Kastberger 2017ff.) oder auf der Konferenz #LiteraturarchivederZukunft des Deutschen Literaturarchiv Marbach – seit einiger Zeit sich intensivierender Austausch entschieden in den Raum der Digital Humanities getragen werden.

Das Panel forciert diesen Austausch durch einen strukturiert-systematischen Impuls, bei dem wir die digitale Transformation in Literaturarchiven auf drei Ebenen adressieren:

Archivbegriff

Das Digitale, das allerorten vermeintliche 'Archive' hervorbringt, erweitert und stellt den gewachsenen Begriff des Archivs in Frage. Eine neue Phase der begrifflichen Reflexion setzt ein, in der auch Literaturarchive ihre Selbstbeschreibung überdenken.

Archivpraktiken

Digitale Werkzeuge und Infrastrukturen durchdringen die Praktiken heutiger ArchivarInnen, die beim Sammeln, Bewahren, Erschließen, Vermitteln immer häufiger zugleich Daten- und Code-expertInnen sein müssen. Im Kontext der Digitalisierung ist ein neues Verständnis von Praktiken und Handlungsprogrammen der Tätigkeiten in Literaturarchiven zu entwickeln.

Archivobjekte

Das Spektrum der Objekte, die von Literaturarchiven 'prozessiert' werden, wandelt und weitet sich. Literaturarchive befinden sich in einer Situation, in der sie die Materialität und Objektivität ihrer Bestände und Sammlungen neu begreifen müssen.

Das Panel versammelt VertreterInnen von bedeutenden Literaturarchiven aus dem DACH-Raum: Bernhard Fetz, (Literaturarchiv und Literaturmuseum der Österreichische Nationalbibliothek, Wien), Marcel Lepper (Goethe- und Schiller-Archiv Weimar), Sandra Richter (Deutsches Literaturarchiv Marbach) und Irmgard Wirtz Eybl (Schweizerisches Literaturarchiv, Bern). Die Moderation übernehmen Anna Busch und Peer Trilcke (Theodor-Fontane-Archiv, Potsdam). Mit theoretischem, konzeptionell-institutionellem und praxeologischem Blick sucht das Panel nach dem neuen Selbstverständnis der 'Digitalen Archive für Literatur'. Um die Diskussion vorzubereiten, haben die vier VertreterInnen Positionierungen und Reflexionen zu den drei systematischen Ebenen ausformuliert.

Positionierungen und Reflexionen

Bernhard Fetz

Archivbegriff

Die digitale Kommunikation des Archivs könnte eine Bewegung auslösen, so die Utopie des Archivs, die die Objekte und deren HüterInnen zu MediatorInnen eines umfassenden Bildungsbegriffs werden lässt, eines Prozesses, der Traditionen, (nationale) kulturelle Repräsentationen und die Werke der 'Großen' fluide macht. Die ubiquitäre Verfügbarkeit der Archivalien geht mit Prozessen der Entkanonisierung einher. Die Hochkultur und das 'Gipfelprinzip' verlieren an Geltung, der Fokus verschiebt sich von einzelnen Werken zu den diversen Lebens- und Arbeitsspuren in digitalen Archiven und sozialen Netzwerken (zu Tage- und Notizbüchern, biografischen Projekten, zu Recherche als Form und Selbstvergewisserung).

Archivpraktiken

Die ArchivarInnen als sichtende, selektierende, bewertende, bewahrende Instanzen mutieren zu den DatenkuratorInnen von morgen. Der Begriff des 'data curators' ist schillernd: Die digitalen KuratorInnen sind die HerrscherInnen über die Schnittstellen, sie sind aber auch SammlungsmanagerInnen, mehr oder weniger kuratorische Freigeister, abhängig vom institutionellen Selbstverständnis der Archive. Sie stellen Corpora zu bestimmten Themen in Labs oder auf Plattformen zusammen, bieten digitale Werkzeuge zu deren Nutzung an und richten die virtuellen Archivräume der Zukunft ein, in denen wir forschen, uns weiterbilden und 'erleben', in denen wir Teilhabe an Kultur erproben sollen.

Archivobjekte

Die Archivzeugen in den Depots, wiedergeboren als digitale Objekte, multiplizieren deren kulturelle und soziale Erscheinungsformen – als visualisierte, transkribierte und kommentierte Handschriften im Rahmen eines digitalen Editionsprojektes, als Ausgangspunkt von Geschichten im analogen und virtuellen Mu-

seum, als Beweisstück aus dem Webportal in der öffentlichen Debatte, als Flaschenpost in den sozialen Medien, als im Kontext eines Nachlasses zu erschließendes Objekt in der Praxis des Archivs. Die digitalen Sammelobjekte der Zukunft – seien es E-Mails, Social Media-Beiträge, Netzliteratur oder Serien – transformieren den traditionellen Literaturbegriff.

Marcel Lepper

Archivbegriff

Die digitale Transformation, die gegenwärtig in den Wissenschafts- und Kultureinrichtungen zu gestalten ist, bringt ihre eigenen öffentlichen Irrtümer und ihre eigenen falschen Begriffe mit (Francis Bacon, *Novum Organum*). Zugleich verändert die Vorstellung vom 'Netz', das angeblich 'nicht vergisst', die Wahrnehmung der Rolle und Notwendigkeit öffentlich finanzierter Archive. Der Archivbegriff verschiebt sich: von der Herrschaftsinstitution zur Beglaubigungsinstitution. Dieser Begriffswandel muss gedacht und gelebt werden.

Archivpraktiken

In der öffentlichen Wahrnehmung kämpft hochgradig ausdifferenzierte Forschung – und nicht allein historische und philologische – aktuell gegen Erfahrbarkeitsdefizite. Archive haben in den vergangenen Jahren dazu beigetragen, den abstrakten sprachlichen Gegenstand erfahrbar und vorstellbar zu machen. Nicht die Wahl zwischen der Welt des Papiers und der Welt der Daten, sondern die erfindungsreiche Gestaltung von Anschaulichkeit und Erfahrung im digitalen Modus ist die Herausforderung, vor der Archive für Literatur gegenwärtig stehen.

Archivobjekte

Flachware war lange die Krux der Literatúrausstellungen. Wie arbeiten Archive im Zeitalter von 3D und 4D? Visualisierungspraktiken, die ein Manuskript nicht mehr als Pixelfläche, sondern als Datenkubus präsentieren, und digital erzeugte Objekte nicht mehr in genetisch-qualitativer, sondern in struktural-quantitativer Form, verändern das Grundverständnis vom Gegenstand der lesenden und schreibenden Fächer.

Sandra Richter

Archivbegriff

Archive waren jahrhundertlang als räumliche Ordnungen gedacht: als Gebäude mit Gängen, Regalen, Schränken, in denen Schätze liegen, die jemand besitzt und die es aufgrund schwieriger konservatorischer Bedingungen am Ort zu untersuchen gilt. Zu den Versprechen des Digitalen gehört die virtuelle Verfügbarkeit, Durchsuchbarkeit und Erweiterbarkeit digitaler Daten, das Archiv als Literaturdatenzentrum. Damit löst sich, pointiert formuliert, der Begriff vom Archiv auf: Literaturdatenzentren kennen nurmehr virtuelle Räume, in denen existierende und künftige Daten überall zugänglich sind, sich teilen, verknüpfen und neu ordnen lassen.

Archivpraktiken

Aus den Praktiken der Datenspender und -nutzer entstehen Korpora und andere Forschungsdaten, die sich durch ihre Qualität und Anschlussfähigkeit zur Nachnutzung empfehlen. Die Vision von einem solchen Literaturdatenzentrum erscheint jedoch in mindestens zweierlei Hinsicht als unrealistisch: Zum einen lässt sich die Datenqualität und -vergleichbarkeit auf Dauer nicht nur durch temporär diese Daten Spendende und Nutzende sicherstellen, und die Daten lassen sich auch nicht einfach erhalten, ergänzen, pflegen.

Archivobjekte

Zum anderen sind die physischen Objekte, die derzeit in Archiven liegen oder dort künftig eingehen, erst in digitale Daten zu übertragen; außerdem werden Sammlungen von Forschungsdaten in einigen Jahren selbst Archivobjekte. Die Objektgruppen der Archive vervielfältigen sich durch das Digitale ein weiteres Mal, und ihre Entwicklung zu Literaturdatenzentren ist möglicherweise bloß ein weiteres historisches Stadium einer erstaunlich stabilen epistemologischen Ordnung.

Irmgard Wirtz

Archivbegriff

Literaturarchive sind mehr als Sammlungen ihrer Vor- und Nachlässe. Sie entwickeln und verwalten Wissen um die Erhaltung und die *mise en valeur* der Sammlungen. Und sie wandeln sich mit ihren Sammlungen wie mit deren Nutzung im wissenschaftlichen oder öffentlichen Interesse. Am Ende des 20. Jahrhunderts haben sich Sammlungen der Literaturarchive ausdifferenziert und ausgeweitet: Neben physischen Objekten empfangen sie zunehmend auch digitale Datenträger mit Texten, Bildern und Tönen. Die Literaturarchive gestalten die Transformationsprozesse zwischen analoger und digitaler Überlieferung und gewährleisten die Lesbarkeit.

Archivpraktiken

Derzeit befinden wir uns in einer langwierigen technischen Transformationsphase, einer *longue durée* (Fernand Braudel), die unterschätzt ist, wenn sie als einfaches Entweder/Oder besprochen wird. Bereits seit drei Jahrzehnten und bis auf weiteres erhalten die Literaturarchive Nachlässe mit polymorphen Medien: Die AutorInnen arbeiten gleichzeitig mit Bleistift, Kugelschreiber, Schreibmaschine und PC. Die Literaturarchive stehen folglich vor der Bewältigung der Simultaneität der Schreibprozesse und deren Multimedialität. Die Herausforderung besteht darin, die individuell konfigurierten Datenträger (hard & soft) zu erhalten, zu erschließen und lesbar zu machen. Dabei praktizieren Literaturarchive weiterhin institutionell die Hoheit über die Auswahl, die Kassation und die Nutzerspuren in ihren Sammlungen.

Archivobjekte

Materielle Dokumente oder digitale Daten – das ist eine falsche Distinktion, auch digitale Daten haben materielle Träger. Daten lassen sich ablösen, unterscheiden sie sich darin vom Original? Wir stehen in einer Verunsicherung in Bezug auf das Original (nicht das Kunstwerk) und seine Reproduzierbarkeit (Walter Benjamin). Dabei ist das Sammeln digitaler Dokumente nicht zu verwechseln mit der Digitalisierung von Dokumenten: Diese setzt

minimal beim Scan ein und geht maximal bis zur genetischen Edition. Eine andere Aufgabe der Literaturarchive ist das Sammeln und Aufbereiten digitaler Daten (digital born). Eine Herausforderung ist es, die technischen Zugänge und die Lesbarkeit der digitalen Datenträger aus den Anfängen des PC, der Mails und der fotografisch-/filmischen Selbstdokumentation zu gewährleisten. Wichtiger wird die Entwicklung der Standards & Normen, die Pflege und Sicherung der Metadaten und ihre intelligente Vernetzung. Erhalten (im doppelten Wortsinn) die Archive und verwalten sie künftig technisch und rechtliche aufbereitete Daten und dokumentieren deren Nutzung?

Bibliographie

Assmann, Aleida (1999): *Erinnerungsräume. Formen und Wandlungen des kulturellen Gedächtnisses*. München: C.H. Beck.

Cook, Terry (2013): "Evidence, memory, identity, and community: four shifting archival paradigms", in: *Arch Sci* (2013) 13: 95–120 <https://doi.org/10.1007/s10502-012-9180-7> [letzter Zugriff 15. Juli 2021].

Dallinger, Petra-Maria / Kastberger, Klaus (2017ff.): *Literatur und Archiv*. Berlin, Boston: De Gruyter.

Dilthey, Wilhelm (1970 [1889]): "Archive der Litteratur in ihrer Bedeutung für das Studium der Geschichte der Philosophie", in: *Archiv für Geschichte der Philosophie* II,3: 343–367.

Goethe, Johann Wolfgang von (1999): "Archiv des Dichters und Schriftstellers", in: *Goethes Werke Band X. Autobiographische Schriften*. Textkritisch durchgesehen von Liselotte Blumenthal und Waltraud Loos. München: C.H. Beck.

Lepper, Marcel / Raulff, Ulrich (2016): "Idee des Archivs", in: Lepper, Marcel / Raulff, Ulrich (eds.): *Handbuch Archiv*. Stuttgart: J.B. Metzler: 1–9 https://doi.org/10.1007/978-3-476-05388-6_1 [letzter Zugriff 15. Juli 2021].

Schögl-Ernst, Elisabeth / Stockinger Thomas / Wührer Jakob (2019): *Die Zukunft der Vergangenheit in der Gegenwart. Archive als Leuchtfeuer im Informationszeitalter*. Wien: Böhlau.

Stäcker, Thomas (2019): "Die Sammlung ist tot, es lebe die Sammlung! Die digitale Sammlung als Paradigma moderner Bibliotheksarbeit", in: *BIBLIOTHEK – Forschung und Praxis* 43(2): 304–310 <https://doi.org/10.1515/bfp-2019-2066> [letzter Zugriff 15. Juli 2021].

Szekely, Ivan (2017): "Do Archives Have a Future in the Digital Age?" in: *Journal of Contemporary Archival Studies*, Vol. 4, Article 1: 1–16 <http://elischolar.library.yale.edu/jcas/vol4/iss2/1> [letzter Zugriff 15. Juli 2021].

Thaler, Jürgen (2011): "Zur Geschichte des Literaturarchivs. Wilhelm Diltheys Archive für Literatur im Kontext", in: *Schiller-Jahrbuch* 55: 361–374.

Theimer, Kate (2012): "Archives in Context and as Context", in: *Journal of Digital Humanities* Vol. 1, No. 2 <http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/> [letzter Zugriff 15. Juli 2021].

Theimer, Kate (2018): "Partizipation als Zukunft der Archive", in: *ARCHIVAR*, 71. Jahrgang, Heft 01: 6–12 <https://archive20.hypotheses.org/files/2018/03/Aufsatz-Theimer.pdf> [letzter Zugriff 15. Juli 2021].

Erinnern durch Vernetzen Digitale Sammlungsforschung

Alschner, Stefan

stefan.alschner@klassik-stiftung.de
Klassik Stiftung Weimar, Forschungsverbund MWW

Baumgarten, Marcus

baumgarten@hab.de
Herzog August Bibliothek Wolfenbüttel, Forschungsverbund MWW

Horstmann, Jan

jan.horstmann@uni-muenster.de
Universität Münster, Universitäts- und Landesbibliothek

Müller, Christiane

christiane.mueller@klassik-stiftung.de
Klassik Stiftung Weimar, Forschungsverbund MWW

Nantke, Julia

julia.nantke@uni-hamburg.de
Universität Hamburg, Institut für Germanistik

Weis, Joëlle

weis@uni-trier.de
Trier Center for Digital Humanities (TCDH)

Wübbena, Thorsten

wuebbena@ieg-mainz.de
Leibniz-Institut für Europäische Geschichte (IEG)

Perspektiven der Sammlungsforschung

Die Vernetzung von Daten ist immer auch eine Vernetzung von Wissensbeständen. In der digitalen Sammlungsforschung eröffnen sich mit verschiedenen Technologien der Referenzierung und Relationierung Möglichkeiten, ein Wissensnetzwerk aufzubauen, das je nach Bedarf durchsuchbar, in seiner Gesamtheit erschließbar und so offen gestaltet ist, dass es in ein globales Netz des Wissens integriert werden kann. Kulturelles Erbe und damit Erinnerungskultur können somit – zumindest in quantitativer Hinsicht – auf ein bislang ungesehenes Niveau gehoben werden. Dass mit der großen Menge an Erinnerungsdaten auch eine verlässliche und einheitliche hohe Qualität einhergeht, bleibt Herausforderung.

Neben großen Infrastrukturprojekten im Sammlungsbereich wie OCLC WorldCat oder Europeana sind etwa Initiativen wie Wikimedia oder Google Arts and Culture hervorzuheben, die auch jenseits der Forschung maßgeblich dazu beitragen, Sammlungsdaten miteinander zu verknüpfen. Internationale Projekte, in denen Forschende und Institutionen sich vernetzen können wie etwa

DARIAH oder UNIVERSEUM, auf deutscher Ebene auch die Koordinierungsstelle für wissenschaftliche Universitätsammlungen, bilden eine entscheidende Infrastruktur zur Vernetzung von Personen und Sammlungen.

Vernetzung und daraus resultierende Projekte finden häufig gerade auch auf persönlicher Ebene statt. Es bedarf nicht immer übergeordneter Strukturen, um wissenschaftlichen Fortschritt zu bewirken. Gerade dieses Vernetzen von Wissensbeständen im Kleinen kann außerordentlich fruchtbar sein, um zu innovativen und flexiblen Methoden der Modellierung und Darstellung des jeweiligen Forschungsvorhabens zu gelangen. Wir betrachten Herausforderungen im Großen wie im Kleinen, mit denen sich einzelne Projekte konfrontiert sehen, als auch Möglichkeiten, wie diese im gemeinsamen Austausch – sowohl in formellen Kontexten als auch auf individueller Ebene – angegangen werden können.

Die Relationierung von Sammlungsdaten ist häufig auch die Vernetzung unterschiedlicher Personen, die an diesen Datenbeständen arbeiten. Ein Netzwerk von Sammlungsdaten ist damit immer auch eine Infrastrukturaufgabe. Projekte, sammlungsführende Gedächtnisinstitutionen, interdisziplinäre Center und ortsübergreifend tätige Verbünde müssen sich nicht nur der Herausforderung stellen, ihre jeweiligen Daten zu erzeugen und in einer weiterbearbeitbaren und Standards entsprechenden Form zugänglich zu machen, sie müssen immer auch die infrastrukturellen Voraussetzungen schaffen oder sich in vorab definierte Infrastrukturen einfinden, um erfolgreich zu arbeiten und verschiedene Ansätze der Sammlungsdigitalisierung, -erschließung und -forschung zu koordinieren.

Im Panel werden diese Arbeitsfelder zusammengeführt und treten in einen Dialog. Zur Sprache kommen dabei nicht nur technische, kommunikative und informative Infrastrukturmaßnahmen, wie sie im Digitalen Labor des Forschungsverbunds Marbach Weimar Wolfenbüttel (MWW) etwa im dort entwickelten Virtuellen Forschungsraum (VFR) ergriffen werden. Die Rede wird auch sein von Möglichkeiten, Besonderheiten und Herausforderungen des vernetzten digitalen Servicemanagements, beispielhaft demonstriert am an der ULB Münster angesiedelten und interdisziplinär agierenden Service Center for Digital Humanities (SCDH). Die unterschiedlichen Tätigkeitsbereiche der digitalen Sammlungsforschung mit ihren je spezifischen Herausforderungen liefern ebenfalls Input: Die Volltextdigitalisierung und Relationierung benannter Entitäten eines handschriftlichen Briefnetzwerks hat sich das Projekt „Dehmel Digital“ an der Universität Hamburg zum Ziel gesetzt. Auch fragen wir, wie die Netzwerkanalyse als Methode der Relationierung von Daten(sätzen) etwa im Zuge von frühneuzeitlichen Bibliotheksrekonstruktionen einen epistemischen Wandel erforschbar und abbildbar machen kann. Schließlich beleuchten wir zwei sich ergänzende Ansätze, digital unterschiedlich (d.h. mit differierenden Metadaten-schemata) erschlossene Sammlungen durchsuchbar und erschließbar zu machen. Der Sammlungsübergreifenden Suche liegt ein im Forschungsverbund MWW entwickeltes gemeinsames Metadatenformat zugrunde, in das auch materiell divergierende Sammlungsbestände und digitale Editionen integriert und dadurch übergreifend durchsuchbar gemacht werden können. Einen anderen Ansatz verfolgt das ebenfalls in MWW angesiedelte Projekt des sogenannten Virtuellen Sammlungsraums (VSR): Die Webanwendung soll es durch eine im Hintergrund agierende Graphdatenbank niedrigschwellig und visuell ermöglichen, auf Sammlungsdatenbanken unterschiedlicher Formate zuzugreifen und semantische Relationen zwischen Objekten erschließen bzw. ergänzen zu können. Bei diesem Ansatz bildet kein gemeinsames Metadatenformat die Orchestrierungsschicht zwischen den einzelnen Datensätzen, sondern die jeweiligen digitalen Sammlun-

gen werden in ihrer ursprünglichen Form angesteuert und durch Breiten- und Tiefensuche als Graph erschlossen, ohne dass es zu einem Informationsverlust durch Angleichung unterschiedlicher Schemata käme.

Beteiligte Teilprojekte

Digitales Labor: Technische, kommunikative und informative Infrastrukturen

Technisches Herzstück des Forschungsverbunds MWW ist das Digitale Labor mit seinem Virtuellen Forschungsraum (VFR: <https://vfr.mww-forschung.de/>). Der VFR bietet Forscher:innen an verschiedenen Standorten digitalen Zugang zu den Sammlungen der drei Verbundeinrichtungen und interaktive computergestützte Arbeitsmöglichkeiten für kooperative Forschungsprojekte (vgl. Dogunke / Steyer 2019). Der Forschungsraum als vernetzende Infrastruktur ist damit auch digitales Zugangs- und Arbeitsportal zu den Forschungsprojekten des Verbunds und den Verbundeinrichtungen. Jeweils projektspezifische Labore bieten Forscherteams ein Set an Tools, um Kommunikation sowie bestandsbezogene Forschung von der Korpusbildung über Analyse- und Auswertungsverfahren bis hin zur Veröffentlichung von Forschungsergebnissen kollaborativ gestalten zu können. Die Integration weiterer Tools in den Virtuellen Forschungsraum erfolgt sukzessive und bedarfsorientiert. Basierend auf der Taxonomie TaDiRAH (<http://tadirah.dariah.eu/vocab/>) werden dabei sämtliche Forschungsprozesse abgedeckt. Die im Forschungsverbund angesiedelten Fallstudien verstehen sich als Anwendungsfelder für die bereits entwickelten digitalen Verfahren und vernetzen durch ihre Impulse die Forschungs- und Anwendungs- mit der Entwicklungsperspektive.

Service Center for Digital Humanities

Disziplinenübergreifend und damit über digitale Methoden vernetzend ist das an der Universitäts- und Landesbibliothek angesiedelte Service Center for Digital Humanities (SCDH, <https://www.uni-muenster.de/SCDH/>) an der Universität Münster. Gerade die Unabhängigkeit einer konkreten fachlichen Zuordnung ermöglicht es den Beratenden des SCDH, methodische Lösungen für konkrete Probleme aus ganz unterschiedlichen Bereichen der geisteswissenschaftlichen Forschung miteinander zu vernetzen und so einen transdisziplinären Austausch zu evozieren. Dabei kommt es einerseits zu disziplinspezifischen Herausforderungen etablierter DH-Methoden (etwa die OCR oder gar HTR mathematischer Formeln in Texten). Andererseits kristallisiert sich ein Set an Basisdiensten heraus, die in verschiedensten fachlichen Kontexten die digitale Arbeit mit kulturellen Artefakten ermöglichen, transformieren und im Idealfall bereichern und verbessern kann. Den Grundsätzen von "Open Science" (vgl. Heise 2018) verpflichtet ermöglicht die Konzeption des SCDH den freien Austausch von Informationen und Wissen.

Dehmel Digital

Ziel des Projekts Dehmel digital ist die sukzessive wissenschaftliche Erschließung, Zugänglichmachung und Beforschung des umfangreichen Korrespondenznetzes von Ida und Richard Dehmel im Rahmen einer digitalen Plattform. Hierbei steht dezi-

diert nicht die Künstlerpersönlichkeit Richard Dehmel, sondern das Netzwerk als dynamischer Verbund von Akteur:innen im Projektfokus. In Kooperation mit der Staats- und Universitätsbibliothek Hamburg werden ca. 35.000 handschriftliche Originalbriefe von namhaften Verfasser:innen wie Stefan Zweig, Rainer Maria Rilke, Else Lasker-Schüler, Arnold Schönberg uvm. an das Ehepaar Dehmel sowie die teilweise archivierten Antworten der Dehmels mittels digitaler Werkzeuge teilautomatisiert transkribiert. Die Transkriptionen bilden im Projekt die Basis für die weitere inhaltliche Erschließung, die ebenfalls mittels computergestützter Verfahren erfolgt (vgl. Baillot 2020, Nutt-Kofoth 2016). Auf einem Webportal sollen ab Anfang 2022 sukzessive die im Projekt erzeugten Daten zum Download verfügbar gemacht sowie das Briefnetzwerk in strukturierter Form zur skalierbaren Rezeption bereitgestellt werden.

Bibliotheksrekonstruktion

Im Kontext des Forschungsverbunds MWW wurde die digitale Rekonstruktion historischer Privatbibliotheken des 17. und 18. Jahrhunderts in den letzten Jahren intensiv verfolgt (<http://bibliotheksrekonstruktion.hab.de/>; vgl. auch Beyer et al. 2017). Über die einzelnen Fallstudien hinaus liegt das große Erkenntnispotential in der Vernetzung dieser Datensätze, die es ermöglicht, Bestände miteinander in Beziehung zu setzen und zu vergleichen: Wie verändert sich das Sammlungsinteresse im Laufe der Zeit? Inwiefern unterscheiden sich Bibliotheken von Männern und Bibliotheken von Frauen? Gibt es Standardwerke? Gibt es die typische Privatbibliothek (vgl. Weis 2021)? Die Basis der Verknüpfung bilden Normdaten auf Werkebene, die in Wikidata flexibel angelegt werden können und den Zugang auf vielfältige Ressourcen des Semantic Web garantieren (vgl. Lemus-Rojas / Pintscher 2018). Das finale Ziel ist ein Sammlungsnetzwerk, das einen systematischen Einblick in die Bücherwelten der Frühen Neuzeit erlaubt.

Digitale Editionen in der Sammlungsübergreifenden Suche

Die Herzog August Bibliothek in Wolfenbüttel erarbeitet seit fast 20 Jahren digitale Editionen und Volltexttranskriptionen im TEI/XML-Format. Diese Vielzahl kleiner und großer Editionsprojekte (vgl. <https://www.hab.de/digitale-editionen/>) stellen innerhalb der Bibliothek mittlerweile eine eigene Sammlung dar, die in der Wolfenbütteler Digitalen Bibliothek präsentiert wird. Mit der Editionsarbeit einher geht eine gewisse Heterogenität bei der Strukturierung und Kodierung der Quellen, die eine einheitliche Suche über alle Editionseinheiten erheblich erschwert. Eine mögliche Lösung für diese Herausforderung stellt die sammlungsübergreifende Suche des MWW-Verbunds (<https://vfr.mww-forschung.de/suche>) dar. Hier werden von den beteiligten Einrichtungen Sammlungen definiert, die über ein gemeinsames Metadatenformat (MMM) für eine sammlungsübergreifende Suche zur Verfügung gestellt werden (vgl. Gradl et al. 2017). Um das Mapping auf ein einheitliches Format in Zukunft zu erleichtern, wird darüber hinaus von DH-Entwicklern parallel ein HAB-Schema entwickelt und die Normierung von Normdaten erarbeitet, die zukünftig für möglichst viele Editionen an der Herzog August Bibliothek gelten sollen. Dabei müssen Fragen beantwortet werden wie z.B.: Was ist der Kern von historisch-kritischen Editionen? Kann es ein Basisformat für Editionen geben? Wie individuell darf eine Edition sein? Welche Fragestellungen bzw.

Suchanfragen muss eine editionsübergreifende Suche beantworten können?

Sammlungserschließung durch Vernetzung heterogener Objekte

Sammlungen lassen sich als komplexe Beziehungsnetzwerke aus Objekten, verschiedenen Akteuren (Sammler:innen, Künstler:innen, Institutionen etc.) sowie weiteren Entitäten (z. B. Geographika) verstehen. Um nach diesem Verständnis der Sammlung als Forschungsgegenstand gerecht zu werden, stoßen Katalogdarstellungen, welche die Bestände nur auflisten, an ihre Grenzen, da diese nur einen Ausschnitt (Mikroperspektive) der Sammlung darzustellen vermögen. Der Virtuelle Sammlungsraum (VSR) ist als Webanwendung konzipiert, welche die Sammlung als Ganzes in den Blick nimmt. Aufbauend auf einer Graphdatenbank soll es möglich sein, Metadaten aus unterschiedlichen Erfassungssystemen zusammenzuführen und Beziehungen zwischen Objekten, Akteuren und weiteren möglichen Informationsknotenpunkten erfassbar und erfahrbar zu machen (vgl. Oldman / Tanase 2018, Werner 2021). Letzteres soll durch interaktive Visualisierungen geschehen, mit welchen sich das Phänomen "Sammlung" aus ganz unterschiedlichen Blickwinkeln beleuchten lässt.

Ablauf

Im Panel sollen die einzelnen Projekte in kürzeren Statements (insg. 30 Minuten) vorgestellt und ihr jeweiliger Bezug zur Vernetzungsthematik deutlich gemacht werden. Im Zentrum soll der multiperspektivische Austausch über das Phänomen sowie Fragen nach Möglichkeiten, Grenzen und institutionelle Bedingungen des systematischen Vernetzens in der Erinnerungsarbeit stehen. Wir fragen etwa ob die Konzeption der unterschiedlichen Bereiche als Netzwerk "nur" ein hilfreiches theoretisches Konstrukt ist oder ein konkreter praktischer Mehrwert entsteht, oder inwieweit digitale Vernetzung ein Verständnis von Sammlung konterkariert bzw. erweitert. Mindestens die letzten 30 Minuten sollen genutzt werden für die Diskussion mit dem Publikum. Die große Bandbreite der Projektvorstellungen soll hervorheben, auf welchen unterschiedlichen Ebenen Vernetzung den Forschungsprozess befördert oder für diesen integraler Bestandteil ist. In der Moderation wollen wir den Status der Beispielhaftigkeit der vorgestellten Projekte betonen und in der Diskussion von ihnen abstrahieren. Das Panel wird unterstützt durch die Möglichkeit der Beteiligung per Twitter, während und nach der Diskussion. Außerdem möchten wir im vorgestellten Virtuellen Forschungsraum (VFR) ein Labor mit Forum- und Etherpad-Funktionalität einrichten, in dem niedrigschwellig weitere Erfahrungen und Argumente gesammelt werden können. Auf diese Weise können wir bereits im Vorfeld des Panels die Diskussion bündeln und die spontane Vernetzung der Diskutant:innen ermöglichen. Das Labor wird zudem dauerhaft öffentlich zugänglich bleiben.

Bibliographie

Baillot, Anne (2020): "Digitalisierung und ihre Einflüsse auf den Umgang mit alten wie neuen ‚Briefen‘ in deutscher wie internationaler Perspektive" in: Matthews-Schlinzig, Marie Isabel / Schuster, Jörg / Steinbrink, Gesa / Strobel, Jochen (eds.): *Handbuch Brief: von der Frühen Neuzeit bis zur Gegenwart. Band 1:*

Interdisziplinarität – systematische Perspektiven – Briefgenres. Berlin / Boston: De Gruyter 387–398.

Beyer, Hartmut / Münkner, Jörn / Steyer, Timo / Schmidt, Katrin (2017): "Bibliotheken im Buch: Die Erschließung von privaten Büchersammlungen der Frühneuzeit über Auktionskataloge" in: Busch, Hannah / Fischer, Franz / Sahle, Patrick (eds.): *Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4*, Norderstedt: Books on Demand 43–70. URN: urn:nbn:de:hbz:38-77794.

Dogunke, Swantje / Steyer, Timo (2019): "Virtuell Zusammenwachsen: Konzeption, Aufbau und Intention der digitalen Forschungsinfrastruktur im Forschungsverbund MWW" in: Huber, Martin / Krämer, Sybille / Pias, Claus (eds.): *Forschungsinfrastrukturen in den digitalen Geisteswissenschaften: Wie verändern digitale Infrastrukturen die Praxis der Geisteswissenschaften?* Symposienreihe „Digitalität in den Geisteswissenschaften“, Frankfurt am Main: CompaRe, 111–128. URN: urn:nbn:de:hebis:30:3-519476.

Gratl, Tobias / Aschauer, Anna / Dogunke, Swantje / Klaffki, Lisa / Schmunk, Stefan / Steyer, Timo (2017): "Daten sammeln, modellieren und durchsuchen mit DARIAH-DE" in: *Konferenzpaper DHd 2017: Digitale Nachhaltigkeit*, Bern (Schweiz), 13. bis 18. Februar 2017. DOI: 10.5281/zenodo.582316.

Heise, Christian (2018): *Von Open Access zu Open Science: Zum Wandel digitaler Kulturen der wissenschaftlichen Kommunikation*. Lüneburg: meson press. DOI: 10.14619/1303.

Lemus-Rojas, Mairelys / Pintscher, Lydia (2018): "Wikidata and Libraries: Facilitating Open Knowledge" in: Proffitt, Merrilee (ed.): *Leveraging Wikipedia: Connecting Communities of Knowledge*. Chicago, IL: ALA Editions 143–158.

Nutt-Kofoth, Rüdiger (2016): "Briefe herausgeben: Digitale Plattformen für Editionswissenschaftler und die Grundfragen der Briefedition" in: Richts, Kristina / Stadler, Peter (eds.): *"Ei, dem alten Herrn zoll' ich Achtung gern": Festschrift für Joachim Veit zum 60. Geburtstag*. München: Allitera. 575–586.

Werner, Claus (2021): "Die Sammlung als Netz" in: Andraschke, Udo / Wagner, Sarah (eds.): *Objekte im Netz*. Bielefeld: transcript Verlag, 247–260. DOI: 10.14361/9783839455715-018.

Oldman, Dominic / Tanase, Diana (2018): "Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace" in: Vrandečić D. et al. (eds.): *The Semantic Web – ISWC 2018. ISWC 2018. Lecture Notes in Computer Science, Part 1*, Cham: Springer, 325–340. DOI: 10.1007/978-3-030-00671-6.

Weis, Joelle (2021): "Women's private libraries as spaces of knowledge making. The cases of Elisabeth Sophie Marie and Philippine Charlotte of Braunschweig-Wolfenbüttel" in: Klein Käfer, Natacha / Silva Perez, Natália da (eds.): *Practices of Privacy: Knowledge in the Making*. Amsterdam: Amsterdam University Press (im Druck).

Kinetik und Methodik

Film als dynamische und multimodale Herausforderung für die DH

Bateman, John

bateman@uni-bremen.de
Universität Bremen

Diecke, Josephine

diecke@staff.uni-marburg.de
Universität Marburg

Ewerth, Ralph

ralph.ewerth@tib.eu
Universität Hannover

Heftberger, Adelheid

adelheidh@gmail.com
Bundesarchiv

Howanitz, Gernot

gernot.howanitz@uibk.ac.at
Universität Innsbruck, Austria

Spiegel, Simon

simon@simifilm.ch
Universität Zürich

Loertscher, Miriam

miriam.loertscher@zhdk.ch
Zürcher Hochschule der Künste

Motivation

Wo sind sie, die ‚kinetischen‘ Methoden für die Filmanalyse, die dem Film als sowohl dynamisches als auch multimodales Medium zumindest ein Stück weit gerecht werden können? Das geplante Panel versucht, der Beantwortung dieser Frage ein Stück näher zu kommen und adressiert damit ein Forschungsdesideratum in den DH: Filme sind – gerade auch aus quantitativer Sicht – notorisch schwer zu bändigen. Einerseits sind sie – multimodal – aus Bild, Text und Ton zusammengesetzt, andererseits bestehen sie – dynamisch – nicht nur aus Bildern, sondern eben aus *bewegten* Bildern, eine Tatsache, die gleich etliche Ebenen an Komplexität hinzufügt.

Verständlicherweise beschäftigen sich viele DH-Projekte gegenwärtig mit Einzelaspekten des Mediums Film, wobei hier in der Regel besonderes Augenmerk auf das Visuelle gelegt wird. Beispielsweise bietet das von Taylor Arnold und Lauren Tilton initiierte *Distant Viewing Toolkit*, ein etabliertes Python-Paket zur quantitativen Filmanalyse, zwar rudimentäre Aggregatoren für

Audio und Untertitel an, stellt aber ungleich mehr Möglichkeiten für die Analyse von Einzelrahmen zur Verfügung (Arnold / Tilton 2019). Diese Fokussierung auf einige wenige Teilelemente des Mediums ‚Film‘ wird häufig mit der im Vergleich zu Text oder Bild vielfach höheren Datenmenge argumentiert; ein Argument, das jedoch letztlich ins Leere laufen muss, können die DH ja gerade hier ihren Trumpf ausspielen.

Immerhin zeigen einige Forschungsprojekte aus der Computer Vision, dass neuronale Netze auch auf das Erkennen komplexer Bewegungsabläufe und letztlich ganzer Szenen bzw. szenischer Handlungen trainiert werden können und konkret bei der Erkennung von Gewaltszenen (Mumtaz et al. 2020) oder bei der automatisierten textuellen Beschreibung von Handlungsabläufen (Park et al. 2018) eingesetzt werden. Diese innovativen Methoden aus dem Deep Learning machen es noch dringlicher, gesamtheitlichere Zugänge zum Medium Film zu finden, und dieses nicht einfach als ‚bag of images‘ misszuverstehen, weil sie aufeinander abgestimmt und in den methodischen Kontext der Geisteswissenschaften eingebettet werden – ein Ansinnen, das uns vor Herausforderungen stellt. So nimmt es nicht wunder, dass die DH nicht die neuesten Methoden der Computer Vision einsetzt (vgl. dazu die Methodenübersicht in Pustu-Iren et al.)

Nun gibt es in den letzten Jahren auch im deutschsprachigen Raum einige große Forschungsprojekte, die eine ganzheitlichere Betrachtung des Mediums Film propagieren. So haben etwa das FilmColors-Projekt mit dem Annotationstool VIAN (ERC Advanced Grant, Zürich; Flückiger / Halter 2020), der von der Volkswagen Stiftung geförderte „Digital Cinema Hub“ (Marburg, Frankfurt und Mainz) sowie „TIB AV Analytics“ (Hannover, DFG) erste vielversprechende Versuche geliefert und setzen gleichzeitig sowohl in methodologischer Sicht als auch hinsichtlich der Forschungsfragen verschiedene Schwerpunkte. Zu klären bleibt deshalb die Frage: Wie führt man Ergebnisse aus Deep Learning mit visuellen Medien, Audio und Text mit empirischen Ergebnissen, Eye-Tracking-Analysen, etc. zusammen?

Das geplante Panel möchte aber nicht einfach die eben skizzierten Probleme beschreiben, sondern Wissenschaftlerinnen und Wissenschaftler aus den unterschiedlichsten Disziplinen und Forschungskontexten zusammenbringen, um mögliche Lösungswege aufzuzeigen und eine dem Medium Film entsprechende ‚kinetische‘ Methodologie zu skizzieren. Dabei wird nicht nur eine (forschungsgeleitete) Anwendungsperspektive eingenommen; vielmehr sollen ebenso Aspekte aus der Lehre, der (Film-)Kulturvermittlung und der Theoriebildung berücksichtigt werden.

Organisation des Panels

Das Panel verschreibt sich ganz einer lebendigen Diskussionskultur. Aus diesem Grund verzichten wir auf Kurzreferate der Teilnehmerinnen und Teilnehmer und setzen auf pointierte Eröffnungsstatements. Dabei bringt jede Teilnehmerin und jeder Teilnehmer eine spezifische Perspektive, eigene Forschungsfragen und Anliegen in die Diskussion mit ein, die im folgenden Abschnitt kurz umrissen werden. Die starke Fokussierung auf das Gespräch soll auch helfen, das Publikum verstärkt anzusprechen. Um der Diskussion einen Rahmen zu geben, verschickt die Moderation im Vorfeld Leitfragen, die helfen sollen das Gespräch in Gang zu setzen, zu fokussieren und – falls notwendig – zu einem roten Faden zurückzukehren.

Der Zeitplan sieht eine fünfminütige Einführung durch den Moderator vor, die von fünf zweiminütigen Kurzstatements der Teilnehmerinnen und Teilnehmer abgerundet werden. Es folgt ein of-

fenes Panelgespräch entlang der vorher verschickten Leitfragen im Umfang von 45 Minuten, bei dem das Publikum natürlich auch gerne eingreifen darf, die letzten dreißig Minuten sind explizit für Fragen aus und Diskussion mit dem Publikum vorgesehen; dabei möchten wir auch mit einer Social-Media-Komponente (Twitter) und anderen mit etablierten digitalen Feedbacklösungen aus der Hochschuldidaktik experimentieren, die das Eintreten in einen Dialog möglichst niederschwellig gestalten sollen.

Spezifische Perspektiven

Prof. Dr. John Bateman (Bremen) widmet sich der Frage, inwieweit Analysetools für dynamische Medien auch dynamisch sein müssen. Sind dynamische Werkzeuge zur Auseinandersetzung mit Daten für die Theoretisierung geeignet oder sind sie eher Hilfsmittel zur Datenexploration in Abgrenzung zur Theoriebildung? Was könnten dynamische Theorien sein? Ist der Begriff in sich kohärent? Inwieweit könnten dynamische Werkzeuge für den Umgang mit Daten und Datenanalysen von unserem Wissen über die Funktionsweise von Film, insbesondere von kinematografischen Ausdruckstechniken, profitieren?

Dr. Adelheid Hefberger (Bundesarchiv Berlin) interessiert, wie es der DH-Community (das schließt nicht nur Wissenschaftler*innen, sondern auch Kulturerbe-Verantwortliche ein) gelingen könnte, von Pilotprojekten, ‚Hacks‘, Einzellösungen für Projekte und fertigen Softwarelösungen – sprich: Einzelforschungsinteressen – hin zu Anwendungen in großem Stil (Fernseharchive) zu kommen. Eine wesentliche Frage ist dabei: Welche Fragestellungen sind überhaupt möglich, wenn eben nicht mit professionellen Tools gearbeitet werden kann?

Josephine Diecke (Marburg) sieht ihren Schwerpunkt in der Reflexion von ‚kinetischen‘ Methoden für die Filmanalyse im Kontext der Lehre. Im Rahmen des VW-Projektes „Digital Cinema Hub“ an den Universitäten Marburg, Frankfurt und Mainz beschäftigt sie sich mit der Integration von digitalen Tools und Methoden in die Lehre und Forschung. Als zentrale Herausforderung für die Film- und Medienwissenschaft sieht sie die Verknüpfung von aktuellen Tools für die Filmanalyse mit etablierten Methoden bzw. auch die bewusste Abgrenzung dieser beiden Pole.

Prof. Dr. Ralph Ewerth (Hannover) weist auf die großen Fortschritte hin, die Deep Learning für die Mustererkennung von audiovisuellen Daten gebracht hat. Von besonderem Interesse ist, für welche Analysen (z.B. Szenengrenzen, Kamerabewegung, Spracherkennung) Methoden direkt anwendbar sind und welche neuen Möglichkeiten sich für die systematische Filmanalyse ergeben. Weiterhin ist eine offene Frage, wie solche Algorithmen auf einfache Weise für Forschende in den Filmwissenschaften verfügbar und anwendbar werden. Hierzu können Ergebnisse des DFG-Projekts „TIB AV Analytics“ diskutiert werden, im Rahmen dessen in interdisziplinärer Kooperation zwischen Filmwissenschaft und Informatik eine offene Web-basierte Analyse-Plattform für diese Zwecke entwickelt wird.

Dr. Miriam Loertscher und PD Dr. Simon Spiegel (Zürich) haben im Rahmen des ERC-Forschungsprojekts „FilmColors. Bridging the Gap Between Technology and Aesthetics“ (PI: Prof. Dr. Barbara Flückiger) zusammen mit anderen die Annotationssoftware VIAN entwickelt, die im Rahmen der Keynote der DHd 2019 vorgestellt wurde und mittlerweile um neue Funktionen erweitert worden ist. So kann VIAN nun auch die Daten von Eye-Tracking-Experimenten verarbeiten und anzeigen; zudem wird die Software aktuell mit Funktionen für die Gesprächsanalyse erweitert. Besonders spannend erscheint es im Kontext des Panels,

auf die Verbindung von ästhetischen Parametern und empirischen Daten einzugehen, die VIAN leisten soll.

Bibliographie

Arnold, Taylor / Tilton, Lauren (2019): „Distant Viewing: Analyzing Large Visual Corpora“, in: *DSH*, fqz013, <https://doi.org/10.1093/digitalsh/fqz013> [letzter Zugriff: 15. Juli 2021].

Burghardt, Manuel / Heftberger, Adelheid / Pause, Johannes / Walkowski, Niels-Oliver / Zeppelzauer, Matthias (2020): „Film and Video Analysis in the Digital Humanities – An Interdisciplinary Dialog“, in: *DHQ* 14.4 <http://digitalhumanities.org/dhq/vol/14/4/000532/000532.html> [letzter Zugriff: 15. Juli 2021].

Flückiger, Barbara / Halter, Gaudenz (2020): „Methods and Advanced Tools for the Analysis of Film Colors in Digital Humanities“, in: *DHQ* 14.4, digitalhumanities.org/dhq/vol/14/4/000500/000500.html [letzter Zugriff: 15. Juli 2021].

Mumtaz, Aqib / Bux Sargano, Allah / Habib, Zulfiqar (2020): „Fast Learning Through Deep Multi-Net CNN Model For Violence Recognition In Video Surveillance“, in: *The Computer Journal*, bxaa061, <https://doi.org/10.1093/comjnl/bxaa061> [letzter Zugriff: 15. Juli 2021].

Park, Jae Sung / Rohrbach, Marcus / Darrell, Trevor / Rohrbach, Anna (2018): „Adversarial Inference for Multi-Sentence Video Description“, in: *arXiv*, 1812.05634, <https://arxiv.org/abs/1812.05634>.

Pustu-Iren, Kader / Sittel, Julian / Mauer, Roman / Bulgakowa, Oksana / Ewerth, Ralph (2020): „Automated Visual Content Analysis for Film Studies: Current Status and Challenges“, in: *DHQ* 14.4, <http://digitalhumanities.org/dhq/vol/14/4/000518/000518.html> [letzter Zugriff: 15. Juli 2021].

Shou, Zheng / Lin, Xudong / Kalantidis, Yannis / Sevilla-Lara, Laura / Rohrbach, Marcus / Chang, Shih-Fu / Yan, Zhicheng (2019): „DMC-Net: Generating Discriminative Motion Cues for Fast Compressed Video Action Recognition“, in: *arXiv*, 1901.03460, <https://arxiv.org/abs/1901.03460> [letzter Zugriff: 15. Juli 2021].

Kultur – Daten –

Kuratierung

Was speichern wir und wozu?

Altenhöner, Reinhard

Reinhard.Altenhoener@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Dieckmann, Lisa

Lisa.dieckmann@uni-koeln.de
Universität zu Köln

Münzmay, Andreas

Andreas.muenzmay@uni-paderborn.de
Universität Paderborn

Pratschke, Margarete

Margarete.pratschke@hu-berlin.de
HU Berlin

Primavesi, Patrick

primavesi@uni-leipzig.de
Universität Leipzig

Richts-Matthaei, Kristina

kristina.richts@uni-paderborn.de
Universität Paderborn, Germany

Röwenstrunk, Daniel

roewenstrunk@uni-paderborn.de
Universität Paderborn

Schulz, Christoph

cbschulz@web.de
FH Düsseldorf

Stellmacher, Martha

Martha.Stellmacher@slub-dresden.de
SLUB Dresden

Was speichern wir und wozu? Was wird verdeckt, was verschwindet? Wie kann der inhaltliche Sinn erhalten bleiben? Welche Daten, Dokumentationen oder Programme sind für das Verstehen der Erkenntnisse notwendig? Was ist es wert, erhalten zu bleiben und für wen? Was macht es mit 'Kultur', wenn wir sie digital speichern, oder nicht speichern, oder nur für eine gewisse Zeit speichern? Besteht die Gefahr, heute etwas 'wegzusortieren', was morgen wichtig ist? Inwieweit überlassen wir das dem Zufall, inwieweit den technischen Routinen, inwieweit den Expert:innen?

Beschreibung des Themas

Die wissenschaftliche Beschäftigung mit Gegenständen der materiellen und immateriellen Kultur basiert in steigendem Maße auf digitalen Prozessen und auf der Arbeit mit digitalen Objekten. Die Bedeutung IT-gestützter Auswertungsverfahren nimmt ebenfalls zu. Häufig entstehen im Zusammenhang solcher Projekte neue Korpora von Digitalisaten, spezifische Modelle, formale Beschreibungen und Verknüpfungen von materiellen oder immateriellen Objekten, Ereignissen und Wissensbeständen, beispielsweise in Form von 3D-Digitalisierungen, Scans, Kodierungen, Audio- und Videoaufnahmen oder Simulationen. Aber auch wenn vorhandene Digitalisate herangezogen, angereichert, transformiert, verknüpft, visualisiert usw. werden, entstehen vielgestaltige Datenbestände von mitunter hohem Komplexitätsgrad, deren Relevanz sich keineswegs in der Funktion beispielsweise des 'Anhangs' zu klassischen Publikationsformen – Journal Papers, Konferenzbeiträge, Qualifikationsschriften usw. – erschöpft. Vielmehr gehen kulturbezogene Datensätze potenziell selbst als Arbeitsdaten, d.h. als Gegenstand, in den wissenschaftlichen, aber auch in den kulturellen Diskurs bis hin zur kulturökonomischen Praxis ein.

Im Allgemeinen stehen bei der Entwicklung von Standards für Datenkuratierung und Langzeitarchivierung technische Belange im Zentrum – also das „Wie?“ der Speicherung und die Beschaffenheit bzw. Qualität der Daten (Zhou 2021). Nicht weniger dringlich für die Aggregation digitaler Kulturdatenkorpora ist aber die vorgelagerte, grundlegende Frage nach der Datenauswahl: Was speichern wir? Warum sollen wir es speichern? Und wie können künftig auch die Perspektiven der Beteiligten an Datenproduktion und -nutzung in den Entscheidungsprozess einbezogen werden?

Das Panel regt daher einen Perspektivwechsel an, hin zu forschungsgeleiteten Entscheidungswegen in Bezug auf die Fragen, was in die nachhaltige Bewahrung eingeht, und welche wahrscheinlichen Anforderungen zukünftiger Forschung an diese Objekte dabei in den Blick genommen werden müssen. Bei aller Schwierigkeit prognostischer Bewertungen geht es hier darum, die Sicht derjenigen besser zu verstehen, die konkret mit diesen Daten arbeiten, also den „designated communities“, wie es der OAIS-Standard formuliert (Reference Model 2012): Wie werden die Perspektiven, die Wissenschaftler:innen auf die Daten haben, angemessen und konkret in die Kuratierungsprozesse einbezogen? Welche Anforderungen für Infrastruktureinrichtungen ergeben sich hieraus?

Damit ist zugleich ein wesentliches Aufgabenprofil des Konsortiums NFDI4Culture (Altenhöner et al. 2020) umschrieben, das es sich zur Aufgabe macht, den Archivierungs- und Publikationsprozess als forschungsgeleitetes Zusammenwirken der verschiedenen Akteure zu verstehen und pragmatische Ansätze für den Aufbau einer in den unterschiedlichsten Dimensionen kohärenten (und nicht nur technisch verstandenen) Infrastruktur zu liefern.

Ablauf des Panels

Die drei Beitragenden werden zunächst in je 10minütigen Impulsvorträgen anhand konkreter Fallbeispiele für 'interessante Ausnahmen' – im Sinne von: Datenkorpora, die gewinnbringend in eine Langzeitarchivierung gebracht werden könnten und sollten, obwohl sie bei den normalen Routinen öffentlicher Archivierung momentan wohl unberücksichtigt blieben – ihre Perspektive auf die übergreifende Thematik darstellen. Im Kern des Panels steht ausgehend von diesen Impulsvorträgen die offene Plenumsdiskussion mit den Panelisten. Die Moderation strukturiert hierbei die Diskussion entlang von Leitfragen (s.u.). Parallel kümmert sich ein Team zugleich um die Dokumentation der Diskussionsbeiträge, insbesondere auch der in der Plenumsdiskussion von den Teilnehmenden eingebrachten weiteren Praxisbeispiele. Die Dokumentation wird in Form eines ausführlichen, zusammenfassenden Berichts im Nachgang den Teilnehmenden zur Verfügung gestellt und im Open Access in kommentierbarer Form öffentlich gemacht. Die Veranstaltung möchte auf diese Weise einen kontinuierlichen Diskussionsprozess der an der Thematik beteiligten Fachcommunities anregen und bestmöglich unterstützen.

Die Leitfragenbündel für die Paneldiskussion lauten:

- Welche Objekte erhalten wir? Wie bewerten wir Daten?
- Wie können inhaltlich die aktuellen wie künftigen Interessen der forschenden Communities Berücksichtigung finden? Wer entscheidet?
- Welche Verwendungszusammenhänge entstandener oder gewonnener Daten sind im eigenen Forschungskontext bzw. im Kontext anderer Projekte absehbar?
- Was bedeutet das für die Ausgestaltung von Förderlinien? Brauchen wir neben der Bewilligung von Projekten (und der Datenmanagementplan gesicherten Datenpublikation)

auch die zyklische 'Wiederbehandlung' von Archiven, um sie nutzbar zu halten?

- Welche Rolle spielen bestimmte Medialitäten und Modalitäten von Daten? Welche Rolle spielen disziplinspezifische bzw. disziplinenübergreifende und -unabhängige Perspektiven?
- Was bedeutet das für die NFDI4Culture und andere geistes- und kulturwissenschaftliche Konsortien? Wie können die Konsortien hier optimal zusammenwirken?

Panelisten

Margarete **Pratschke**, Berlin: Digitale Bildkultur im Netz (Arbeitstitel)

Patrick **Primavesi**, Leipzig: Tanz - Theater - Performance. Digitale Repräsentation performativer Künste (Arbeitstitel)

Christoph **Schulz**, Wuppertal: Vergleich verschiedener Modi der digitalen Präsentation von Bewegungsbüchern mit Perspektiven auf die *gaming arts* (Arbeitstitel)

Moderation

Reinhard **Altenhöner**, Berlin

Lisa **Dieckmann**, Köln

Andreas **Münzmay**, Paderborn

Bibliographie

Altenhöner, Reinhard et al. (2020): "NFDI4Culture – Consortium for research data on material and immaterial cultural heritage", in: *Research Ideas and Outcomes* 6 (Juli 2020), e57036. <https://doi.org/10.3897/rio.6.e57036>

Keitel, Christian (2013): "Der Nestor-Leitfaden zur Digitalen Bestandserhaltung und seine Folgen für die Archive", in: Naumann, Kai (ed.): *Digitale Archivierung in der Praxis*. 16. Tagung des Arbeitskreises "Archivierung von Unterlagen aus digitalen Systemen" und Nestor-Workshop "Koordinierungsstellen", Stuttgart 2013, S. 267–277.

Primavesi, Patrick et al. (2016): "Archiv/Praxis. Verkörpertes Wissen in Bewegung", in: Cairo, Milena / Hannemann, Moritz / Haß, Ulrike / Schäfer, Judith Schäfer (eds.): *Episteme des Theaters*. Bielefeld 2016, S. 425–450. <https://doi.org/10.14361/9783839436035-030>

Reference Model For An Open Archival Information System (OAIS) (2012). CCSDS Secretariat. Juni 2012. Zugleich ISO 14721:2012. <https://public.cds.org/Pubs/650x0m2.pdf>

Schulz, Christoph Benjamin (2019): "Die Geschichte(n) gefalteter Bücher. Gefaltete Texte und Leporellos in literarischer Avantgarde und experimenteller Poesie", in: Schulz, Christoph Benjamin (ed.): *Die Geschichte(n) gefalteter Bücher. Leporellos, Livres-Accordéon und Folded Panoramas in Literatur und Bildender Kunst*. Hildesheim / Zürich / New York 2019, S. 11–121 und S. 437–486.

Schulz, Christoph Benjamin (2021a): "Bookishness and Digitally Enhanced Publications Against the Backdrop of Apologies of the Book During the Advent of Digital Media", in: Bazarnik, Katarzyna / Hildebrand-Schat, Viola / Schulz, Christoph Benjamin (eds.): *Refresh the Book. On the Hybrid Nature of the Book*

in the Age of Electronic Publishing (Critical Studies 41). Leiden 2021, S. 133–161. https://doi.org/10.1163/9789004443556_008

Schulz, Christoph Benjamin (2021b): "Von einem Phänomen der Buchgeschichte zur Herausforderung für die Digital Humanities: Leporellos in der Kinder- und Jugendliteratur", in: Schmideler, Sebastian / Helm, Wiebke (eds.): *BildWissen – KinderBuch*. Studien zu Kinder- und Jugendliteratur und -medien, Bd. 5, . Stuttgart 2021, S. 23–44. https://doi.org/10.1007/978-3-476-05758-7_3

Zhou, Peter (2021): "Towards a Sustainable Infrastructure for the Preservation of Cultural Heritage and Digital Scholarship", in: *Data and Information Management* 5, no.2 (2021), S. 253–261. <https://doi.org/10.2478/dim-2020-0052>

Offen für alle(s)? Open Identities im Reviewprozess der DHd-Konferenz

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de
Universität Leipzig

Czmiel, Alexander

czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Dieckmann, Lisa

lisa.dieckmann@uni-koeln.de
Universität zu Köln

Guhr, Svenja

svenja.guhr@tu-darmstadt.de
Technische Universität Darmstadt

Jacke, Janina

janina.jacke@uni-goettingen.de
Technische Universität Darmstadt

Reiter, Nils

nils.reiter@uni-koeln.de
Universität zu Köln

Scholger, Walter

walter.scholger@uni-graz.at
Universität Graz

Wuttke, Ulrike

ulrike.wuttke@fh-potsdam.de
Fachhochschule Potsdam

Einleitung

Die Begutachtung von Forschungsbeiträgen ist ein zentraler Pfeiler wissenschaftlicher Qualitätssicherung, sei es für Zeitschriften, Konferenzen oder drittmittelfinanzierte Forschungsprojekte. Dafür, *wie* diese Begutachtung konkret abläuft, gibt es unterschiedliche Modelle, Gepflogenheiten, Erfahrungen und Erwartungen.

Das bei DHd-Konferenzen bis inklusive 2020 verwendete Modell sah eine Teilanonymisierung vor, d.h. Autor:innen waren den Gutachter:innen namentlich bekannt, jedoch nicht umgekehrt (sog. *single-blind*-Modell). Die Gutachten selbst (Text und zahlenmäßige Bewertung) wurden nur den Autor:innen, allen Gutachter:innen des Beitrags und dem Programmkomitee mitgeteilt. Dieses Modell war Gegenstand von Diskussionen auf den DHd-Mitgliederversammlungen 2019 und 2020, bis 2020 beschlossen wurde, für die nächste DHd-Konferenz ein *zero-blind*-Modell zu erproben – mit der Einschränkung, dass die Namen der Gutachter:innen nur den Autor:innen des Beitrags und dem Programmkomitee bekannt sind, aber nicht insgesamt veröffentlicht werden.

Wir möchten mit diesem Panel der DHd-Community die Möglichkeit geben, sich über das Begutachtungsverfahren der DHd-Jahrestagungen zu informieren und auszutauschen. Zum einen ist das Zeitkorsett einer Mitgliederversammlung eng begrenzt, was eine wirkliche Diskussion schwierig macht. Zum anderen betrifft das Begutachtungsverfahren auch zahlreiche Nicht-Mitglieder.

Begriffe und Problemstellung

Für eine zielführende und produktive Diskussion sollen auf dem Panel zunächst einige zentrale Begriffe geklärt werden, die in der Diskussion häufig vorkommen. Unter dem Begriff "open peer review" wird ein Bündel unterschiedlicher konkreter Verfahren zusammengefasst, die unterschiedliche Aspekte des Reviewprozesses öffnen:

Open Reports

Mit "open reports" oder "open reviews" ist gemeint, dass die zu einem Beitrag geschriebenen Gutachten öffentlich einsehbar sind, also über den Kreis der Autor:innen des Beitrages hinaus. Ob diese als eigenständige Publikationen adressier- und zitierbar sind oder dem Beitrag quasi als eine Art Anhang beigelegt werden, ist nicht festgelegt, wirft aber neue Fragen auf: Können die Gutachten als eigenständige Publikation anonym bleiben? Können sie dem Beitrag als Anhang angefügt werden, sodass der Beitrag immer mit den Gutachten gelesen werden wird? Und, im konkreten Verfahren: Wenn die Beiträge nach Kenntnisnahme der Gutachten von den Autor:innen überarbeitet werden, müssen dann auch die Gutachten nochmal überarbeitet werden?

Trotz dieser Fragen liegen Vorteile offener Reviews klar auf der Hand: Spätere Leser:innen der begutachteten Beiträge können kritische oder umstrittene Punkte direkt identifizieren, und nicht zuletzt können offene Reviews auch als Beispiele für Erst-Gutachter:innen dienen, wie Reviews aussehen (können).

Open Identities

Bei "open identities" handelt es sich um den am kontroversesten diskutierten Aspekt. Grundsätzlich können analog zur Verwen-

dungsweise in den experimentellen Wissenschaften verschiedene Stufen von "blindness" unterschieden werden: Beim **double-blind-Verfahren** (Abb. 1) sind sowohl Autor:innen als auch Gutachter:innen gegenseitig anonym. Aus praktischen Gründen (etwa um Interessenkonflikte bei der Zuordnung zu verhindern) muss eine Instanz existieren, die Autor:innen und Gutachter:innen namentlich kennt (das Programm- oder Organisationskomitee, bei großen Konferenzen auch sog. *area chairs*, die für einen thematischen, abgegrenzten Bereich zuständig sind). Das *double-blind*-Verfahren wird z.B. bei den Konferenzen der *Association for Computational Linguistics* (ACL) verwendet, wobei die Bekanntheit zwischen den Gutachter:innen unterschiedlich gehandhabt wird.

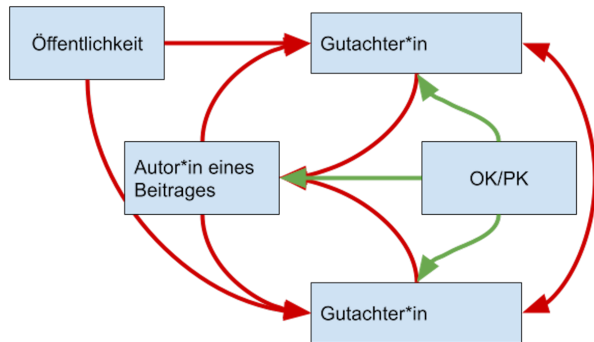


Abb. 1: *Double-blind*-Verfahren. OK/PK: Organisations- bzw. Programmkomitee. Roter Pfeil: Name nicht bekannt, grüner Pfeil: Name bekannt. Die Abbildung bezieht sich auf jeweils einen Beitrag, d.h. der Pfeil zwischen den Gutachter:innen repräsentiert die Bekanntheit zwischen den Gutachter:innen, die für den gleichen Beitrag zur Begutachtung eingeteilt wurden.

Ein sog. **single-blind-Verfahren** wurde seit Anbeginn für die DHd-Konferenzen angewendet. Dabei kennen die Gutachter:innen eines Beitrages dessen Autor:innen, aber nicht umgekehrt (Abb. 2).¹

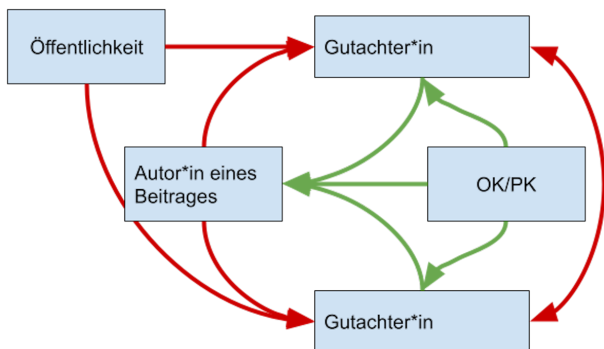


Abb. 2: *Single-blind*-Verfahren, das für die DHd-Konferenzen bis zur DHd 2020 angewendet wurde.

Im **zero-blind-Verfahren** sind sich sowohl Gutachter:innen als auch Autor:innen gegenseitig bekannt (Abb. 3). Dieses Verfahren wird bei der DHd2022 erstmals und testweise angewendet. Unterschiedlich gehandhabt wird, für wen genau die Gutachter:innen namentlich bekannt sind. Bei der DHd2022 werden die Gutachter:innen namentlich nur den jeweiligen Autor:innen (und dem Programmkomitee) bekannt gemacht.

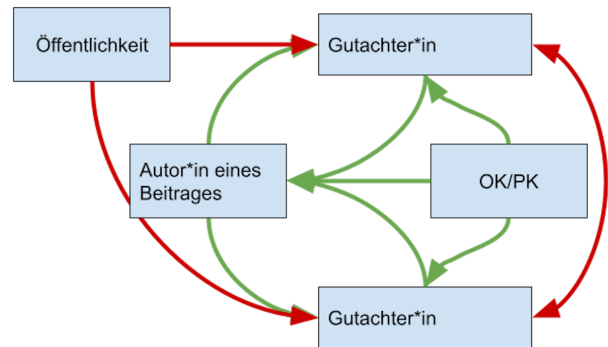


Abb. 3: DHd-Reviewing für die DHd 2022.

Open Participation

Bei den meisten Begutachtungsverfahren ist eine Instanz vorgesehen, die Gutachter:innen Beiträge zuweist, basierend auf Kompetenz, Interesse und/oder der Vermeidung von Interessenkonflikten. Bei "open participation" entscheiden Gutachter:innen selbst, ob sie einen Beitrag begutachten oder nicht, wobei die Kenntnis über mögliche Beiträge vorausgesetzt wird (d.h. die Beiträge müssen den potentiellen Gutachter:innen zugänglich sein). In der Praxis zieht dieses Verfahren weitere Fragen nach sich: Wie wird sichergestellt, dass alle Beiträge begutachtet werden und auch gleich viele Gutachten erhalten? Was ist die Motivation von Gutachter:innen, Beiträge zu begutachten, und wird dadurch eine (neue) Schieflage eingeführt?

Wir werden uns im Folgenden auf den Aspekt der *identities* und hier auf die Optionen *single*- und *zero-blind* konzentrieren, da sie für die DHd-Konferenzen im Mittelpunkt der Diskussion stehen, und typischerweise die größten Kontroversen auslösen (vgl. Ross-Hellauer/Görög, 2019). Im Panel allerdings können auch andere Aspekte im Verlauf der Diskussion aufgegriffen werden.

Erfahrungen

In verschiedenen Disziplinen wurde und wird mit unterschiedlichen Begutachtungsverfahren experimentiert. Die Erfahrungen dabei sind unterschiedlich, und oft wird darüber vor allem anekdotisch oder in semi-öffentlichen Kreisen berichtet (wie z.B. von der DH2020). Ein Grund dafür ist sicher, dass es schwer ist, belastbare, vergleichende Studien zum Thema durchzuführen, da der Review-Prozess als sensibel gilt und eine große Zahl an Faktoren die Ergebnisse beeinflussen. Ein solches Experiment – das sich nicht mit open vs. blind beschäftigt hat, sondern mit der Konsistenz von Begutachtung im Allgemeinen – ist das sog. NeurIPS-Experiment (Cortes/Lawrence 2021): Dabei wurden 10% oder 170 der Einreichungen bei der *machine-learning*-Konferenz NeurIPS zwei unabhängigen Programmkomitees zugewiesen, die selbstständig ihre Entscheidung getroffen haben (angenommen wurden Beiträge, wenn sie von mind. einem Komitee angenommen wurden). Bei ca. einem Viertel der Beiträge kamen die Komitees zu unterschiedlichen Entscheidungen. Auch wenn dieses Experiment sich nicht mit open vs. blind beschäftigt, zeigt es vielleicht einen Weg auf, wie generell eine Qualitätssicherung des Reviewprozesses aussehen könnte. Zu Bedenken ist allerdings, dass eine Übertragung von Verfahren aus anderen Disziplinen schon deswegen schwierig ist, weil diese sich in Größe,

Diversität, Kompetitivität und Publikationspraxis massiv unterscheiden.

Argumente für die Bekanntmachung von Gutachter:innen (pro Open Identities)

Eine einseitige Anonymität der Gutachter:innen, wie sie bislang bei den DHd-Jahrestagungen praktiziert wurde, bringt Nachteile für die Autor:innen mit sich: Ein in den Mitgliederversammlungen 2019 und 2020 benannter Kritikpunkt ist die Abgabe von "Einsatz-Gutachten" ohne ernsthafte Auseinandersetzung mit den eingereichten Beiträgen, geschweige denn konstruktiven Verbesserungsvorschlägen für die Autor:innen. Im Gegensatz dazu steigen Umfang und Konstruktivität der Gutachten in offenen Begutachtungsprozessen nachweislich, da die Verbindlichkeit und Rechenschaftspflicht der Bewertungen von Gutachter:innen durch die Offenlegung ihrer Identität deutlich zunimmt (Besançon/Rönnberg/Löwgren 2020:7 und Pucker/Schilbert/Schumacher 2019:3). Eine mildere Bewertung der Beiträge ging damit bei der DH2020 zumindest nicht einher (Guiliano/Estill 2020:13, interner Bericht). Des weiteren konnte auch kein Rückgang der Bereitschaft zur Begutachtung bei der DH2020 festgestellt werden: Nur fünf der vorjährigen Gutachter:innen verweigerten die Begutachtung mit Hinweis auf das offene Verfahren, während die Gesamtzahl der Gutachter:innen von 779 auf 974 anstieg.

Befürworter:innen eines anonymen Begutachtungsverfahrens argumentieren häufig damit, dass sich Autor:innen durch die Offenlegung ihrer Forschung sowie die Reviewer:innen durch die Formulierung ihrer Reviews angreifbar machen und eine Anonymisierung des Reviewverfahrens als Schutzmechanismus fungiere. Gerade in einer verhältnismäßig überschaubaren Fachgemeinschaft wie den Digital Humanities kann diese Anonymität jedoch nur geringfügig sichergestellt werden, wodurch eine tatsächliche Anonymität der Autor:innen fragwürdig ist oder nur durch umständlichen Anonymisierungsaufwand in den Beiträgen gewährleistet werden.

Gerade bei einer fachlich weit gestreuten Konferenz wie der DHd-Reihe – und einem ebenso breit gefächerten Pool an Gutachtenden – dürfte die Offenlegung der Identitäten und damit der fachlichen Expertisen dazu beitragen, dass Gutachter:innen nur Beiträge begutachten, zu denen eine fachliche Nähe gegeben ist (Pucker/Schilbert/Schumacher 2019:3), was nicht nur konstruktive Rückmeldungen begünstigt, sondern auch negative Gutachten aufgrund fachlicher Unkenntnis verhindert.

Nicht zuletzt kann die offene Kommunikation zwischen Autor:innen und Gutachter:innen nicht nur zu einer Qualitätssteigerung des eingereichten Beitrages und zu weiterführenden Diskussionen führen, sondern auch Impulse für zukünftige Forschungs- und Publikationsvorhaben sowie Kollaborationen geben (Besançon/Rönnberg/Löwgren 2020: 5). Offene Identitäten im Reviewprozess ermöglichen somit generell eine transparentere Gestaltung, die eine bessere Nachvollziehbarkeit der Annahmen und Ablehnungen von Beiträgen zur DHd-Jahrestagung gewährleistet.

Argumente für die Anonymität von Gutachter:innen (Contra Open Identities)

Mit der Ablehnung von Beiträgen müssen naturgemäß auch schlechte Nachrichten überbracht werden. Autor:innen wiederum reagieren unterschiedlich auf ablehnende Bewertungen, was von den Gutachter:innen schlecht oder gar nicht vorgesehen werden kann. 'Blindness' erlaubt es den Gutachter:innen, ihre Gutachten unabhängig von solchen Erwägungen zu schreiben. Prekär beschäftigte Nachwuchswissenschaftler:innen müssen keine Angst haben, Beiträge etablierter Wissenschaftler:innen negativ zu beurteilen, selbst wenn diese ggf. für das berufliche Fortkommen (z.B. für Forschungsanträge oder Empfehlungsschreiben) relevant sind. Es ist, damit dieser Effekt eintritt, auch nicht entscheidend, ob die etablierten Wissenschaftler:innen in solchen Positionen sind, oder ihre Macht für diese Art von 'Racheaktionen' ausnutzen würden: Die 'Beißhemmung' tritt ja ein, wenn Gutachter:innen negative Reaktionen nur befürchten, unabhängig davon, ob sie eintreten. Ein möglicher langfristiger Effekt ist die Verkleinerung des Pools an Gutachter:innen, sowie dessen fortschreitende Ent-Diversifikation, da z.B. Professuren nach wie vor überdurchschnittlich häufig von Männern besetzt werden.

Selbst wenn Indizien dafür sprechen, dass offene Gutachter:innen-Namen zu besseren Reviews führen, besteht die Gefahr, dass Gutachten zu oberflächlich bleiben, weil sich Gutachter:innen nicht mehr trauen, subjektive und ggf. spekulative Aspekte anzumerken, welche die Qualität von Beiträgen erheblich steigern können. Insgesamt ist vermutlich eine starke Tendenz zu positiven Reviews zu erwarten, was die Unterscheidbarkeit der Qualität der Beiträge deutlich erschweren würde. Dieser Effekt wurde bereits anhand eines kontrollierten Experiments des British Journal of Psychiatry empirisch untermauert (Walsh et al., 2018): "Reviewers who signed were more likely to recommend publication." Wenn mehr Beiträge positiv (und damit ähnlich) bewertet werden, müssen mehr Entscheidungen durch das Programmkomitee getroffen werden. Dies steigert nicht nur den Arbeitsaufwand für wenige Personen, sondern ist auch für Autor:innen weniger transparent.

Analyse der DHd2022-Erfahrungen

Nach Abschluss des Review-Prozesses werden die Reviews vergleichend analysiert, um auf die Argumente pro und contra *open identities* unterstützend/entkräftend reagieren zu können. Die Reviews sowie Bewertungen der vergangenen DHd in Köln (2018) und Paderborn (2019) stehen als Vergleichswerte zur Verfügung. Die Bewertungen können direkt quantitativ verglichen werden. Bei den Reviews kommen Faktoren wie die Länge, genannte Referenzen, oder auch *sentiment*-Scores in Betracht (die aber mit Vorsicht zu interpretieren sind). Darüber hinaus soll ein Interview mit der Ombudsstelle in die Analyse einfließen.

Besetzung und Ablauf des Panels

Die Panelist:innen werden mit einleitenden Impulsvorträgen aus ihrer Expertise und Erfahrungen die Ausgangslage und Problemfelder skizzieren und nach einem Austausch innerhalb des Panels in eine offene Diskussion mit dem Plenum übergehen.

Der Ablauf des Panels ist wie folgt geplant (Gesamtdauer 90 Minuten):

- Vorbereitete Beiträge (zusammen ca. 30 Minuten)
 - Einleitung und Begriffsklärung (Alexander Czmiel)
 - Pro Bekanntgabe der Gutachter:innen (Svenja Guhr, Walter Scholger)
 - Pro Anonymität der Gutachter:innen (Janina Jacke, Nils Reiter)
 - Analyse der DHd2022-Erfahrungen (Lisa Dieckmann)
- Diskussion innerhalb des Panels (30 Minuten)
- Plenardiskussion (30 Minuten)

Fußnoten

1. Theoretisch denkbar wäre auch eine Anonymität in umgekehrte Richtung, also bekannte Gutachter:innen und anonyme Autor:innen. Dieses Verfahren wird von Pontille/Torny (2014) als "Blind review" bezeichnet.

Bibliographie

Besançon, Lonni / Rönnerberg, Niklas / Löwgren, Jonas et al. (2020): "Open up: a survey on open and non-anonymized peer reviewing", in: *Res Integr Peer Rev* 5(8) 10.1186/s41073-020-00094-z

Burghardt, Manuel / Dieckmann, Lisa / Reiter, Nils / Steyer, Timo / Scholger, Walter / Trilcke, Peer / Wuttke, Ulrike (2021): *Besseres Reviewing für die DHd (Version 1.0)*. Zenodo 10.5281/zenodo.4633633

Cortes, Corinna / Lawrence, Neil D. (2021): Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. arXiv:2109.09774v1. <https://arxiv.org/abs/2109.09774>

Guiliano, J. / Estill, L. (2020): DH2020 Report to ADHO CCC Nature America Inc. (1999): "Pros and cons of open peer review", in: *Nat Neurosci* 2: 197–198 10.1038/6295

Pontille, David / Torny, Didier (2014): "The blind shall see! The question of anonymity in journal peer review", in: *Ada: A Journal of Gender, New Media, and Technology* 4 10.7264/N3542KVV

Pucker, Boas / Schilbert, Hanna Marie / Schumacher, Sina Franziska (2019): "Integrating Molecular Biology and Bioinformatics Education", in: *Journal of Integrative Bioinformatics* 16(3) 10.1515/jib-2019-0005

Ross-Hellauer, Tony / Görögh, Edit (2019): "Guidelines for open peer review implementation", in: *Res Integr Peer Rev* 4(4) 10.1186/s41073-019-0063-9

Ross-Hellauer, Tony (2017): "What is open peer review? A systematic review", in: *F1000Research* 6: 588 10.12688/f1000research.11369.2

Walsh, E. / Rooney, M. / Appleby, L. / Wilkinson, G. (2000): "Open peer review: A randomised controlled trial", in: *British Journal of Psychiatry*, 176(1): 47–51 10.1192/bjp.176.1.47

Protokolle

Modellierung einer administrativen Textsorte

Arndt, Nadine

nadine.arndt@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften

Baddack, Cornelia

c.baddack@bundesarchiv.de

Bundesarchiv

Fischer-Nebmaier, Wladimir

wladimir.fischer-nebmaier@oeaw.ac.at

Österreichische Akademie der Wissenschaften, Austria

Gleixner, Sebastian

s.gleixner@bundesarchiv.de

Bundesarchiv

von Hindenburg, Barbara

hindenburg@kgparl.de

Kommission für Geschichte des Parlamentarismus und der politischen Parteien e. V.

Jüngerkes, Sven

juengerkes@kgparl.de

Kommission für Geschichte des Parlamentarismus und der politischen Parteien e. V.

Kurz, Stephan

stephan.kurz@oeaw.ac.at

Österreichische Akademie der Wissenschaften, Austria

Schrott, Maximilian

maximilian.schrott@ndb.badw.de

Historische Kommission bei der Bayerischen Akademie der Wissenschaften

Protokolle, Protokolledition, analog/digital

Administrative Textsorten, darunter prominent das Protokoll, stehen im Zentrum jeder Bürokratie, sie sorgen für das Funktionieren von Wissenstransfer innerhalb von Institutionen über Zeit und zwischen Akteuren. Was sie festhalten, ist relevant: Sie sind Quellen für die Forschung, sie sorgen in unterschiedlichen Verhältnissen aber auch für Organisationsgedächtnis, Verbindlichkeit und Transparenz innerhalb und zwischen Institutionen, Verwaltungseinheiten und der Gesellschaft. Das Panel fokussiert die Verantwortung der Überlieferung von Protokollen und fragt, wie

Daten und Applikationen verfasst sein müssen, um der Gedächtnisfunktion von historischen „open governmental data“ gerecht zu werden. Es geht dabei von einer relativen Stabilität der Formen protokollarischen administrativen Handelns aus, die weitgehend unabhängig davon ist, ob das betreffende Protokoll z.B. staatliche, privatwirtschaftliche, akademische, glaubensgemeinschaftliche oder sonstige Gremien betrifft. An der Konzeption des Panels sind einige große Editionsprojekte aus dem deutschsprachigen Raum beteiligt, die bereits mit digital veröffentlichten Protokollkorpora hervorgetreten sind; die Anwendbarkeit der im Panel erarbeiteten Ergebnisse auf andere protokollführende Körperschaften ist Gegenstand der geplanten Diskussion.

Der Textsorte Protokoll wurde zuletzt auch vonseiten der Text- und Medienwissenschaften mit je unterschiedlicher Schlagseite auf literaturwissenschaftlichen, textpragmatischen und geschichtswissenschaftlichen Fragestellungen Aufmerksamkeit zugewendet, die zumindest grobe Umriss einer Bestimmung des Protokolls als Textsorte administrativen oder bürokratischen sozialen Handelns ermöglicht (cf. Niehaus/Schmidt-Hannisa 2005, in Vorbereitung Plener/Werber/Wolf 2022).

Gemeinsam ist Protokollen (vom Beschluss- zum Verbatimprotokoll, vom Tonbandmitschnitt zu Sitzungsberichten) im Wesentlichen auch über Grenzen der sie verhandelnden Körperschaften hinweg, dass sie sich auf ein temporal bestimmtes Ereignis beziehen, an dem eine endliche Anzahl von Teilnehmer/innen beteiligt war, das im häufigsten Fall eine mehr oder weniger formalisierte Agenda mit einer endlichen Zahl an Tagesordnungspunkten aufweist und das sich üblicherweise in einem definierten institutionellen Zusammenhang (einer Behörde, einem Berufsverband, einer Körperschaft etc.) ereignet hat. Die Überlieferung der materiellen Fixierung durch bspw. Protokollführer/innen ist in der Regel innerhalb der Institution festgesetzt, ebenso wie ihre Ablage und archivarische Behandlung reglementiert sind.

Der Fokus des Panels liegt auf dem Abgleich der Herangehensweisen von bestehenden Editionsprojekten, die sich vorrangig mit Protokollen auseinandersetzen – sie haben, wie Vorgespräche zeigten, ähnliche Ausgangslagen bei der Bewältigung von „Altbausanierung“ (Neuber/Schafan/Kasper/Gödel/Stäcker 2020), aber auch bei der Erstellung von Druckvorlagen im Rahmen ihrer Umstellung in hybrides Single-source-Publishing.

Die meisten dieser Protokoll-editionsprojekte sind als Digitale Editionen angelegt und verwenden TEI-XML unterschiedlicher Auszeichnungstiefe. Überlappungen bestehen unter anderem mit den Bestrebungen, die Parlamentssitzungen verschiedener europäischer Staaten in einem TEI-Schema miteinander vergleichbar zu machen (Parla-CLARIN/teiParla, cf. Erjavec / Pančur, Andrej. (2019), mit den Interessen von Korpora aus Politik- und Rechtswissenschaften (vgl. etwa *saschagobel/legislatoR: Interface to the Comparative Legislators Database* oder <https://zenodo.org/communities/sean-fobbe-data/>), oder mit den Überlegungen zur Überarbeitung des TEI-Elements <event> (cf. Fritze/Klug/Kurz/Steindl, 2019 und Fritze/Kurz/Klug/Schlögl/Steindl 2020).

Beteiligte

Die Ministerratsprotokolle Österreichs und der österreichisch-ungarischen Monarchie 1848–1918 – Digitale Edition

Die digitale Edition umfasst zwei der vier Ministerräte der Habsburgermonarchie bzw. Österreich-Ungarns, die in der editorischen Verantwortung der Österreichischen Akademie der Wissenschaften liegen: Den österreichischen Ministerrat 1848–1867, dessen Protokolle in 28 Bänden abgeschlossen vorliegen und die in TEI retrodigitalisiert wurden, sowie den „cisleithanischen“ Ministerrat 1867–1918, wobei der erste Band dieser Serie in TEI vorliegt, während zum Zeitpunkt der Einreichung die Bände II, III/1 sowie VIII im seit 2018 erarbeiteten Editionsworkflow in hybrider Erscheinungsweise in der Endbearbeitung sind. Weiters sind unkorrigierte Volltexte zu jenen Bänden der Protokolle des Gemeinsamen Ministerrats 1867–1918 verfügbar, die von der Ungarischen Akademie der Wissenschaften ediert werden konnten. Zum Ungarischen Ministerrat 1867–1918 liegen Links zu Bilddigitalisaten und Listen der Sitzungen und Tagesordnungspunkte ebenfalls in der Webapplikation des Editionsprojekts unter <https://mrp.oew.ac.at> vor. Prosopographische und bibliographische Auxiliardaten zu verwendeter Literatur, identifizierten Personen, Institutionen und Orten sind ebenfalls verfügbar. Mehrere API-Endpoints ergänzen das Angebot. Das Editionsteam, das mit zwei Perspektiven am Panel vertreten sein wird, einer geschichtswissenschaftlich-editorischen (Fischer-Nebmaier) und einer eher technischen (Kurz), ist überzeugt, dass im Austausch mit anderen Protokoll-editionen Verbesserungen in der Datenmodellierung, in der editorischen Arbeit, aber auch an den Interfaces möglich ist.

Editionsprogramm Fraktionen im Deutschen Bundestag 1949–2005

Im Rahmen des vom Deutschen Bundestag geförderten Editionsprogramms erschließt die „Kommission für Geschichte des Parlamentarismus und der politischen Parteien“ (KGParl) seit 1993 in unregelmäßigen Abständen, seit 2013 systematisch einen einzigartigen Quellenbestand zur Geschichte der parlamentarischen Kultur und des Parlamentarismus in der Bundesrepublik für Forschung und Öffentlichkeit: Seit 1949 führen praktisch alle Fraktionen und größeren (Landes-)Gruppen des Bundestags Protokolle ihrer Sitzungen, die sukzessive für die Zeit von 1949 bis 2005 vollständig ediert und veröffentlicht werden. Seit Ende der 1960er Jahre treten zu den schriftlichen Protokollen unterschiedlichster Ausprägung (Wort-, Verlaufs- oder Ergebnisprotokoll bzw. Mischformen) Tonbandaufzeichnungen, die teilweise das schriftliche Protokoll als Basis des Organisationsgedächtnisses ergänzen oder ersetzen. Die zunächst klassisch in Buchform erschienene Edition wurde seit 2017 digitalisiert und ist in PDF(A)-Form mit Volltextsuche im Netz abrufbar. Seit Ende 2020 wird das Vorhaben in eine digitale Edition auf Basis von TEI-XML überführt, neuere Zeiträume ab 1976 erscheinen genuin digital und nur noch als Auswahl-edition im Print. Besonderes Augenmerk wurde bei der semantischen Auszeichnung auf die Identifizierung von Parlamentaria (Gesetzesvorhaben etc.) und v.a. von Personen (SprecherInnen und Erwähnte) gelegt, wobei auf Austauschbarkeit über Normdaten zunächst mit der GND und den MdB-Stammdaten des Open-data-Portals des Bundestags Wert gelegt wurde. (In Überlegung ist, in mittlerer Perspektive, soweit möglich, weitere Verknüpfungen z.B. mit *legislatoR* oder entstehenden (OAI-)Schnittstellen der Parteiarchive, die die Quellen für die Edition liefern, hinzuzufügen). So entsteht eine umfassende digitale Quellenbasis zu zentralen parlamentarischen Organisationseinheiten und damit zur Funktionsweise des politischen Systems der Bundesrepublik, die interdisziplinär nutzbar ist.

Die Protokolle des Bayerischen Ministerrats 1945–1962 Online

Anfang der 1990er Jahre begann die Historische Kommission bei der Bayerischen Akademie der Wissenschaften in Zusammenarbeit mit der Generaldirektion der Staatlichen Archive Bayerns mit der historisch-kritischen Aufbereitung der Sitzungsprotokolle des Bayerischen Ministerrats nach 1945. Bis 2017 erschienen acht gedruckte Bände, die Diskussionen, Kontroversen und Beschlüsse der Bayerischen Staatsregierung in der Zeit der Besatzung und der frühen Bundesrepublik dokumentieren (retrodigitalisiert unter <https://www.bayerischer-ministerrat.de>).

Während diese Bände auf klassische Weise erstellt wurden, wurde 2014 beschlossen, die Edition auf einen digitalen, TEI-XML-basierten Ansatz und die Veröffentlichung im Internet umzustellen. Dabei mussten unterschiedliche Ziele in Einklang gebracht werden. Einerseits sollte durch umfängliche Anreicherung des Texts mit Metadaten und Verknüpfungen möglichst großer Nutzen aus den neuen Möglichkeiten gezogen werden. Andererseits galt es Rücksicht auf bestehende Strukturen des Projekts zu nehmen und den bisherigen Workflow möglichst wenig zu verändern oder durch übermäßigen Zusatzaufwand auszubremsten. Auch sollen die neuen Bände in möglichst unveränderter Form weiterhin noch gedruckt erscheinen.

Als Ergebnis dieser Anforderungen entstand das Editionskonzept "Oxydation" (<https://www.bayerischer-ministerrat.de/oxydation>), das inzwischen auch bei anderen Projekten der Kommission eingesetzt wird. Im Sommer 2019 wurde der neunte Band der Edition im Druck erfolgreich fertiggestellt. Die Onlineveröffentlichung ist für 2022 geplant.

Mit den Protokollen des Bayerischen Staatsrats 1799–1817 betreibt die Historische Kommission eine zweite Protokolledition. Seit 2006 wurden vier Bände im Druck und online (<https://www.bayerischer-staatsrat.de>) veröffentlicht, der fünfte und letzte Band ist derzeit in Arbeit. Die TEI-XML-Fassung wird im Zuge der Drucklegung mit erzeugt. Sie umfasst lediglich typographische Auszeichnungen. Hinzu kommt eine Anreicherung mit Verknüpfungen zwischen Editionsdocumenten und Register, externe Verlinkungen auf häufig zitierte, online verfügbare Ressourcen und Metadaten.

Die Kabinettsprotokolle der Bundesregierung

Mit der Edition „Die Kabinettsprotokolle der Bundesregierung“ erfüllt das Bundesarchiv den per Kabinettsbeschluss erhaltenen Auftrag der Bundesregierung, die über 30 Jahre alten Sitzungsniederschriften der Beratungen des Kabinetts und seiner Ausschüsse in wissenschaftlich kommentierter Form zu veröffentlichen. Seit Erscheinen des ersten Bandes 1982 wurden 32 Bände, zuletzt der Jahresband 1974, publiziert. Eine seit 2003 bestehende Online-Edition präsentiert die Inhalte der gedruckten Bände jeweils 18 Monate nach Veröffentlichung. Darüber hinaus stellt sie seit 2013 die bereits schutzfristfreien, aber noch unkommentierten Jahrgänge der Kabinettsprotokolle textkritisch bearbeitet zur Verfügung.

Um das veraltete, proprietäre Verfahren der Online-Stellung durch ein XML-basiertes Redaktions- und Publikationssystem abzulösen, wird seit Mai 2019 in Kooperation mit der Berlin-Brandenburgischen Akademie der Wissenschaften die Arbeitsumgebung ediarum für die Edition der Kabinettsprotokolle weiterentwickelt, seit Juli 2020 im produktiven Betrieb. Ediarum ist eine von TELOTA an der BBAW seit 2012 entwickelte Lö-

sung, die es Wissenschaftler*innen erlaubt, Transkriptionen in TEI-XML zu bearbeiten, mit Apparaten sowie Registern zu versehen und zu veröffentlichen. Dabei setzt TELOTA auf existierende Softwarekomponenten von Dritten auf, um die Entwicklungsarbeit auf spezifische Bedürfnisse und Anforderungen der editions-wissenschaftlichen Fachcommunity zu fokussieren.

Der Dokumententyp „Protokoll“ ist im Kontext der ediarum-Entwicklung neuartig, sodass die ediarum-Basiskomponenten angepasst und erweitert werden. Es entsteht eine generisch konzipierte Editions-umgebung für (Sitzungs-)Protokolle, die perspektivisch die Integration weiterer Editionen des Bundesarchivs ermöglicht und von anderen Akteneditionen nachgenutzt werden kann. Eine Veröffentlichung von ediarum.MINUTES ist für 2021 geplant.

Themen und Herausforderungen

Die vertretenen Protokoll-editionsprojekte stehen vor ähnlichen Herausforderungen, etwa in den Bereichen

- Quellenkritik (Was ist ein verbranntes Protokoll? Was ist der Quellenstatus eines Spiegel-Artikels, der anstelle eines Protokolls abgedruckt wird?)
- Modellierung (Wie ist die Textsorte Protokoll in allen ihren Formen in Markup zu fassen? Was an Beilagen, welche in Bezug auf Textsorten und Medien divers sind, nehmen die Projekte wie auf?)
- Dateneingabe (HTR? Transkription von Stenographie? OCR? Abtippen?)
- Verarbeitung und Datenanreicherung (NER, Metadatenerfassung, Erkennung textueller Muster; Nutzung von Norm- und Auxiliardaten)
- der visuellen Präsentation, Webseite, Schnittstellen usw.

Primäres Ziel ist die Entwicklung gemeinsamer Best-practice-Mittel bei der Edition von protokollartigen Texten, insbesondere im Hinblick auf die Möglichkeiten von Mapping oder Harmonisierung der Datenmodelle der Projekte (etwa im Hinblick auf Metadatenerstellung oder APIs für Datenaustausch/Harvesting sowie Anbindung an Open Data-Portale). Daneben erwarten wir Hinweise zu wiederverwendbaren Werkzeugen, die sich in der Bearbeitung (nicht nur) dieser Textsorte bewährt haben.

Der Austausch zwischen den teilnehmenden Editionsprojekten soll öffentlich geführt werden, um anderen ähnlich gelagerten Projekten die Möglichkeit zu geben, sich an der Diskussion zu beteiligen.

Bibliographie

Erjavec, Tomaž, & Pančur, Andrej. (2019). *Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings*. Zenodo. <http://doi.org/10.5281/zenodo.3446164>

Fritze, Christiane / Helmut W. Klug / Stephan Kurz / Christoph Steindl. *Recreating history through events* (TEI 2019, Graz, 2019, <https://gams.uni-graz.at/o:tei2019.141>)

Galloway, Alexander: *Protocol. How Control Exists after Decentralization*. Cambridge, Mass: MIT Press 2004.

Fritze, Christiane / Stephan Kurz / Helmut W. Klug / Matthias Schlögl / Christoph Steindl. *Panel: Events: Modellierungen und Schnittstellen* (DHd 2020, Paderborn <https://doi.org/10.5281/zenodo.3666690>)

Neuber, Friederike / Thorsten Schaßan / Dominik Kasper / Martina Gödel / Thomas Stäcker. (2020). *Altbausanierung mit*

Niveau – die Digitalisierung gedruckter Editionen (DHD 2020, <http://doi.org/10.5281/zenodo.4621822>)

Niehaus, Michael; Hans-Walter Schmidt-Hannisa (2005). *Das Protokoll: Kulturelle Funktionen einer Textsorte*. Frankfurt: P. Lang.

Niehaus, Michael (2011). "Epochen des Protokolls". *Zeitschrift für Medien- und Kulturforschung*, hg. von Lorenz Engell und Bernhard Siegert, 2/2011: *Medien des Rechts*. Hamburg: Meiner, 141–156.

Siegert, Bernhard / Vogl, Joseph (2003, Hg.). *Europa. Kultur der Sekretäre*. Zürich und Berlin: diaphanes.

Vorträge

AdA Annotation Explorer

Ein Framework für zeitbasierte Linked Open Data-Annotationen zur Analyse audiovisueller Korpora

Agt-Rickauer, Henning

henning.agt@gmail.com

Hasso-Plattner-Institut, Universität Potsdam, Germany

Scherer, Thomas

scherer.thomas@fu-berlin.de

Freie Universität Berlin, Germany

Stratil, Jasper

jasper.stratil@fu-berlin.de

Freie Universität Berlin, Germany

Die Medienspezifität zeitbasierter performativer Künste, wie Musik, Theater, Tanz und auch Film, stellt Analyseansätze, die nach der ästhetischen Erfahrung fragen, vor zahlreiche Herausforderungen: Es geht in diesen Ansätzen nicht um die 'objektiv-messbaren' Eigenschaften der Artefakte und Aufführungen, sondern darum, die intersubjektive Erfahrungsdimension in den Blick zu nehmen – sie zielen auf das Denken und Fühlen, das im Filme-Sehen (Kappelhoff 2018) oder Musik-Hören entsteht. Ein tradiertes analytisches Vorgehen besteht in der Dekonstruktion solcher Erfahrungsprozesse in detaillierte Beschreibungen zeitlicher Prozesse – von Partituren über Tanznotationen bis hin zu mehrdimensionalen Einstellungsprotokollen. Solche Instrumente und Techniken rücken vermehrt in den Fokus der Digital-Humanities-Forschung, wobei die Verbindung ästhetischer Theorien und Techniken der Datengewinnung, -haltung, -verarbeitung und -aufbereitung eine der zentralen Herausforderungen des Feldes markiert (Arnold et al. 2019, Flückiger/Halter 2020, Freedman et al. 2019). Drei Schlüsselaspekte eines solchen Methodendesigns werden im Folgenden anhand der Begriffe der *Synchronizität*, *Prozessualität* und *Komparabilität* beleuchtet. Audiovisuelle Bilder werden erst dann analytisch handhabbar, wenn sie auf unterschiedlichen Beschreibungsebenen gleichzeitig erfasst werden, um dann deren Zusammenspiel in den Blick zu nehmen (*Synchronizität*). Dabei sind einzelne Beschreibungen immer kontextabhängig und nicht objektorientiert, sondern auf die zeitliche Entfaltung eines Ganzen bezogen (*Prozessualität*). Das Forschungsinteresse ist dabei selten auf ein einzelnes Werk begrenzt, sondern richtet sich an ein Genre, einen Themenkomplex oder einen audiovisuellen Diskurs (*Komparabilität*). Es besteht so die Anforderung nach komplexen, vielschichtigen Beschreibungen, die gleichzeitig jedoch auf einen größeren Korpus bezogen werden können. Dies übersteigt häufig die Kapazitäten einzelner Forscher*innen, jedoch ermöglichen systematisierte filmanalytische Annotationen gleichermaßen die Vergleichbarkeit der analytischen Beobachtungen mehrerer Annotator*innen, als auch die Integration (semi-)automatisch erzeugter Annotationen.

Einer solchen Korpusstudie widmete sich die Nachwuchsgruppe "Affektrhetoriken des Audiovisuellen" (2016–2021, Freie Universität Berlin/ Hasso-Plattner-Institut, kurz: "AdA"), die den audiovisuellen Diskurs zur globalen Finanzkrise (nach 2007) hinsichtlich wiederkehrender affektrhetorischer Muster untersuchte.

Das Projekt gründete auf theoretischen Annahmen zum Filme-Sehen, dem Verhältnis der Wahrnehmung von Zuschauenden und sich zeitlich entfaltenden audiovisuellen Bilder und knüpfte an die eMAEX-Methode an (Kappelhoff et al. 2011-2016), die auf die analytischen Grundprinzipien von Segmentierung und Qualifizierung zurückgreift, um Erfahrungsfigurationen zu rekonstruieren (Bakels et al. 2020a, 2020b).

Um eine feingliedrige und multidimensionale Beschreibung mit interpersoneller Annotation für einen größeren Korpus leisten zu können, galt es das filmanalytische Beschreibungsvokabular in der AdA-Filmontologie maschinenlesbar zu systematisieren (für eine ausführlichere Darstellung siehe Bakels et al. 2020c). Das resultierende systematische Vokabular orientiert sich ebenso an methodischen Standardbegriffen der Filmwissenschaft (z.B. Einstellungsgrößen oder Montagefiguren), wie an ansatzspezifischen Fachbegriffen der eMAEX-Analyse (z.B. 'Ausdrucksbewegung' oder 'Dynamiken des Bildraums'). Die AdA-Filmontologie umfasst 502 einzelne Annotationswerte, die 78 Annotationstypen zugeordnet sind, die wiederum auf 8 Beschreibungsebenen wie Akustik, Montage, Bildkomposition oder Kamera organisiert sind (siehe Abb. 2).

Mit einem auf Basis der AdA-Filmontologie generierten Template konnten in der Annotationssoftware *Advene* (Aubert/Prié 2005) semantisch strukturierte Linked Open Data-Annotationen erstellt werden. Insgesamt wurden mehr als 125.000 Annotationen manuell erstellt für 4 komplette Filme, 2 Webvideos, Beiträge aus 35 Nachrichtensendungen sowie einzelne Szenen aus 10 weiteren Filmen. Darüber hinaus wurden über 440.000 Annotationen für 76 Filme sowie 315 Nachrichtensendungen automatisch generiert. Um die affektrhetorischen Muster auf Grundlage dieser manuellen und automatischen Annotationen allerdings korpusübergreifend analysieren zu können, insbesondere hinsichtlich ihrem inhärenten multidimensionalen Zusammenspiel, d.h. auf unterschiedlichen Beschreibungsebenen, und kontextabhängig im zeitlichen Verlauf, war ein Framework zur Exploration des umfassenden Annotationsdatensatzes notwendig, für das verfügbare Annotations- und Visualisierungssoftware (wie *Advene*, *ELAN*, *Dive+* der *Clariah Media Suite*) nicht ausgelegt war. Die zeitgleich entwickelte *VIAN WebApp* (<https://www.vian.app/>) legt im Zusammenspiel mit der *VIAN*-Annotationssoftware wiederum den zentralen Untersuchungsfokus auf die Farbanalyse.

Entwicklung & Architektur

Der AdA Annotation Explorer¹ wurde entwickelt, um sämtliche im AdA-Projekt erstellten zeitbasierten Videoannotationen zu verwalten und über eine webbasierte Oberfläche für die Exploration und Analyse zugänglich zu machen. Hierbei stand einerseits im Fokus, alle im Projekt entstehenden Daten zur freien Nutzung und Weiterverwendung zu veröffentlichen und andererseits die Anforderungen aus der korpusübergreifenden Filmanalyse umzusetzen, um Annotationsdaten gezielt nach bestimmten Kriterien abzufragen, zu filtern und die Ergebnisse kontextabhängig (d.h. in ihrer Einbettung in szenische Kompositionen) zu visualisieren.

Der Annotation Explorer wurde über eine entkoppelte Client-Server-Architektur mit RESTful API (Masse 2011) realisiert (siehe Abb. 1), die auf offene Standards für den Datenaustausch und auf Verwendung von Open-Source-Komponenten setzt.

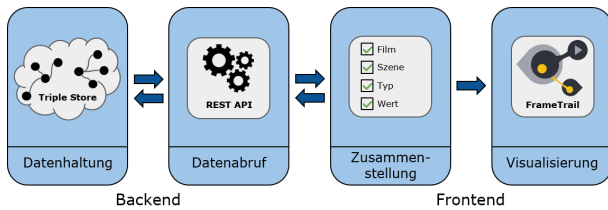


Abb. 1: Architektur des Annotation Explorers

Die Serverkomponenten im Backend sind für die Speicherung der Daten in einer Graphdatenbank (Triplestore) und für die Bereitstellung von Funktionen zum Datenabruf (REST-API) zuständig. Die Clientkomponenten des Frontends laufen ausschließlich im Browser und bilden die Benutzerschnittstelle, die eine Zusammenstellung von Annotationen mittels Eingabe von Auswahlkriterien (Filme, Szenen, Annotationswerte, etc.) ermöglicht und die Filterung und Weiterverarbeitung der Ergebnisse der REST-API übernimmt. Die aufbereiteten Daten werden mit der Open-Source-Software FrameTrail² visualisiert, die es erlaubt, interaktive Videos und damit verlinkte Inhalte direkt im Browser anzusehen.

Datenhaltung

Das Datenmanagement des Annotation Explorers folgt den Linked Open Data-Prinzipien³, um die in aufwändiger Arbeit erstellten Annotationen und Metadaten als maschinenlesbare, offene Daten für die Nachnutzung zu veröffentlichen. Sie werden mithilfe von Openlink Virtuoso⁴ gespeichert und mit LodView⁵ angezeigt. Der öffentliche SPARQL-Endpunkt⁶ erlaubt auch das Abfragen und Nutzen der Daten mit selbst entwickelten Queries. Die Nutzung von W3C Standards, die Bereitstellung von öffentlichen URIs für Annotationsdaten, Metadaten und Vokabular, sowie die dokumentierte Ontologie, ermöglichen die Nutzung der Projektergebnisse im Sinne der FAIR Data Prinzipien.

Zu den Daten des Projekts gehört die *AdA-Filmontologie*⁷, ein systematisches Vokabular auf Basis von OWL und RDF, welches filmanalytische Begriffe – geordnet nach Annotationsebenen, Annotationstypen und Annotationswerten (siehe Abb. 2) – für feingranulare semantische Videoannotationen definiert und per URIs erreichbar macht⁸.

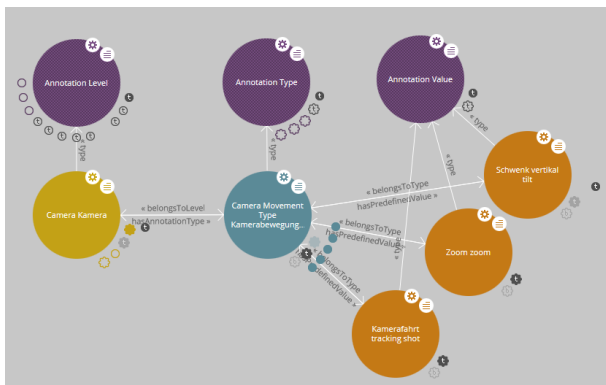
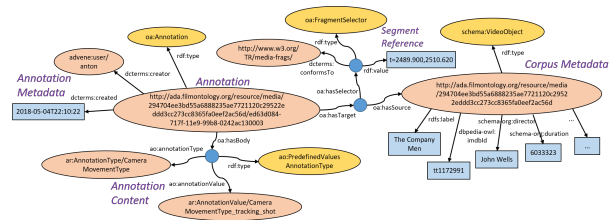


Abb. 2: Klassenstruktur der AdA-Filmontologie am Beispiel kameraspezifischer Konzepte

Die *Korpus-Metadaten* umfassen beschreibende Information (wie z.B. Titel, Laufzeit, Veröffentlichungsjahr, Regisseur) zu allen Filmen im Videokorpus zur globalen Finanzkrise. Die Metadatenfelder werden mit bestehenden Vokabular-Properties aus Dbpedia, schema.org und Dublin Core kodiert und sind ebenfalls per URI abrufbar⁹. Der Korpus umfasst 391 Filme in den Kategorien Spielfilme, Dokumentarfilme und Nachrichten¹⁰.

Semantische Videoannotationen sind der Hauptteil der Daten, die im Triplestore gespeichert werden. Sie erfassen die zeitbezogenen Anmerkungen zu den Filmen auf den unterschiedlichen Beschreibungsebenen unter Verwendung des systematischen Vokabulars. Das Annotationsmodell basiert auf dem W3C Web Annotation Data Model¹¹ und der Media Fragments URI Spezifikation¹² (siehe Abb. 3).

Abb. 3: Beispielannotation einer Kamerafahrt für den Film "Company Men"¹³ als RDF-Graph visualisiert.

Datenabruf

Die im Projekt entwickelte REST-API bildet die Schnittstelle zwischen dem User Interface und den im Triplestore gespeicherten Videoannotationen, Metadaten und Ontologiedaten. Sie bietet über einen Web-Dienst per HTTP eine Reihe von Funktionen an, die entsprechend der Anfrageparameter SPARQL-Datenbankabfragen generieren, ausführen und direkt im Browser weiterverarbeitbare JSON bzw. JSON-LD Formate zurückliefern. Sie ist in Java implementiert und verwendet die Frameworks Apache Jena¹⁴, Javalin¹⁵ und JSONLD-JAVA¹⁶. Neben der Bereitstellung der Ontologiedaten und der Verwaltung der Metadaten, ist die Hauptaufgabe der REST-API, die feingliedrige und multidimensionale Beschreibung der Filme korpusübergreifend anhand bestimmter Kriterien abrufbar und durchsuchbar zu machen.

Videoannotationen können pro Film, pro Szene und pro Annotationstyp abgerufen werden (siehe Abb. 4). Die REST-API ermöglicht außerdem eine Volltextsuche nach einem oder mehreren Schlüsselwörtern und liefert alle Trefferannotationen korpusübergreifend zurück (Beispiel: "crisis", nur ganze Worte, nur im Annotationstyp 'Dialogtext')¹⁷. Um Muster im Korpus zu finden, ist eine Value Search nach ein oder mehreren Annotationswerten möglich. Dabei werden die Stellen zurückgeliefert, bei denen Werte *zeitgleich* (auch teilweise überlappend) auftreten (z.B. Musikstimmung "traurig" zusammen mit nahen Einstellungsgrößen).

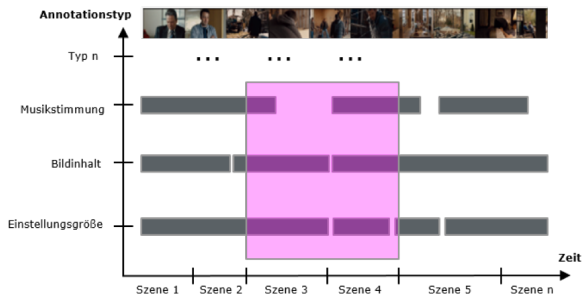


Abb. 4: Veranschaulichung des Datenabrufs für drei Annotationstypen aus zwei Szenen eines Films

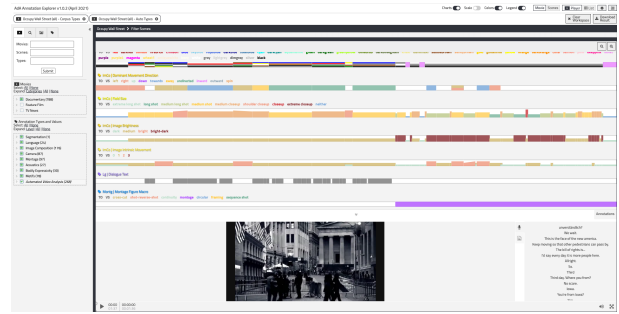


Abb. 6: Workspace des Annotation Explorers

Frontend (Zusammenstellung & Visualisierung)

Die *Zusammenstellungskomponente* im Annotation Explorer Web-Frontend erlaubt der Benutzer*in Anfragen an die Datenbasis zu stellen. Dafür werden Formularfelder zur Film-, Text-, Bilder- und Wertesuche (siehe Abb. 5) mit nutzerfreundlicher Auto-complete-Funktion angeboten, um ein schnelles Auswählen aus der großen Menge von Metadaten zu ermöglichen. Entsprechend der Nutzereingaben werden REST-API Aufrufe erzeugt, Annotationen geladen und für den Arbeitsbereich zusammengestellt.

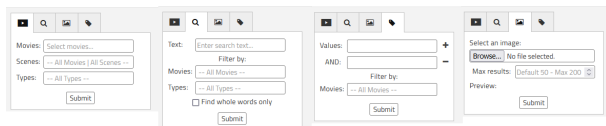


Abb. 5: Übersicht der vier Anfragemöglichkeiten im Annotation Explorer

Filmmetadaten und Ontologiedaten werden im Filterbereich in zwei Baumansichten (Facettensuche) dargestellt, geben Aufschluss darüber, wie viele Annotationen in den jeweiligen Kategorien gefunden wurden, und ermöglichen das weitere Einschränken des Suchergebnisses.

Die kontextabhängige *Visualisierung* der Annotationen erfolgt in einer Timeline-Darstellung zusammen mit einem Videoplayer, bei der angefragte Annotationstypen als Spuren dargestellt werden (siehe Abb. 6). Für diesen Zweck wird FrameTrail eingesetzt, das im Rahmen des Projekts durch den Autor Joscha Jäger hinsichtlich der Darstellung spezieller Annotationstypen, Sortiermöglichkeiten, Szenen- bzw. Filmvergleich und Exportmöglichkeiten erweitert wurde.

Anwendung & Analyse

Das graphische User-Interface des Annotation Explorers – der sogenannte Workspace – ist zentral für die Nutzer*inneninteraktion, das meint in erster Linie Film- und Medienwissenschaftler*innen ohne Programmierkenntnisse (siehe Abb. 6). Dabei kommen die drei zu Beginn genannten Schlüsselaspekte des Methodendesigns in der konkreten Nutzung des Annotation Explorers zum Tragen.

Prozessualität

Gegenstand der eMAEX-basierten Filmanalyse ist eine sich in der Wahrnehmung entfaltende Bewegung. In diesem Kontext ist nicht nur die enge Verknüpfung der Annotationen mit dem interaktiven Videoplayer herauszustellen, der die direkte Rückbindung der Analysebeobachtungen an das Bewegtbild ermöglicht und dabei deren Einbettung erlebbar macht, vielmehr sind die Einzelannotationen nicht als statische Eigenschaften oder isolierte Zeitpunkte von Interesse, sondern stets in der zeitlichen Entfaltung und in Bezug auf größere Zeitsegmente (etwa szenische Kompositionen).

Entscheidend ist dabei auch die Skalierbarkeit der Ansicht: so lassen sich sowohl einzelne Szenen als auch ganze Filme durch eine Szenenauswahlleiste oder freies Zoomen in den Blick nehmen.

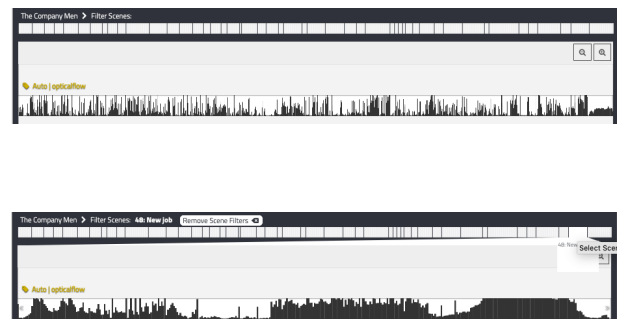


Abb. 7: Optischer Fluss-Makroprofil eines Films (oben) und Detailansicht einer Szene (unten)

Mit dem Wechsel zwischen Makro und Mikro in der Ansicht ganzer Filme können Schlüsselsequenzen (z.B. dramaturgische Umschlagpunkte) anhand Extrema einzelner Analysedimensionen (etwa extreme Nahaufnahmen, besonders (un-)bewegte Abschnitte, sehr leise Passagen oder schnelle Wortwechsel auf Ebene des Dialogs) identifiziert werden.

Der Annotation Explorer ist so gestaltet, dass nicht der einzelne analytische Befund in Annotationsform grafisch isoliert dargestellt wird, sondern stets als Teil eines komplexen Gefüges. So lassen sich durch unterschiedliche Darstellungsoptionen beispielsweise Einstellungsgrößen als farbiges Balkendiagramm anzeigen, bei der die Höhe des Balkens Auskunft über die Nähe der Kamera zum Referenzobjekt und die Breite Auskunft über die Dauer der einzelnen Einstellungen gibt (siehe Abb. 8). Auf diese Weise wird der Blick der Analysierenden auf den Prozess der Entfaltung im Hinblick auf spezifische Gestaltungsdimensionen gelenkt: Im Fall

von Einstellungsgrößen die Modulation von Nähe-Distanz-Verhältnissen.

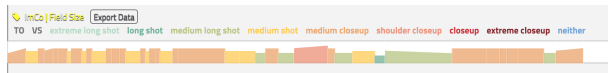


Abb. 8: Darstellung einer Abfolge von Einstellungsgrößen

Synchronizität

Die Expressivität audiovisueller Bilder ist nicht auf eine Gestaltungsdimension zu reduzieren, sondern erst der Blick auf das Zusammenspiel unterschiedlicher Dimensionen gibt Aufschluss über die Affektdramaturgie einer Sequenz. So lassen sich im Annotation Explorer unterschiedliche Kombinationen von Annotationstypen zusammenstellen und über die Baumansichten (Facettensuchen) detailliert anpassen (siehe Abb. 9).

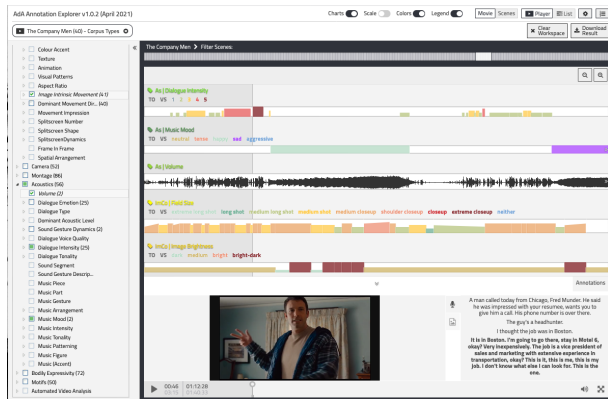


Abb. 9: Annotations-„Partitur“ im Annotation Explorer.

Es lässt sich beispielsweise untersuchen wie ein akustisches Phänomen (z.B. Einsatz und Taktung von Musik) mit einem optischen Phänomen (z.B. Schnittrhythmus) interagiert, dieses aufgreift, verstärkt oder konterkariert. Die FrameTrail-Implementation im Workspace bietet die Möglichkeit unterschiedliche Visualisierungsformen auf den so abgerufenen Annotationsdatensatz anzuwenden und mit wenigen Klicks unterschiedliche Annotations-„Partituren“ zu erstellen. Diese können entweder über die URL geteilt und gespeichert, oder als offline verfügbare HTML-Pakete heruntergeladen werden.

Komparabilität

Die Untersuchungsperspektive des AdA-Projekts und die Frage nach affektrhetorischen Mustern zielte aber nicht nur auf einzelne Filme und Sequenzen, sondern korpusübergreifende vergleichende Analysen. Dazu galt es die Möglichkeit bereitzustellen, Analyseabfragen verschiedener Filme direkt nebeneinander zu legen ohne an Funktionalität (interaktiver Videoplayer, Filter- und Suchoptionen, Darstellungsoptionen) einzubüßen. So lassen sich in der Playeransicht zwei Filme bzw. Szenen nebeneinander anzeigen, während weitere Vergleichsfilme/-szenen durch seitliches Scrollen ins Bild gebracht werden können (siehe Abb. 10).

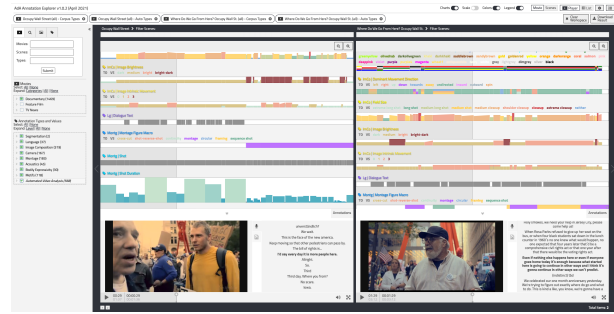


Abb. 10: Vergleich zweier Filme.

In einer Listenansicht können wiederum mehrere Filme und Szenen untereinander angezeigt werden (siehe Abb. 11).

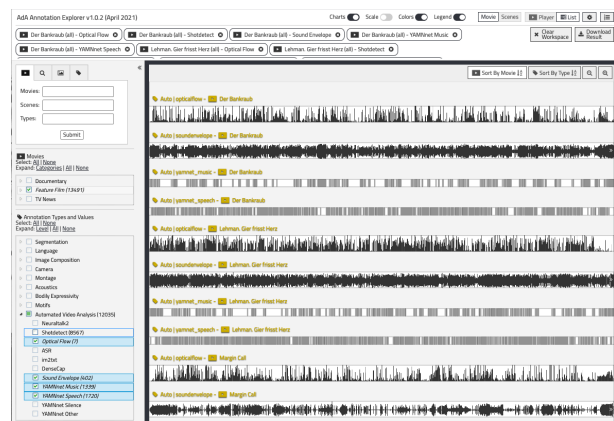


Abb. 11: Vergleich in der Listenansicht.

Zur Entwicklung und Überprüfung von Analysehypothesen gibt es unterschiedliche Suchfunktionen: etwa die Value Search oder die Volltextsuche, die für die Suche nach Schlagworten oder erwähnten Personen im Dialog (sei es auf Basis manueller Transkripte, redigierter Untertitel oder automatischer Spracherkennungsalgorithmen) genutzt werden kann oder auch für die Ergebnisse von automatischen Image Captioning-Extraktoren, das Aufspüren wiederkehrender oder ungewöhnlicher Bildelemente.

Eine weitere Möglichkeit ist die Reverse Image Search, bei der eine Bilddatei als Suchanfrage hochgeladen werden kann. Diese wird dann mit Keyframes jeder Einstellung des gesamten Korpus abgeglichen und die ähnlichsten Bilder werden mit Metadaten versehen ausgegeben (siehe Abb. 12).

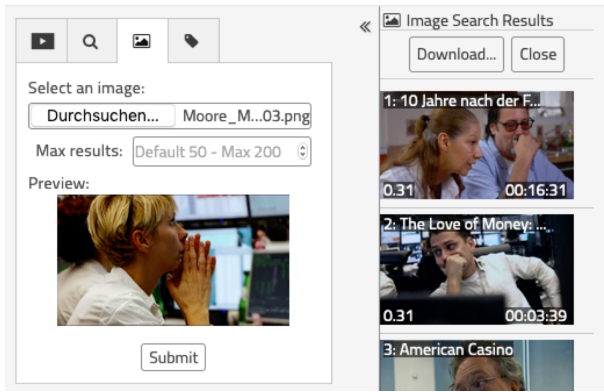


Abb. 12: Reverse Image Search im Annotation Explorer.

Die jeweiligen Anfrageergebnisse lassen sich wiederum im Framework des Annotation Explorers mit weiteren Annotationen kombinieren und können somit innerhalb sich zeitlich entfaltender Dynamiken, die auf dem Zusammenspiel unterschiedlicher Dimensionen beruhen, analysiert werden.¹⁸

Forschungsperspektiven

In den Pilotstudien zum audiovisuellen Diskurs zur globalen Finanzkrise (2007–) erwies sich der Annotation Explorer insbesondere im Hinblick auf die Korpusexploration und die Überprüfung von filmübergreifenden Forschungshypothesen als nützliches Werkzeug der Mustererkennung (im Sinne von Bakels et al. 2020a, 115f).¹⁹ Im Kontext der Frage nach der audiovisuellen Gestaltung und Motivatik gesellschaftlicher Krisenkommunikation in Nachrichten, Spiel- und Dokumentarfilmen bot das Framework neue Möglichkeiten der Exploration in unterschiedlichen Zusammenhängen. Werden einzelne Aspekte eines audiovisuellen Motivs durch die bereits implementierten (semi-)automatischen Erkenner erfasst, so ermöglichen diese das schnelle Auffinden weiterer Vorkommnisse. So zeigte eine Studie zum Stilmittel des „Brokerface“, dass Bildkompositionen von angespannten Gesichtern von Aktienhändlern im Zusammenspiel mit Monitoren einen wiederkehrendes Topos in der Verhandlung der Frage nach den realen Auswirkungen virtueller Marktentwicklungen spielt (vgl. Scherer & Stratil (in Ersch.)). Eine weitere Studie zum Motiv der plötzlich hereinbrechenden BREAKING NEWS als Kipp- und Höhepunkt von Krisendramaturgien veranschaulicht hingegen wie das Zusammenspiel von automatischer Schnitt-, Bildbewegungs- und Lautstärkeerkennung Rhythmusprofile hervorbringt, die Aufschluss über deren affektive Dimension geben (Stratil, in Ersch.). Ein solcher dynamischer ‚Fingerabdruck‘ einer Szene, der sich durch die (Darstellungs-)Prinzipien der Prozessualität, Synchronizität und Komparabilität in den Annotationsdaten als Muster zu erkennen gibt, ermöglicht es Aussagen zur Affektorientierung audiovisueller Sequenzen direkt am Material zu plausibilisieren und Charakteristika und Vergleichspunkte herauszuarbeiten. Die grundlegende Architektur des Annotation Explorers ermöglicht die flexible Kombination von (semi-)automatischen und manuellen Annotationen. Werden Kernkompositionsprinzipien allerdings nicht durch (semi-)automatische Erkenner tangiert, ist die Analyse auf die zeitaufwändigeren manuellen Annotationen angewiesen. Der Nutzen des Frameworks im Bereich der Korpus-Exploration verschiebt sich dann hin zu mikroanalytischen Detail- und Vergleichsstudien, anhand derer nachvollzo-

gen werden kann, wie kompositorische Prinzipien quer zu einzelnen Gestaltungsebenen als übergreifendes dynamisches Muster in Erscheinung treten. Es konnte so beispielsweise gezeigt werden, wie das Gefühl der Enttäuschung in einem Finanzkrisenfilm nicht als narrative Information zum Ausdruck kommt, sondern als filmische Ausdrucksbewegung (Bakels et al. 2020b).

Die Architektur des Annotation Explorers ermöglicht die Verknüpfung semantischer Daten mit einer geisteswissenschaftlichen Perspektive auf die Expressivität audiovisueller Medien und stellt so Fragen nach der Anwendbarkeit solcher Technologien in qualitativen Ansätzen – nicht als Alternative, sondern als ergänzendes Instrument zur Weitung und Systematisierung des analytischen Blicks.

Fußnoten

1. <http://ada.cinepoetics.org/explorer/>
2. <https://frametrail.org/>
3. <https://www.w3.org/DesignIssues/LinkedData.html>
4. <https://github.com/openlink/virtuoso-opensource/>
5. <https://github.com/LodLive/LodView>
6. <https://ada.cinepoetics.org/sparql/>
7. Download der Ontologie unter <https://github.com/ProjectAda/ada-ae/tree/main/filmontology>, Visualisierung unter <https://ada.cinepoetics.org/ontoviz/>
8. Beispiel einer Ontologie URI: Typ der Kamerabewegung <http://ada.cinepoetics.org/resource/2021/05/19/Annotation-Type/CameraMovementType>
9. Beispiel einer Metadaten URI: Occupy Wall Street: <https://ada.cinepoetics.org/resource/media/39953b6cce-a8c49b0a119f1715aab20818e4564cc4b2c2e8567722c9f418f1b9>
10. Download der Metadaten unter https://github.com/ProjectAda/ada-ae/tree/main/corpus_metadata
11. <https://www.w3.org/TR/annotation-model/>
12. <https://www.w3.org/TR/media-frags/>
13. Die Annotation ist auch per URI abrufbar: <http://ada.cinepoetics.org/resource/media/294704ee3b-d55a6888235ae7721120c29522eddd3cc273cc8365fa0ee-f2ac56d/ed63d084-717f-11e9-99b8-0242ac130003>
14. <https://jena.apache.org/>
15. <https://javalin.io/>
16. <https://github.com/jsonld-java/jsonld-java>
17. [https://project1.ada.cinepoetics.org/explorer/?r\[\]=t_n_crisis_w_48&ui=kfc&unit=movie&view=movie](https://project1.ada.cinepoetics.org/explorer/?r[]=t_n_crisis_w_48&ui=kfc&unit=movie&view=movie)
18. Detaillierte Anleitung zur Benutzung des Annotation Explorer findet sich in Pfeilschifter et al. 2021, S. 151ff.
19. Vgl. dazu den Themenschwerpunkt der *mediaesthetics* -Ausgabe „Digitale Filmanalyse und Bilder der Krise“.

Bibliographie

- Arnold, Taylor / Tilton, Lauren / Berke, Annie** (2019): „Visual Style in Two Network Era Sitcoms“, in: *Journal of Cultural Analytics* 1(2) 10.22148/16.043.
- Aubert, Olivier / Prié, Yannick** (2005): „Advene: active reading through hypervideo“, in: *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia* 235-244.
- Bakels, Jan-Hendrik / Grotkopp, Matthias / Scherer, Thomas / Stratil, Jasper** (2020a): „Digitale Empirie? Computergestützte Filmanalyse im Spannungsfeld von Datenmodellen und Gestalttheorie“. In: *montage/AV* 29(1) 99-118.

Bakels, Jan-Hendrik / Grotkopp, Matthias / Scherer, Thomas / Stratil, Jasper (2020b): "Matching Computational Analysis and Human Experience. Performative Arts and the Digital Humanities". In: *Digital Humanities Quarterly* 14(4) <http://www.digitalhumanities.org/dhq/vol/14/4/000496/000496.html> [letzter Zugriff: 26.11.2021].

Bakels, Jan-Hendrik / Scherer, Thomas / Stratil, Jasper / Agt-Rickauer, Henning (2020): "AdA Filmontology – a machine-readable Film Analysis Vocabulary for Video Annotation", in: *Book of Abstracts: DH2020 – carrefours/intersections* https://dh2020.adho.org/wp-content/uploads/2020/07/488_AdAFilmontologyamachinereadableFilmAnalysisVocabularyforVideoAnnotation.html [letzter Zugriff: 26.11.2021].

Flückiger, Barbara / Halter Gaudenz (2020): "Methods and Advanced Tools for the Analysis of Film Colors in Digital Humanities". In: *Digital Humanities Quarterly* 14(4) <http://www.digitalhumanities.org/dhq/vol/14/4/000500/000500.html> [letzter Zugriff: 26.11.2021].

Freedman, Richard / Fiala, David / Micah, Walter (2019): "The Quotable Musical Text in a Digital Age: Modeling Complexity in the Renaissance and Today". In: *Book of Abstracts DH2019*, <https://dev.clariah.nl/files/dh2019/boa/0402.html> [letzter Zugriff: 26.11.2021].

Kappelhoff, Hermann (2018): *Kognition und Reflexion: Zur Theorie filmischen Denkens*. Berlin/Boston, MA: De Gruyter.

Kappelhoff, Hermann / Bakels, Jan-Hendrik / Berger, Hanno / Brückner, Regina / Böhme, Dorothea / Chung, Hye-Jeung / Dang, Sarah-Mai / Gaertner, David / Greifenstein, Sarah / Gronmaier, Danny / Grotkopp, Matthias / Haupts, Tobias / Illger, Daniel / Lehmann, Hauke / Lück, Michael / Pogodda, Cilli / Roleff, Naomi / Rook, Stefan / Rositzka, Eileen / Scherer, Thomas / Schlochtermeyer, Lorna / Schmitt, Christina / Steininger, Anna / Tag, Susanne (2011-2016): *Empirische Medienästhetik. Datenmatrix Kriegsfilm – eMAEX*. <https://www.empirische-medienaesthetik.fu-berlin.de/emaex-system/index.html> [letzter Zugriff: 26.11.2021].

Masse, Mark (2011): *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. Sebastopol, CA: O'Reilly Media.

Pfeilschifter, Yvonne / Prado, João / Zorko, Rebecca / Buzal, Anton / Scherer, Thomas / Stratil, Jasper / Bakels, Jan-Hendrik (2021) *Manual: Annotieren mit Advene und der AdA-Filmontologie*. <https://www.ada.cinepoetics.fu-berlin.de/ada-toolkit> [letzter Zugriff: 26.11.2021].

Scherer, Thomas / Stratil, Jasper (in Ersch.): "Can't Read my Broker Face? – Tracing a Motif and Metaphor of Expert Knowledge Through Audiovisual Images of the Financial Crisis". In *Literature Compass*.

Stratil, Jasper (in Ersch.): "Geteilte (Medien-)Erinnerung und die Zeiten der Krise. Zum Diskurs audiovisueller Finanzkrisendarstellungen anhand von 'Breaking News'", in: *mediaesthetics* (4).

Adapting Coreference Algorithms to German Fairy Tales

Schmidt, David

david.b.schmidt@uni-wuerzburg.de
Universität Würzburg, Germany

Krug, Markus

markus.krug@uni-wuerzburg.de
Universität Würzburg, Germany

Puppe, Frank

frank.puppe@uni-wuerzburg.de
Universität Würzburg, Germany

Introduction

Coreference Resolution has been posing an ongoing challenge to researchers for more than 50 years. It is the task of grouping mentions (concrete or abstract references, represented as textual spans) into clusters representing entities. The approaches for solving this problem have been manifold and range from rule-based approaches (Lee et al., 2013) over classical machine learning approaches (Rahman and Ng, 2009) to modern approaches based on Deep Learning (Lee et al., 2017; Joshi et al., 2020). Coreference Resolution can act as a "glue" between information that is extracted on a local level (usually sentences) in order to obtain representations for an entire document or a collection of documents. This enables many interesting downstream applications such as the creation of character networks (Elson et al., 2010; Krug, 2020) or tracking of events involving central objects in textual media (such as the dagger in Emilia Galotti) (Hatzel and Biemann, 2021). The transfer of existing approaches to new domains or types of text usually comes with a drop in performance. In this work, we examine the performance of a rule-based and an end-to-end Deep Learning algorithm and their adaptability to the domain of German fairy tales. These experiments should provide insight into: a) the drop experienced from one kind of texts to another b) the reliability of state-of-the-art Deep Learning approaches compared to rule-based approaches for a change of texts and c) Capabilities for the adaptation to mitigate this natural drop in performance. This helps to estimate the required amount of manual work that is to be expected when transferring to a new kind of text, especially when the new type features a low number of annotated documents.

For this we use fragments of German novels provided from the DROC corpus (Krug et al., 2018) and as target domain we make use of annotated fairy tales by the Brothers Grimm. In the next section we present our data, followed by the coreference algorithms as well as the methods for the domain adaptation in more detail. We conclude the paper by presenting and discussing the results of our experiments and potential follow up work.

Related Work

There are several recent works that evaluate the performance of coreference resolution models when applied to a different domain than they have been trained on. Srivastava et al. (2018) examine the performance of several coreference resolution systems (rule-based, statistical and projection-based) on English and German out-of-domain data and find that the rule-based system is the best choice for their use cases. Han et al. (2021) train a coreference resolution model based on c2f (Lee et al., 2018) and SpanBERT (Joshi et al., 2020) on two different corpora, Ontonotes (Hovy et al., 2006) and their new corpus FantasyCoref. They then evaluate both models on FantasyCoref and find that the model trained on the same domain outperforms the other one. (Toshniwal et al., 2021) examine the generalization capabilities of coreference models by evaluating the performance of longdoc (Toshniwal et al., 2020) on out-of-domain data using several English datasets and find that models which have been trained on several datasets jointly perform better than those trained on a single dataset.

Data

The data sets for our experiments were the DROC corpus (Krug et al., 2018), comprising 90 fragments of German novels, and 46 tales from the seventh edition of the Children’s and Household Tales by the Brothers Grimm¹. 40 of these fairy tales have been released together with networks of the important characters and their relations (Schmidt et al., 2021). The mentions and their coreference ids have been annotated by human annotators in both data sets. A notable difference between the two data sets is that DROC has gold information about direct speeches, speakers and addressees while for the fairy tales this information was generated automatically. All other information that the algorithms might use, e.g. POS tags or dependency parse trees, has been annotated automatically in both data sets.

Tab. 1: Average number of mentions per document in DROC and the fairy tales, and the ratios of names, noun phrases and pronouns

	Number of Mentions	Names	Noun Phrases	Pronouns
DROC	579	11.4%	20.1%	68.5%
Fairy Tales	296	5.2%	31.0%	63.8%

Table 1 shows some statistics about the mentions in the documents of DROC and the fairy tales. One can see that a document in DROC is on average about twice as long as a document of the fairy tales. Names are used a lot less often in fairy tales, while the usage of noun phrases increases and that of pronouns is comparable.

There is also an important difference regarding the entities that are referred to by the annotated mentions: In DROC, only human characters are annotated. In the fairy tales, animals and legendary beings (like giants) are also annotated because they are important (and sometimes the only) characters (e.g. the Wolf in *Little Red Riding Hood/Rotkäppchen* or the main characters in *Town Musicians of Bremen/Die Bremer Stadtmusikanten*).

A notable difference of both corpora to a lot of other corpora like OntoNotes (Hovy et al., 2006) and LitBank (Bamman et al., 2020) is how the mentions are annotated: OntoNotes annotates the maximal extent of a span (e.g. ‘[eine kleine süße Dirne]’) while DROC and the fairy tales only annotate the heads (‘eine kleine süße [Dirne]’).

For the experiments, DROC was split (a fix split) into a training set and a test set in a ratio of 80% to 20%: 72 documents for training and 18 for evaluation. The fairy tales were evaluated via five-fold cross validation.

Method

In order to assess the capabilities of domain adaptation from German novels to German fairy tales, we made use of a rule-based coreference resolution system and a model based on neural networks. We briefly present both methods followed by the way of adaptation.

The rule-based approach we use is an adaptation of the sieves algorithm by (Lee et al., 2013) to German (Krug et al., 2015). It partitions its rules into so-called sieves, which are ordered by the precision of their rules and applied one after the other to a document. This enables the rules to make use of the decisions of previously applied rules. Most rules use string matching to resolve names and noun phrases. Among the first sieves is also one that uses information about direct speeches to resolve all first person pronouns to the speaker and all second person pronouns to the addressee, and another that resolves relative and reflexive pronouns based on dependency parse trees. All other pronouns are resolved at the end since they do not possess much helpful information and can only be resolved unreliably (compared to a lot of names and noun phrases).

As Deep Learning architecture, we decided to use c2f (Lee et al., 2018)². It is based on e2e (Lee et al., 2017), which was the first end-to-end neural network-based architecture for coreference resolution. e2e begins by building span representations for all spans up to a pre-defined length. All span representations are scored by a feed-forward neural network and only the top-scoring spans are kept (usually about 40% of all spans), all others are discarded. Each remaining span representation is then paired with a pre-defined number of potential antecedents and the pairs are scored by another feed-forward neural network (aside from the span representations the feed forward neural network (FFNN) also receives additional information like the domain of the document and whether both spans have the same speaker). Since not all span representations actually have an antecedent they are also paired with a dummy antecedent. For each span, the highest-scoring partner is picked as antecedent (unless the dummy antecedent was the highest-scoring partner, then the span does not have an antecedent). The architecture of c2f extends e2e in several ways, the two most important are the following: First, it uses a coarse bilinear scoring function, which is easier to compute, to prune the span representations before they are scored by the FFNN. Secondly, it scores the span representation pairs more than once and refines the span representations based on the scoring results between the iterations.

The adaptation of both approaches was done as follows:

Rule-based approach: Most rules in the sieves algorithm previously skipped family relation words and did not try to resolve them to an antecedent. In fairy tales, family relation words are most often unique (e.g. there is only one character called mother and one called father), so family relation words now are resolved to an antecedent if they are preceded by a definite article. Reflexive pronouns are resolved with the help of a dependency parse tree, which was not possible for several reflexive pronouns in the fairy tales. These are now resolved together with most other pronouns (lacking information about gender and number, hardly any antecedent can be ruled out, so they are usually resolved to the first that is checked). In addition to that, there were a few small

changes that were done as the result of an error analysis on the fairy tales but are not motivated by the domain (they would probably also slightly improve the results on DROC).

Deep Learning approach: We trained and evaluated three variants of the c2f algorithm: c2f trained on DROC (c2f D) for 75000 steps, c2f trained on the fairy tales (c2f FT) for about 50000 steps and c2f pre-trained on DROC for 75000 steps and fine-tuned on the fairy tales for an additional 20000 steps (c2f D+FT)³. The maximum number of words a span may contain was reduced from 30 to 4 since the mentions annotated in DROC and the fairy tales are significantly shorter than the mentions in most English corpora. As language model we used a German ELMo model trained on Wikipedia (May, 2019)⁴.

Results and Discussion

Table 2 displays the results of the sieves algorithm (old and adapted version) and c2f (trained on DROC, the fairy tales or both) on DROC (first two rows) and the fairy tales. As metrics we use MUC (Vilain et al., 1995), B³ (Bagga and Baldwin, 1998), CEAF_E (Luo, 2005) and as well as LEA (Moosavi and Strube, 2016).

	MUC			B ³			CEAF _E			LEA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Sieves old	85.2	77.7	81.2	68.8	43.5	52.3	32.3	56.0	39.6	57.5	37.1	44.3
c2f	89.8	91.2	90.5	48.5	59.8	52.7	31.0	46.2	36.6	45.8	57.7	50.1
Sieves old	86.1	77.5	81.5	67.5	44.7	52.9	28.2	49.9	34.9	55.6	38.0	44.2
Sieves	87.3	80.9	84.0	67.1	47.6	55.0	33.9	54.5	40.9	57.8	41.3	47.4
c2f D	88.6	90.2	89.4	56.2	46.6	49.5	47.8	26.2	32.4	53.9	42.7	46.3
c2f FT	90.4	90.7	90.5	59.0	56.4	57.0	52.2	30.4	37.1	57.1	53.3	54.4
c2f D+FT	91.2	91.6	91.4	64.2	61.6	62.5	56.5	34.1	40.9	62.4	58.7	60.1

Tab. 2: Results (Precision, Recall and F1 score) of the rule-based sieves algorithm and the coarse-to-fine algorithm on DROC (first two rows) and on the fairy tales (all other rows). c2f was either trained on DROC (D), the fairy tales (FT) or both. Note that the sieves algorithm uses gold mentions while c2f does not. The best results on the fairy tales are marked in bold

The results show multiple interesting aspects:

1. The results of the non-adapted rule-based system appear to be rather stable between domains (with the exception of CEAF_E, which drops surprisingly), while the Deep Learning model has a significant drop when it is evaluated on a domain it has not been trained on (e.g. 46.3% LEA F1 vs 50.1% on DROC and 54.4% on the fairy tales). One reason for this drop is that it did not recognize a lot of references to animals as mentions since animals are not annotated in DROC. The sieves algorithm did not suffer from this because it uses gold mentions.
2. Training and evaluating c2f on fairy tales yields a performance better than doing both on DROC.
3. Domain Adaptation of a Deep Learning model is pretty easy by just training on different data first and then fine-tuning on in-domain data.
4. While requiring more effort, the adaptation of the rule-based system also yields significant improvements on the domain of the fairy tales. And so far only very rudimentary changes have been made and it is to be expected to further improve the results with more in-depth analysis.
5. On DROC, c2f shows overall better performance than the sieves algorithm. On the fairy tales, the performance gap (when c2f is trained only on fairy tales) is even larger. This is even though in both cases c2f is at a disadvantage since the sieves algorithm uses gold mentions and c2f does not.
6. The version of c2f that is pre-trained on DROC and fine-tuned on the fairy tales (c2f D+FT) outperforms all other systems (by over 5% when measured with LEA or B³). This (unsurprisingly) shows that the neural network profits from larger data sets.

We have shown that domain adaptation of both, a rule-based system and a Deep Learning based system, yields substantial improvements to coreference resolution on a target domain (in our case fairy tales). The evaluation also opens possibilities for further combination of the results of the rule-based system and the Deep Learning based system, which we leave for further work.

Footnotes

1. We use only 46 tales because the other documents have not been annotated yet. The data used can be found at <https://gitlab.informatik.uni-wuerzburg.de/kallimachos/coref-adaptation>
2. Most other NN architectures require even more memory and time to train. We also spent some effort experimenting with a more memory-efficient architecture (Kirstain et al., 2021) but could not get any substantial results.
3. Since we do not have any GPUs with sufficient memory capacity (more than 24 GB) c2f D was trained on CPUs, which took about one week. Training on the fairy tales was done on an RTX3090 in a few hours.
4. <https://github.com/t-systems-on-site-services-gmbh/german-elmo-model>

Bibliography

- Bagga, A. and Baldwin, B.** (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.
- Bamman, D., Lewke, O., and Mansoor, A.** (2020). An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.
- Elson, D. K., Dames, N., and McKeown, K. R.** (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 138–147. Association for Computational Linguistics.
- Han, S., Seo, S., Kang, M., Kim, J., Choi, N., Song, M., and Choi, J. D.** (2021). FantasyCoref: Coreference resolution on fantasy literature through omniscient writer’s point of view. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hatzel, H. O. and Biemann, C.** (2021). Towards layered events and schema representations in long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39.
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R.** (2006). OntoNotes: The 90% solution. In *Proceedings*

of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pages 57–60, New York City, USA. Association for Computational Linguistics.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Kirstain, Y., Ram, O., and Levy, O. (2021). Coreference resolution without span representations. *arXiv preprint arXiv:2101.00434*.

Krug, M. (2020). *Techniques for the Automatic Extraction of Character Networks in German Historic Novels*. Doctoral thesis, Universität Würzburg.

Krug, M., Puppe, F., Jannidis, F., Reger, I., Weimer, L., and Macharowsky, L. (2015). Rule based coreference resolution in german historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.

Krug, M., Puppe, F., Reger, I., Weimer, L., Macharowsky, L., Feldhaus, S., and Jannidis, F. (2018). Description of a corpus of character references in german novels - DROC [Deutsches Roman Corpus]. In *DARIAH-DE Working Papers*. DARIAH-DE.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational linguistics*, 39(4):885–916.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2 (Short Papers), pages 687–692.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

May, P. (2019). German ELMo Model.

Moosavi, N. S. and Strube, M. (2016). Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 632–642.

Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977. Association for Computational Linguistics.

Schmidt, D., Zehe, A., Lorenzen, J., Sergel, L., Düker, S., Krug, M., and Puppe, F. (2021). The FairyNet corpus - character networks for German fairy tales. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 49–56, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Srivastava, A., Weber, S., Bourgonje, P., and Rehm, G. (2018). Different german and english coreference resolution models for multi-domain content curation scenarios. In *Language Technologies for the Challenges of the Digital Age*, pages 48–61, Cham. Springer International Publishing.

Toshniwal, S., Wiseman, S., Ettinger, A., Livescu, K., and Gimpel, K. (2020). Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 8519–8526, Online. Association for Computational Linguistics.

Toshniwal, S., Xia, P., Wiseman, S., Livescu, K., and Gimpel, K. (2021). On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Anzeigen als Daten Dynamisches Tagging und iterative Auswertung eines frühneuzeitlichen Intelligenzblattes

Serif, Ina

ina.serif@unibas.ch

Universität Basel, Switzerland

Reimann, Anna

anna.reimann@unibas.ch

Universität Basel, Switzerland

Engel, Alexander

alexander.engel@unibas.ch

Universität Basel, Switzerland

Proto-Industrialisierung, Industrielle Revolution, Konsumrevolution, Handelsrevolution: Solche Forschungsparadigmen haben das 18. und 19. Jahrhundert zunehmend als eine Zeit des Übergangs zum industriellen Zeitalter gezeichnet, als eine Zeit des institutionellen Wandels, der Intensivierung von Produktion und Konsum, als eine Zeit der verstärkten Arbeitsteilung, letztlich: der Vermarktlichung (Grundlegend Mendels 1972; McKendrick/Brewer/Plumb 1982; Mui/Mui 1989; de Vries 2012; einen aktuellen Forschungsüberblick bieten Blondé/Van Damme 2018). Das SNF-Projekt "Märkte auf Papier – das Basler Avisblatt 1729–1844"¹ nähert sich diesen (und anderen) Themen, indem es ein wöchentlich erscheinendes Anzeigenblatt untersucht, das über einen Zeitraum von 116 Jahren einen erheblichen Teil des sozioökonomischen Austauschs in einer Schweizer Grossstadt widerspiegelt, und zwar hinsichtlich Angebot und Nachfrage von Waren, gebrauchten Gütern oder Dienstleistungen, des Wohnungs- und Stellenmarkts, des Geldverleihs und vieler anderer Aspekte (zu Intelligenzzeitungen grundlegend Blome 2006; Tantner 2015).

Das Projekt zielt auf eine digitale Aufbereitung der Quelle ab, die nicht nur in der Bereitstellung digitaler Bilder besteht, sondern die Zeitungsanzeigen in die strukturierte Form einer Datenbank überführt, die als Grundlage für verschiedene historische Studien dient; über unterschiedliche Plattformen und Werkzeuge werden diese auch über das Projektende im Dezember 2022 hinaus zur Verfügung gestellt und analysiert werden.²

In einem ersten Schritt hat die Universitätsbibliothek Basel qualitativ hochwertige Digitalisate aller erschienenen Avisblatt-Ausgaben produziert und mitfinanziert. Das technische Rückgrat der

daraus resultierenden digitalen Sammlung ist eine von der Data Futures GmbH³ entwickelte iiif-basierte Annotationsinfrastruktur, genannt Freizo (MongoDB). Sie bietet in Kombination mit dem Mirador-Viewer neben Anzeige- und Speichermöglichkeiten auch Authentifizierungs- und Bearbeitungsmöglichkeiten sowie Import- und Exportfunktionalitäten.

Ausgehend von den einzelnen Anzeigen als Analyseeinheiten haben wir einen eindeutig referenzierbaren Datensatz für jede einzelne Anzeige erstellt: mit Anzeigen-ID, Transkription, Zeitspempel, iiif-Bildfragment und Annotator:in-ID. Für die Layouterkennung wurde dhsegment verwendet, um automatisch Bounding Boxes für die Anzeigen zu erstellen (Ares Oliveira/Seguin/Kaplan 2018). Diese Boxes wurden in Transkribus korrigiert und verfeinert, um dort danach die automatische Texterkennung durchzuführen.⁴ Für die Texterkennung haben wir zwei Modelle mit einer Zeichenfehlerrate (CER) im Validierungsset von weniger als 1,7 Prozent trainiert. Das resultierende page-xml wurde dann in die Forschungsumgebung Freizo eingespeist, wo es im Mirador-Viewer angezeigt oder als TSV für die weitere Analyse mit R exportiert werden kann. Durch diesen Prozess haben wir Zugriff auf fast eine Million Datensätze, die aus 6.600 Avisblatt-Ausgaben bzw. 48.000 Seiten extrahiert wurden: der komplette Bestand aller im Avisblatt über die Laufzeit von 116 Jahren veröffentlichten Anzeigen.

Wir nutzen R und Github als unsere Data-Science-Umgebung, um zusätzliche Metadaten zu den Datensätzen hinzuzufügen und spezifische Werkzeuge und Methoden für die Analyse zu entwickeln; diese Skripte und die Daten werden später als öffentliches Repository zur Verfügung gestellt.

Da die Anzeigen sowohl extrem zahlreich als auch in ihrer Art sehr unterschiedlich sind, gehört deren Klassifizierung zu den wichtigsten Metadaten, die hinzugefügt werden müssen: Für Untersuchungen beispielsweise des Basler Wohnungs- oder Büchermarkts, von Aktivitäten im Rahmen der Herbstmesse oder von Auktionen als Transaktionsform wird jeweils nur die relevante Teilmenge an Anzeigen als Forschungsgrundlage benötigt.

Im Avisblatt selbst sind die Anzeigen bereits klassifiziert, wenn auch in allgemeiner, pragmatischer und wechselnder Weise: Die Hauptrubriken, unter denen Anzeigen abgedruckt werden, sind "Zum Verkauf wird angetragen", "Zum Ausleihen wird offeriert", "Zu kaufen begehrt", "Kost, Informationen und Bedienung", "Verlorene und gefundene Sachen" und schliesslich die gut durchmischte Rubrik "Allerhand Nachrichten" – und es gibt weitere, die im Laufe der 116 Jahre neu eingeführt wurden oder wieder verschwunden sind. Wir haben die verschiedenen Hauptrubriken über Texterkennung identifiziert und diese als Metadaten für jede Anzeige als erste Klassifizierung aufgenommen – die aber allein nicht ausreicht. Eine vollständige manuelle Klassifizierung aller Datensätze wäre jedoch allein aufgrund ihrer schier Anzahl nicht durchführbar – deswegen sind Anzeigenblätter bisher kaum als wirtschafts- und konsumhistorische Massenquelle in entsprechender Breite und Tiefe analysiert worden (eine der wenigen Arbeiten dazu ist Homburg 1991).

Stattdessen haben wir die Strategie des algorithmischen Taggings entwickelt. Mit Hilfe von R haben wir eine Klasse von Funktionen definiert ("Tag-Filter"), die jeweils ein positives und ein negatives Dictionary mit regular expressions enthalten, um Anzeigen zu erfassen, die wir entsprechend taggen wollen (z.B. "Kleidung" oder "Mietangebot"). Jede Funktion kann dabei auch auf Anzeigen beschränkt werden, die nur unter bestimmten Überschriften, also Rubriken, erscheinen. Aufgrund des skriptbasierten Ansatzes von R können wir die Metadaten jederzeit aktualisieren, wenn wir Tagfilter hinzufügen oder ändern, sodass hier ein dyna-

mischer statt eines statischen Tagging-Ansatzes zur Anwendung kommt.

Die Vorteile dieser dynamischen, algorithmischen Verschlagwortung sind enorm. Erstens ist sie skalierbar: Anders als bei einem manuellen Tagging ist die Gesamtzahl der zu klassifizierenden Anzeigen praktisch irrelevant; der einzige Unterschied zwischen der Verschlagwortung von einigen hundert oder einigen hunderttausend Anzeigen sind ein paar Minuten Rechenzeit. Zweitens sind die Ergebnisse, anders als bei den Entscheidungen, die einer manuellen Klassifizierung zugrundeliegen, vollständig reproduzierbar und immer eindeutig nachvollziehbar. Drittens ist der Ansatz extrem flexibel: Anstatt ex ante eine feste Klassifikation der Datensätze zu erstellen, die dann die Analyse vorgibt und einschränkt (d.h. was gezählt oder zusammengefasst werden kann und was nicht), können die Klassifikationen angepasst und weiterentwickelt werden, wenn sich im Analyseprozess neue Erkenntnisse und Ideen ergeben. Jeder:r Forscher:in, der:die mit den Daten arbeitet, kann seine:ihre eigene Klassifizierung in einem überschaubaren Zeitrahmen erstellen.

Obwohl die Vorteile überwiegen, gibt es auch einen Nachteil des algorithmischen Ansatzes: Im Gegensatz zur manuellen Klassifizierung und zur Verwendung der Rubrikeninformation aus der Quelle selbst produziert er zwangsläufig einige Klassifizierungsfehler, sowohl false positives (Anzeigen, die fälschlicherweise als einem Tag zugehörig getaggt sind) als auch false negatives (Anzeigen, die fälschlicherweise nicht als einem Tag zugehörig getaggt sind). Die Optimierung eines Tagfilters zur Vermeidung von false positives führt zu einem übermässig konservativen und vorsichtigen Filter, der die Erkennungsrate absenkt (und mehr false negatives produziert). Die Optimierung zur Vermeidung von false negatives führt wiederum dazu, dass er zu viel einbezieht und mehr Beifang (d.h. false positives) produziert.

Eine Strategie, damit umzugehen, wäre, einzelne fehlende Datensätze manuell ein- und falsch markierte manuell auszuschliessen. Dies vermindert allerdings nicht nur die Reproduzierbarkeit, auch der Zeit- und Arbeitsaufwand steigt mit der Anzahl der zu prüfenden Datensätze. Da dies unter Umständen in Einzelfällen dennoch sinnvoll sein kann, haben wir die Möglichkeit ebenfalls in unseren Tagfilter-Skripten implementiert.

Als guter Weg, um sowohl false negatives als auch false positives zu minimieren, hat sich ein Bottom-up-Ansatz für den Aufbau und die Kombination von Tagfiltern bewährt: Anstatt einen Filter aufzubauen, der eine grosse und allgemeine Gruppe von Anzeigen umfasst, wie z.B. alle Anzeigen, die sich auf materielle Objekte beziehen, oder nur all diejenigen, die Möbel anonncieren, bauen wir viele Filter, die jeweils einen recht engen Anwendungsbereich haben (wie z.B. Einzelfilter für Betten, Schränke, Stühle, Tische usw.) und kombinieren die resultierende Vielzahl von Tags unter Oberbegriffen (wie z.B. "Möbel"). Filter mit einem engen Anwendungsbereich sind in der Regel selektiver, d.h. sie haben gleichzeitig eine niedrige Fehlerquote für false negatives und false positives.

Dieser Bottom-up-Ansatz ist sehr stark auf spezifische Forschungsfragen und -interessen der Beteiligten ausgerichtet und ignoriert alle Anzeigen und potenziellen Gruppierungen von Anzeigen ausserhalb des eigenen Fokus – genau das aber ist beabsichtigt, da jedes Forschungsinteresse durch eine eigene Klassifikation bedient werden kann. Dies unterscheidet sich von einer einzigen, a priori gesetzten und manuellen one-serves-all Klassifikation, die universeller und interoperabler sein und normalerweise die Gesamtheit aller Anzeigen vollständig aufteilen muss, d.h. jede einzelne Anzeige in eine (allgemein gültige, vorgängig festgelegte) Kategorie einordnen muss – und wenn dies nicht der

Fall ist, schränkt dies die Nutzbarkeit der Datenbank über ein spezifisches Forschungsprogramm hinaus ein.

In den einzelnen Forschungsprojekten innerhalb des Gesamtprojekts dient das dynamische Tagging verschiedenen methodischen Zwecken und zielt auf unterschiedliche Aspekte:⁵ So gibt es (1) eine Reihe von Tagfiltern, die Anzeigen nach ihrem Thema klassifizieren – auf einer allgemeinen Ebene, indem sie (zum Beispiel) den Arbeitsmarkt, den Wohnungsmarkt oder den Tausch von Gegenständen voneinander unterscheiden; oder spezifischer, indem sie verschiedene Arten von Jobs oder verschiedene Kategorien von Dingen klassifizieren. Dann gibt es (2) Tagfilter, die unterschiedliche Arten von Austausch und austauschbezogenen Absichten identifizieren: Sie unterscheiden zwischen dem Verkauf, dem Verleih und der Rückgabe von verlorenen (oder gestohlenen) Gegenständen oder zwischen Angeboten und Gesuchen – einige dieser Informationen sind bereits in den quelleneigenen Überschriften enthalten, allerdings nicht immer zuverlässig bzw. standardisiert. Genauer gesagt erkennen solche Tagfilter verschiedene Formen von Transaktions- oder Marketingpraktiken wie Auktionen, Warenlotterien, Wettbewerbe usw. Sie unterscheiden auch Kleinanzeigen, die auf eine einzelne Transaktion abzielen, von kommerziellen Werbeanzeigen für Unternehmen und Dienstleistungen, die das Ziel hatten, regelmässige zukünftige Geschäfte anzubahnen. Schliesslich gibt es (3) Tagfilter, die bestimmte Umstände prüfen: Ist die Anzeige beispielsweise anonym oder wird der:die Inserent:in genannt, der:die sie ins Avisblatt gestellt hat? Handelt es sich um eine bestimmte Personengruppe – wie z.B. Witwen, die einer Vielzahl von wirtschaftlichen Aktivitäten nachgehen –, oder werden Orte erwähnt (die dann für die Verwendung in GIS-Ansätzen extrahiert werden können)?

Methodisch dient die Identifikation relevanter Teilmengen von Datensätzen als Basis für sehr unterschiedliche Forschungsansätze, die natürlich auch kombiniert werden können und werden: Eine bestimmte Menge relevanter Anzeigen kann einzeln gelesen und hermeneutisch analysiert werden (wenn sie auf einige hundert Datensätze beschränkt ist), sie ist offen für natural language processing oder auch für statistische Analysen.

Eine derartige iterative Analyse kann zu neuen Forschungsfragen führen und neue Forschungsrichtungen aufzeigen. Die Tatsache, dass es dabei mehrere Stränge gibt, mehr als eine Herangehensweise und einen Blickwinkel, um die Quelle zu nutzen, schafft eine zusätzliche Dynamik: Tagfilter und andere Werkzeuge können wiederverwendet und zwischen verschiedenen Forschungsprozessen und Projekten ausgetauscht werden, wodurch die unterschiedlichen Forschungsstränge miteinander verwoben werden. Auch ist eine Weiterverwendung der entwickelten Filter und Analysefunktionen auf weitere Publikationen denkbar – Anzeigenblätter kamen im 18. Jahrhundert in ganz Europa auf und waren gerade auch im deutschsprachigen Raum weit verbreitet; hier finden sich bereits digitalisierte Bestände für viele lokale Intelligenzzeitungen, deren weitere Aufbereitung und anschließende Analyse sich an den für das Avisblatt entwickelten Erschließungsschritten orientieren könnte.

Das Avisblatt-Projekt wird weder von einer spezifischen Forschungsfrage angetrieben, die einen bestimmten und speziell zugeschnittenen Datensatz zur Beantwortung erfordert, noch ist es ein Projekt, das sich auf die Erstellung der Edition einer seriellen Quelle ohne spezifische Forschungsanwendung beschränkt. Letzteres liefe Gefahr, einen Datenfriedhof zu produzieren, ersteres, Single-Use-Datensätze zu produzieren (die später dann zu Datenfriedhöfen werden, nachdem sie verwendet wurden). Indem wir einen Rahmen für dynamisches Tagging und eine skriptbasierte Verarbeitung bereitstellen, fügen wir der digitalisierten Quelle nicht nur nützliche Metadaten hinzu, sondern hoffen, eine gute

Grundlage für weitere Forschung und Nachnutzung zu schaffen, mithilfe derer sich neue iterative Analysen entfalten können.

Fußnoten

1. <https://avisblatt.ch/>.
2. Die Digitalisate sollen in einem nationalen Repositorium bereitgestellt werden; das aktuell noch private Git-Repository, in dem sich die vollständige Datengrundlage und der Code zur Analyse befinden, wird geöffnet.
3. <https://www.data-futures.org/>.
4. <https://readcoop.eu/transkribus>.
5. Das Folgende betrifft Tagfilter, die auf die Anzeigen angewendet werden, aber das Konzept besitzt auch eine grundlegendere Anwendung: Der Wortlaut der Überschriften der einzelnen Rubriken ändert sich über die Jahre, sodass ein spezifisches Tagfilter-Set nur für die Klassifizierung der jeweiligen Überschriften erstellt wurde.

Bibliographie

- Ares Oliveira, Sofia / Seguin, Benoît / Kaplan, Frédéric** (2018): "dhSegment: A generic deep-learning approach for document segmentation", in: *Frontiers in Handwriting Recognition (ICFHR)*. 16th International Conference 2018: 7–12.
- Blome, Astrid** (2006): "Vom Adressbüro zum Intelligenzblatt. Ein Beitrag zur Genese der Wissensgesellschaft", in: *Jahrbuch für Kommunikationsgeschichte* 8: 3–29.
- Blondé, Bruno / Van Damme, Ilja** (2018): "From Consumer Revolution to Mass Market", in: *The Routledge Companion to the History of Retailing*. New York: Routledge: 31–49.
- de Vries, Jan** (2012): *The Industrious Revolution. Consumer Behavior and the Household Economy, 1650 to the Present*. Cambridge: Cambridge University Press.
- Homburg, Heidrun** (1991): "Warenanzeigen und Kundenwerbungen in den ‚Leipziger Zeitungen‘ 1750–1800. Aspekte der inneren Marktbildung und der Kommerzialisierung des Alltagslebens", in: Dietmar Petzina (Hg.): *Zur Geschichte der Ökonomie und Privathaushalte*, Berlin: Duncker & Humblot: 109–131.
- McKendrick, Neil / Brewer, John / Plumb, John H.** (1982): *The Birth of a Consumer Society. The Commercialization of Eighteenth-Century England*. London: Europa Publ. Ltd.
- Mendels, Franklin F.** (1972): "Proto-industrialization. The First Phase of the Industrialization Process", in: *The Journal of Economic History* 32:1: 241–261.
- Mui, Hoh-Cheung / Holbrook Mui, Lorna** (1989): *Shops and Shopkeeping in Eighteenth-Century England*. Kingston / Montreal / London: McGill Queen's University Press Routledge.
- Tantner, Anton** (2015): *Die ersten Suchmaschinen. Adressbüros, Fragämter, Intelligenzcomptoirs*. Berlin: Verlag Klaus Wagenbach.

Auf den Spuren einer altnordischen Saga-Ästhetik Poetologische Aussagen in den Erzählerbemerkungen der Isländersagas

Göggelmann, Michael

michael.goeggelmann@uni-tuebingen.de
Universität Tübingen, Germany

Heiniger, Anna Katharina

anna-katharina.heiniger@uni-tuebingen.de
Universität Tübingen, Germany

Reiter, Nils

nils.reiter@uni-koeln.de
Universität Köln, Germany

Zirker, Angelika

angelika.zirker@uni-tuebingen.de
Universität Tübingen, Germany

Einleitung¹

Der Vortrag stellt die systematischen Annotationen von Erzählerbemerkungen in den anonym überlieferten, mittelalterlichen *Isländersagas* (altnord. *Íslendingasögur*) vor und geht dabei vor allem der Frage nach, ob und inwieweit diese als Teil einer Literarisierungsstrategie wirksam werden und damit Aussagen über ein den Isländersagas möglicherweise inhärentes Konzept von Autorschaft ermöglichen.² Die Untersuchung erfolgt mit quantitativen Methoden auf Grundlage der systematischen Annotation der Texte. Dabei liefert die qualitative Analyse einzelner Erzählerbemerkungen erste Anhaltspunkte für eine solche Literarisierungsstrategie, die sich etwa intratextuell in Form von Verben oder Verbalphrasen (z.B. „sem áðr var“ sagt; ‚Wie zuvor erzählt wurde‘) manifestiert. Der Einsatz quantitativer Methoden ermöglicht folglich erstmals eine textübergreifende Analyse, anhand derer nachgewiesen werden soll, inwiefern sich diese in Einzelbelegen bereits sichtbar werdende Literarisierungsstrategie im Verlauf eines Gesamttextes zu einer poetologischen Aussage verdichtet, aufgrund derer sich das ästhetische Selbstverständnis der *Isländersagas* erschließen lässt. Unsere Annahme ist, dass – wenn auch alle Sagas die gleichen Typen von Kommentaren verwenden – sich aus der Menge der gesammelten Daten ein für jede Saga jeweils individuelles Profil in der Verwendung der Erzählerbemerkungen erkennen lässt.

Während narrative Texte und deren systematische Annotation bereits vielfach Untersuchungsobjekt innerhalb der Digital Humanities waren (Zinsmeister 2016; Gius/Jacke 2017; Adelman et al. 2018; Ketschik et al. 2020), zeigt sich das Innovationspotenzial der vorliegenden Studie in zweierlei Hinsicht: es wird sowohl das bislang quantitativ gänzlich unerschlossene **Korpus** der Islän-

dersagas untersucht, wie auch mit der Annotation **intratextueller Verweise** ein narratologisches Phänomen operationalisiert und in den Blick genommen, das insbesondere hinsichtlich der ästhetischen Faktur der Texte neue Aussagen erlaubt.³

Das Korpus der Isländersagas

Die ca. 40 überlieferten *Isländersagas* (altnord. *Íslendingasögur*) sind wichtige Repräsentanten erzählender volkssprachiger Literatur nicht nur des isländischen, sondern generell des skandinavischen Mittelalters. Es handelt sich um anonym überlieferte Texte, die zwischen dem 13. und 15. Jh. verschriftlicht wurden und die sich hinsichtlich ihres Umfangs zum Teil beträchtlich unterscheiden (Rowe 2017: S. 157). Als entsprechend unterschiedlich gestaltet sich auch die strukturelle Komplexität der Sagas, die zwar in der Regel chronologisch linear erzählen, aber doch häufig mehrere narrative Stränge verfolgen (Clover 1982). Lange Zeit prägten intensive Debatten hinsichtlich des Verhältnisses zwischen Oralität und Literarizität der Sagas die Forschung. Obwohl von den anfänglichen Extrempositionen – Sagas als rein literarische Werke bzw. Sagas als direkte Verschriftlichung oraler Überlieferung – abgerückt wurde, gibt es nach wie vor keinen Konsens bezüglich des Saga-Ursprungs. Viele Forscher sprechen sich mittlerweile für ein Zusammenspiel mündlich geprägter Elemente und einer literarischen Entwicklung der *Isländersagas* aus (Ólason 2005: S. 112-114; Callow 2017).

Obwohl sich die *Isländersagas* kaum explizit zu poetischen Fragen äußern, lassen kurze Bemerkungen der Erzählstimme wie auch weitere narrative Techniken – z.B. die Organisation des Erzählten, der Spannungsaufbau und die dramatische Inszenierung oder auch die Selbstrepräsentation der Erzählstimme – das Bewusstsein von Gattungsregeln ebenso erkennen wie das Bestreben, bestimmte Erwartungen des Publikums zu erfüllen. In unserer Analyse gehen wir deshalb, wie eingangs dargestellt, von der These aus, dass sich poetologische Aussagen in den *Isländersagas* insbesondere dort manifestieren, wo sich durch intratextuelle Verweise Literarisierungsstrategien andeuten.

Die Annotation von Erzählerbemerkungen

Bei intratextuellen Verweisen handelt es sich um in der Forschung bislang kaum beachtete Phänomene in den Erzählerbemerkungen, die wir als Mittel des produktiven Austauschs zwischen der intradiegetischen literarischen Praxis und der extradiegetischen Welt des Publikums betrachten. Um diese zu sammeln, zu systematisieren und zu kontextualisieren sowie im Hinblick auf die narrative (Selbst-)Reflexion in den Isländersagas auszuwerten, wurden deshalb solche Äußerungen der Erzählstimme als Ausgangspunkt gewählt, die sich mit dem Erzählen selbst befassen.

In Vorarbeiten zu dieser Studie wurden Erzählerbemerkungen in den Isländersagas in fünf Kategorien eingeteilt, die ihrerseits die Grundlage für die ersten Annotationsrichtlinien bilden. In der vorliegenden Studie liegt das Augenmerk auf vier näher untersuchte Sagas. Bereits zu Beginn des Annotationsprozesses (wobei wir der Anleitung in Reiter 2020 folgten) wurde deutlich, dass diese für eine produktive Umsetzung in mehreren Schritten geschärft und durch zusätzliche Kategorien ergänzt und ausdifferenziert werden müssen. Die Überarbeitung der Richtlinien ist bisher in fünf aufeinanderfolgenden Runden vorgenommen worden, so dass diese

nun eine erste stabile Form mit sechs Annotationskategorien und meist mehreren Unterkategorien erreichten. Nachfolgend konzentrieren wir uns aus Platzgründen auf die Kategorie der intratextuellen Bezüge – die Annotationen der anderen Kategorien⁴ lassen sich in ähnlicher Weise analysieren.

Die Annotation der Erzählerbemerkungen in den Isländersagas wurde mit Hilfe der Software CorefAnnotator (Reiter 2018) vorgenommen.

Intratextuelle Verweise

Die Kategorie umfasst alle intratextuellen Bezüge, die die Erzählstimme in einer Saga herstellt und gehört zu den am häufigsten annotierten Kategorien. In mehreren Unterkategorien werden bei der Annotation der Sagas verschiedene Arten intratextueller Verweise erfasst. Auf der intratextuellen Ebene nimmt die Erzählstimme eine narrative Selektion vor, erinnert an frühere Geschehnisse, kündigt Geschehnisse an, die erst noch erzählt werden und informiert darüber, welche Figuren neu eingeführt werden oder für die weitere Handlung keine Rolle mehr spielen. Als Beispiele dieser Kategorie lassen sich oft verwendete Phrasen wie „sem fyrr var sagt“ (Laxdœla saga: S. 71; ‚Wie zuvor erzählt wurde‘),⁵ „Nú er at segja frá“ (Reykðæla saga: S. 157; ‚Nun ist zu berichten von‘) oder „ok nefnu vér hana eigi“ (Laxdœla saga: S. 48; ‚Wir nennen sie aber nicht‘) anführen. Gerade durch Bemerkungen wie der letztgenannten wird die durch die Auswahl vorgenommene Rezeptionslenkung in den Sagas deutlich.

Ebenfalls zu den intratextuellen Verweisen zählen häufig verwendete formelhafte Phrasen. Mit Phrasen dieser Art werden zum einen neue Figuren eingeführt („M[aðr] er nefndr Bárðr Heyangrs-Bjarnarson“ (Bárðar saga Snæfellsáss: S. 107) ; ‚Ein Mann wird Bárðr Heyangrs-Bjarnarson genannt‘), zum anderen werden damit Zeitangaben gemacht („Þá var þat á einni nótt“ (Bárðar saga Snæfellsáss: S. 104); ‚Dann war es eines Nachts‘).

Eine weitere intratextuelle Spezifizierung ist die Vorahnung (*foreshadowing*). Mit dieser Unterkategorie werden Textpassagen annotiert, in denen entweder die Erzählstimme oder auch eine Figur eine Vorahnung möglicher bevorstehender Ereignisse ausdrückt, die dann meistens auch eintreffen („mikit illt mun af Hánef hljótask“ (Reykðæla saga: S. 165); ‚Viel Unglück wird durch Hánef gebracht werden‘).

Erste Auswertungen der Annotationen

Die folgenden Analysen wurden auf Basis der bisherigen manuellen Annotationen vorgenommen. Zum jetzigen Zeitpunkt wurden vier Sagas vollständig annotiert, die in der Forschung als ‚randständig‘ innerhalb der Gattung der *Isländersagas* gelten.⁶

Tab. 1: Grundlegende Korpuseigenschaften

Saga	Anzahl Sätze	Anzahl tokens
<i>Bárðar saga Snæfellsáss</i>	959	14 850
<i>Grænlandings saga</i>	352	6 720
<i>Reykðæla saga</i>	1 163	25 549
<i>Stjórnu-Odda draumr</i>	225	5 343

Die unterschiedlich langen Sagas machen einen direkten Vergleich der absoluten Zahlen von Annotationen in ihnen schwierig;

in den folgenden Auswertungen werden Häufigkeiten daher normalisiert. Wir betrachten zunächst die Häufigkeit der Annotationen (Abb. 1).

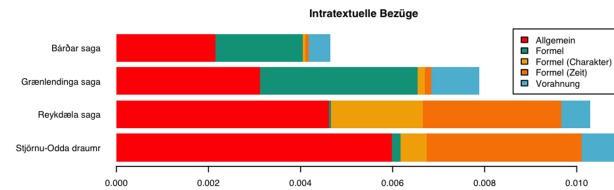


Abb. 1: Anzahl der Annotationen der intratextuellen Verweise mit Unterkategorien.

Bei den *intratextuellen Bezügen* fällt ins Auge, dass formelhafte Referenzen auf Charaktere vor allem in der *Grænlandings saga* und der *Bárðar saga* annotiert wurden, während zeitliche Formeln dort kaum auftauchen. Vorahnungen (*foreshadowing*) machen in allen Sagas einen ähnlichen Anteil der intratextuellen Bezüge aus.

Diese Auswertung der vier Sagas deutet somit darauf hin, dass die Erzählerbemerkungen in jeder Saga ein eigenständiges Profil bilden. Die Hypothese, nach der wir den Isländersagas eine individuelle Ausgestaltung trotz den allen gemeinsamen Typen von Erzählerkommentaren attestierten, konnte also bereits an dieser Stelle plausibilisiert werden.

Verteilung der Annotationen im Text

Einen visuellen Eindruck von der Verteilung der Annotationen im Textverlauf liefert Abb. 2. Jeder senkrechte Strich markiert dabei die Annotation eines intratextuellen Verweises, wobei die verschiedenen Farben die Unterkategorien der intratextuellen Bezüge repräsentieren.

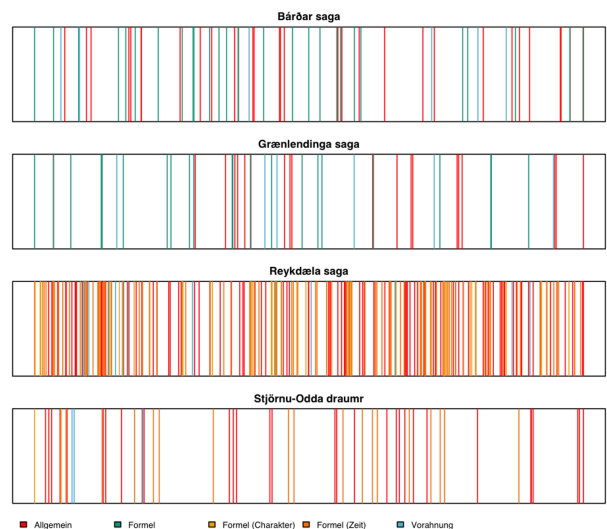


Abb. 2: Grafische Darstellung der Verteilung der Annotationen der Kategorie „intratextuelle Bezüge“ innerhalb der vier exemplarisch diskutierten *Isländersagas*

Grundsätzlich verteilen sich die Annotationen, wie zu erwarten war, über den gesamten Text. Die sich dazwischen befindlichen,

teilweise recht großen Lücken sollen kapitelweise anhand der Annotationsdichte nachfolgend genauer untersucht werden.

Annotationsdichte je Kapitel

Die Dichte der Annotationen wird in Abb. 3 gezeigt. Für jedes Kapitel und jede Kategorie ergibt sich dabei ein Datenpunkt, die Datenpunkte einer Kategorie sind dann durch Linien verbunden. Die Dichte der Annotationen ist hierbei als relative Anzahl an Annotationen pro Kapitel definiert.

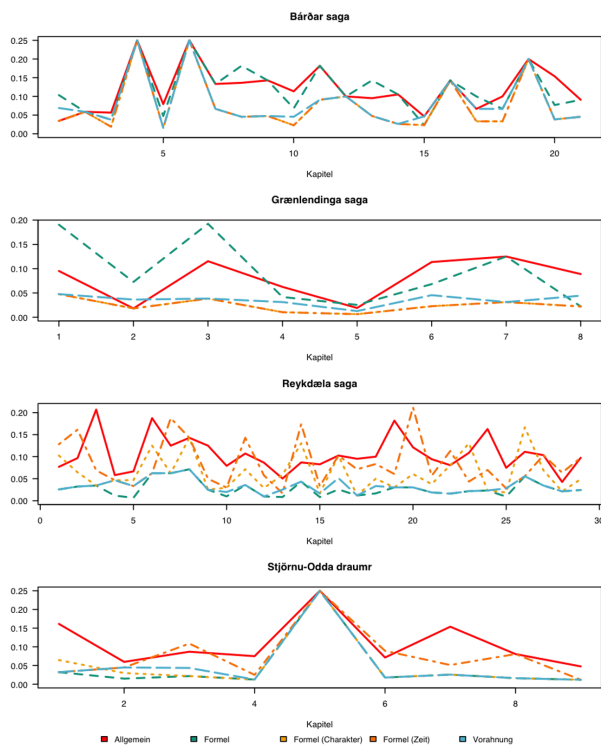


Abb. 3: Annotationsdichte je Kapitel. Visualisiert ist die Anzahl der Annotationen intratextueller Bezüge je Kapitel, normalisiert anhand der Kapitellänge (in *tokens*)

Anhand der Grafiken lassen sich die Höhe- und Wendepunkte der jeweiligen Saga ablesen und nachvollziehen. Die ersten zwei großen Ausschläge der *Bárðar saga* in Kapitel vier und sechs repräsentieren zum ersten Bárðors Ankunft in Island und seine Inbesitznahme von Land auf der Halbinsel Snæfellsnes (Kp. 4), zum zweiten wird der Zeitpunkt von Bárðors Rückzug in die Berge nach dem dramatischen und schmerzvollen, jedoch nur vermeintlichen Verlust seiner Tochter Helga stark markiert (Kp. 6). Die Spitzen im letzten Drittel der Saga zeigen die Herausforderung des Wiedergängers Raknarr und im Zuge dessen den Aufbruch Gestrs, Bárðors Sohn, um Raknarr aufzusuchen und zu bezwingen. In der *Reykðæla saga* stehen die Spitzen der Annotationsdichte bei den Kapiteln sieben, zwölf und dreizehn und zwanzig ebenfalls für drei zentrale Ereignisse.

Die hier für die *Bárðar saga* und *Reykðæla saga* skizzierten Interpretationen der obigen Grafiken weisen darauf hin, dass mittels der quantitativen Analysen nicht nur bisherige qualitative und hermeneutische Auswertungen bestätigt werden können, sondern darüber hinaus eine Möglichkeit eröffnet wird, um den Prozess des Erzählens und somit auch die Struktur der Sagas besser zu ergründen. Die systematische Annotation wirft somit in vielerlei

Hinsicht neues Licht auf die altnordischen Texte: Zum einen ermöglichen Auswertungen wie in Abbildung 3 einen (textübergreifenden) Überblick über die Annotationen. Die Visualisierungen der Daten verdeutlichen abermals die Heterogenität der Textgestaltung innerhalb der Gattung der *Isländersagas*. Zum anderen lässt sich auf Ebene der einzelnen Sagas die Verknüpfung von Erzählerkommentaren und dem Plot anhand der Ausschläge in den Grafiken nachvollziehen. In ihrer Funktion als Gestaltungsmittel begleiten die Erzählerkommentare die Höhe- und Wendepunkte und prägen somit die Handlungsstränge.

Zu bedenken ist, dass Abbildung 3 gegenwärtig ausschließlich die Annotationsdichte der intratextuellen Verweise zeigt. Um ein umfassendes Bild der Annotationsverteilung zu erhalten, müssen in einem nächsten Analyseschritt auch die anderen Annotationskategorien berücksichtigt werden. Daran anschließend wird sich zeigen, ob die Schlüsselstellen in erster Linie mit intratextuellen Kommentaren versehen, oder ob diese auch mit anderen Arten der Erzählerkommentare gekoppelt sind.

Fazit

Durch die vorliegende quantitative Analyse auf Grundlage systematischer Annotationen konnten für die Erzählstimme der Isländersagas Textmerkmale aufgezeigt werden, die zwar zuvor im Einzelfall erkannt, aber nicht im Hinblick auf einen oder mehrere Gesamttexte systematisch erfassbar waren. Mithilfe ihrer visualisierten Distributionen wurde eine neue Perspektive auf die Erzählerbemerkungen geschaffen, deren textübergreifende Bedeutsamkeit bislang nicht erkannt worden war. So lässt die vorliegende Untersuchung annehmen, dass die Distribution der Erzählerkommentare nicht zufällig, sondern an den Handlungsverlauf der Isländersagas gekoppelt ist. Diese Verbindung aus Form und Inhalt wird in einem nächsten Schritt einerseits auf Grundlage der Ergebnisse aus der quantitativen Analyse wieder in die genaue Textanalyse (*close reading*) zurückgeführt. Andererseits gilt es, diese Beobachtung in einem größeren Korpus mess-, beobacht- und beschreibbar zu machen. Dazu planen wir das Verfahren dahingehend weiterzuentwickeln, dass wir Erzählerkommentare auch als Marker für eine automatisierte Erkennung von Handlungseinheiten im Sinne von Zehe et al. (2021) verwenden können.

Bereits jetzt aber zeigen sich die Erzählerbemerkungen in ihrem regelmäßigen Auftreten als Gestaltungsmittel von Höhe- und Wendepunkten der Sagas als derart prägend, dass sich diese als Teil einer Literarisierungsstrategie über Einzelbelege hinweg tatsächlich zu einer poetologischen Aussage verdichten und deren Annotation damit zur ästhetischen Verortung dieser Texte beitragen kann. Auf der Suche nach ästhetischem Potenzial in den Isländersagas trafen wir anhand unserer Methode auf differenzierte Ergebnisse, die auf ein großes Bewusstsein hinsichtlich der Literarisierung und der Ästhetisierung hinweisen. Weitere Analyseschritte zielen auf eine Korpuserweiterung und eine vergleichende Auswertung der Ergebnisse über ein größeres Textkorpus hinweg sowie einer Verfeinerung der Annotations-Tools und ggf. -Kategorien. Darüber hinaus prüfen wir die Möglichkeit einer automatischen Erkennung von Erzählerkommentaren mittels maschineller Lernverfahren auf Grundlage unserer Annotationen. Bei einer hinreichenden Erkennungsrate können ggf. weitere Sagas automatisch annotiert werden. Durch die bei maschinellen Lernverfahren zunächst oft fehlerhaften Verallgemeinerungen erhoffen wir uns zudem auch weitere Einsichten in die Annotationskategorien und deren Verwendung.

Fußnoten

1. Die vorliegende Arbeit wurde gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1391 – Projektnummer 405662736.
2. Hierbei ist nicht Intention, die Urheberschaft der Sagas im Sinne einer Personenbestimmung zu klären, vielmehr geht es um den möglichen Nachweis pluraler Autorschaftskonzepte bzw. Vielstimmigkeit in den Isländersagas.
3. Der Beitrag fasst erste Teilergebnisse zusammen. Neben intratextuellen Verweisen wurden auch andere Arten von Erzählerkommentaren annotiert, die allerdings noch nicht systematisch ausgewertet wurden. Die Annotation intratextueller Verweise wird auch in TEASys praktiziert, dort allerdings als Kategorie der erklärenden Annotation, und nicht als Mittel der Texterschließung. Vgl. hierzu Bauer/Zirker 2020.
4. Bei den nicht näher dargestellten Annotationskategorien handelt es sich um Intertextuelle Verweise, Referentielle Bezüge, Ironische Distanzierung bzw. Erzählstimme, Öffentliche Meinung und Superlative bzw. Hyperbolisches.
5. Die Übersetzungen der altnordischen Zitate stammen von Anna Katharina Heiniger.
6. *Stjörnu-Odda draumr* wird zwar oftmals zum erweiterten Kreis der *Isländersagas* gezählt, ist aber im Grunde ein *pátttr*, ein kurzer, erzählender Prosatext. In der handschriftlichen Überlieferung sind die Texte dieser Art meist als Teile grosser Sagakompendien. Zur Problematik des *pátttr*-Genre vgl. Ármann Jakobsson 2013.

Bibliographie

- Adelmann, Benedikt / Andresen, Melanie / Begerow, Anke / Gaidys, Uta / Gius, Evelyn / Koch, Gertraud / Menzel, Wolfgang / Orth, Dominik / Topp, Sebastian / Vauth, Michael / Zinsmeister, Heike (2018): "hermA. Zur Rolle von Annotationen in hermeneutischen Prozessen", in: Vogeler, Georg (ed.): *DHd 2018: Kritik der digitalen Vernunft. Konferenzabstracts* 412–414.
- "Bárðar saga Snæfellsáss", in: Vilmundarson, Þórhallur / Vilhjálmsson, Bjarni (eds.) (1991): *Harðar saga. Bárðar saga, Þorskfirðinga saga, Flóamanna saga* (Íslenzk fornrit XIII). Reykjavík: Hið íslenzka fornritafélag 99–172.
- Bauer, Matthias / Zirker, Angelika (2020): *TEASys Style Guide - Explanatory Annotation* <http://www.annotation.es.uni-tuebingen.de/wp-content/uploads/2020/09/Styleguide-2020-08-11.pdf> [Letzter Zugriff: 12. Juli 2021].
- Callow, Chris (2017): "Dating and Origins", in: Jakobsson, Ármann / Jakobsson, Sverrir (eds.): *The Routledge Research Companion to the Medieval Icelandic Sagas*. London: Routledge 15–33.
- Clover, Carol (1982): *The Medieval Saga*. Ithaca / London: Cornell University Press.
- Gius, Evelyn / Jacke, Janina (2017): "The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis", in: *International Journal of Humanities and Arts Computing* 11(2): 233–254.
- "Grænlendinga saga", in: Sveinsson, Einar Ól. / Þórðarson, Matthías (eds.) (1935): *Eyrbyggja saga. Brands þátttr orva, Eiríks saga rauða, Grænlendinga saga, Grænlendinga þátttr* (Íslenzk fornrit IV). Reykjavík: Hið íslenzka fornritafélag 239–269.

Jakobsson, Ármann (2013): "The Life and Death of the Medieval Icelandic Short Story", in: *The Journal of English and German Philology* 112(3): 257–291.

Ketschik, Nora / Overbeck, Maximilian / Murr, Sandra / Pichler, Axel / Blessing, André (2020): "Interdisziplinäre Annotation von Entitätenreferenzen: Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung", in: Reiter, Nils / Pichler, Axel / Kuhn, Jonas (eds.): *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Berlin: De Gruyter 203–236.

"Laxdæla saga", in: Sveinsson, Einar Ól. (ed.) (1934): *Laxdæla saga. Halldórs þátttr Snorrasonar, Stífs þátttr* (Íslenzk fornrit V). Reykjavík: Hið íslenzka fornritafélag 1–229.

Ólason, Vésteinn (2005): "Family Sagas", in: McTurk, Rory (ed.): *A Companion to Old Norse-Icelandic Literature*, Malden (MA): Blackwell Publishing 101–118.

Reiter, Nils (2018): "CorefAnnotator - A New Annotation Tool for Entity References", in: *Abstracts of EADH: Data in the Digital Humanities*.

Reiter, Nils (2020): "Anleitung zur Erstellung von Annotationsrichtlinien", in: Reiter, Nils / Pichler, Axel / Kuhn, Jonas (eds.): *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Berlin: De Gruyter 193–202.

"Reykðæla saga", in: Sigfússon, Björn (ed.) (1940): *Ljósvefninga saga með þáttum. Reykðæla saga ok Víga-Skiðu. Hreiðars þátttr* (Íslenzk fornrit X). Reykjavík: Hið íslenzka fornritafélag 149–243.

Rowe, Elizabeth Ashman (2017): "The Long and the Short of it", in: Jakobsson, Ármann / Jakobsson, Sverrir (eds.): *The Routledge Research Companion to the Medieval Icelandic Sagas*. London: Routledge 151–163.

"Stjörnu-Odda draumr", in: Vilmundarson, Þórhallur / Vilhjálmsson, Bjarni (eds.) (1991): *Harðar saga. Bárðar saga, Þorskfirðinga saga, Flóamanna saga* (Íslenzk fornrit XIII). Reykjavík: Hið íslenzka fornritafélag 477–481.

Zehe, Albin / Konle, Leonard / Dümpelmann, Lea / Gius, Evelyn / Hotho, Andreas / Jannidis, Fotis / Kaufmann, Lucas / Krug, Markus / Puppe, Frank / Reiter, Nils / Schreiber, Anneke / Wiedmer, Nathalie (2021): "Detecting Scenes in Fiction: A new Segmentation Task", in: Merlo, Paola / Tiedemann, Jorg / Tsarfaty, Reut (eds.): *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* 3167–3177.

Zinsmeister, Heike (2016): "Modellierung von Forschungsdaten durch Annotation", in: Burr, Elisabeth (ed.): *DHd 2016: Modellierung – Vernetzung – Visualisierung. Konferenzabstracts* 437–438.

Aufführungsinformationen in der Mixed Music Systematische Herausforderungen als Indikatoren musikpraktischer Tendenzen

Akkermann, Miriam

miriam.akkermann@tu-dresden.de
TU Dresden, Germany

Eine aktuelle Herausforderung ist in vielen Bereichen der Musik und der Darstellenden Künste die Strukturierung und Kategorisierung von vorhandenen Aufführungsinformationen. Diese ermöglichen es nicht nur, entsprechende Aufführungen zu beforschen, sondern erlauben es auch perspektivisch, Querverweise und Verlinkungen zu anderen, bereits bestehenden Datenbanken, Archiven und Bibliothekskatalogen zu setzen (Dennerlein 2020). Für musikalische Aufführungen sind solche Informationen sehr heterogen und reichen von Angaben zu Ort und aufführenden Personen bis hin zur verwendeten Instrumentierung und Tonaufnahmen von Einzelaufführungen. Für Verweise zu anderen Informationsquellen ist daher, neben einer guten systematischen Aufbereitung der Informationen selbst, auch die Etablierung übergreifender Categoriesysteme und (Daten-)Modelle nötig, die eine Verlinkung und auch eine adäquate Darstellung der vorhandenen Inhalte ermöglichen. Besonders herausfordernd ist diesbezüglich die Betrachtung Elektroakustischer Musik und Computermusik, zwei Bereiche, die neben der Kategorisierung bekannter musikbezogener Inhalte auch eine Auseinandersetzung mit den sich schnell entwickelnden Technologien erfordern, die diese Musik grundlegend prägen.

Dies wird unter anderem an der im deutschsprachigen Raum genutzten Gemeinsame Normdatei GND deutlich, einem institutionsübergreifenden Format um Inhalte wissenschaftlicher und kultureller Art zu beschreiben und entsprechend die beschriebenen Inhalt kooperativ nutzbar zu machen. Elektroakustische Musik und Computermusik werden bisher in der GND nicht explizit berücksichtigt (Bircher & Wiermann 2018: 2), sondern unter ‚Elektronische Musik‘¹ subsummiert. Dies bringt grundlegende terminologische Herausforderungen mit sich, sowohl in Bezug auf eventuelle Zuordnungen zu Normdaten als auch hinsichtlich der in den Quellen vorhandenen Bezeichnungen: zum einen ist der Begriff ‚Elektronische Musik‘ in der Musikforschung nicht eindeutig und verändert im Laufe des 20. Jahrhunderts gruppenabhängig die Bedeutung, zum anderen spiegeln gerade die vielfältigen wie nicht einheitlichen Bezeichnungen der technischen Instrumentierung implizit die (Weiter-)Entwicklung der Technologien und den jeweiligen Blick auf dieselben wider – zwei Aspekte, die wiederum als Anzeichen musikpraktischer Tendenzen gelesen werden können.

Am Beispiel von Mixed Music Kompositionen, die in den 1980er Jahren am IRCAM entstehen, werden im Folgenden einige begriffliche Unschärfen aufgezeigt und anhand eines bestehenden Aufführungsdatensatzes Herausforderungen für die Kategorisierung der Instrumentierung veranschaulicht. Abschließend wird kurz diskutiert, welche Auswirkungen diese Konstellation auf die Aufführungsinformationen haben kann und inwiefern diese Inkonsistenzen in einer etwas weiter gefassten Betrachtung als Indikatoren aufführungspraktischer Aspekte gelesen werden können.

Herausforderungen der Kategorisierung

Für Werkbeschreibung musikalischer Arbeiten wird im Rahmen der GND auf zwei Listen verwiesen, eine beinhaltet normierte Besetzungsangaben (AH-001) und eine Begriffen für die Kompositionsart (AH-002). In den Besetzungsangaben werden für die Instrumentierung vier Bezeichnungen gelistet, die für den Bereich Elektroakustische Musik und Computermusik zutreffen: ‚Elektronik (Musik)‘, ‚Live Elektronik‘, ‚Tonband‘ und ‚Zuspielaufnahme‘. Die Bezeichnungen ‚Lautsprecher‘ und ‚Tonspur (Zuspielaufnahme)‘ werden zwar genannt, aber nicht als Be-

setzung klassifiziert (DNB 2020b: 17 und 27). Unter ‚Verwendungshinweise und Erläuterungen‘ sind Zusätze dargelegt, die eine Zuordnung zu den Begriffen erleichtern sollen. Für ‚Elektronik‘ ist beispielsweise konkretisiert: ‚Für nicht-spezifische, passive elektronische Klang- und Bilderzeugung (Computermusik, konkrete Musik u. a.). Elektronik wird immer als ein Instrument gezählt. Nicht zu verwechseln mit Live-Elektronik oder Zuspielaufnahme. Bloße Verstärkung wird nicht als Elektronik erfasst.‘ (DNB 2020b: 11)

Computermusik scheint damit der Besetzungsangabe ‚Elektronik‘ zugeordnet zu sein, wobei die ‚aktive‘ Klangerzeugung ebenso ausgeschlossen wird, wie das Zuspielden eines vorgefertigten Audio-Tracks, was als ‚Zuspielaufnahme‘ definiert wird (DNB 2020b: 29). Noch größer wird die Herausforderung einer passenden Zuordnung bei der Betrachtung von sogenannten ‚Mixed Music‘-Kompositionen. Dies bezeichnet musikalische Arbeiten, die – nach der heute gängigen, breiten Definition – zwei Bereiche vereinen: einen instrumentalen Teil, bei dem die Klänge live von Musikerinnen und Musikern erzeugt werden und einen elektronischen Teil, bei dem Klänge elektroakustischen Ursprungs – von einem Tonband bis hin zu digitaler Klangsynthese – über Lautsprecher verbreitet werden (Boutard & Féron 2017; Boutard & Marandola 2014). Frühe Mixed Music Kompositionen sind oft für traditionelle Musikinstrumente und Zuspielband bzw. Tonband komponiert, enthalten also vorproduzierte audiovisuelle Tracks oder Tonaufnahmen, die teilweise im Original mit der Angabe ‚Tonband‘ versehen sind (DNB 2020b: 27). Diese Beschreibung unterliegt jedoch verschiedenen Kategorien: Während die Bezeichnung ‚Tonband‘ für Stücke verwendet werden soll, die für Tonband komponiert sind, soll ‚Zuspielaufnahme‘ verwendet werden, wenn das Tonband nur zur Zuspieldung Verwendung findet. Diese Unterscheidung ist nicht aussagekräftig, da auch eine ‚Zuspielaufnahme‘ vorproduziertes audiovisuelles oder auditives Material umfasst (DNB 2020b: 29). Noch komplizierter wird es, wenn in späteren Aufführungen vormalige Tonbandeinspielungen mit Live Elektronik-Equipment substituiert werden. Für die Instrumentierung ‚Live Elektronik‘ liegen keine weiteren technischen Definitionen vor (DNB 2020b: 18).

Eine andere Perspektive auf die Aufführungsinformationen von Mixed Music präsentieren Serge Lemouton und Samuel Goldszmidt in dem von ihnen erarbeiteten Datenmodell für die interne IRCAM Datenbank Sidney (Lemouton und Goldszmidt 2016). In Sidney können dezidiert Dokumentationen zu den Aufführungen einer Komposition, beispielsweise technische Pläne, Set-up Beschreibungen und auch Programmcode abgelegt werden, die im Laufe der Zeit entstehen. Das Modell ist darauf ausgerichtet, die Informationen insbesondere mit Berücksichtigung der verschiedenen Fassungen, unterschiedlichen Beschreibungsarten und implementierten Technologien zu archivieren. Lemouton und Goldszmidt entwerfen hierfür ein Datenmodell, das die Informationen auf zwei Ebenen in verschiedene Kategorien (Gruppe: work, event, natural person, person function, user; und Inhalt: WorkSidney, WorkFile, VersionFile, Version, Version-NaturalPerson, VersionEquipment) untergliedert und miteinander in Verbindung bringt (Lemouton und Goldszmidt 2016: 4). In Sidney selbst gibt es keine weitere Klassifizierung der (elektroakustischen bzw. Computer Musik-)Kompositionen, die Suche ist nach Personen, Kompositionen und Aufführungsdatum strukturiert. Auch die Benutzeroberfläche der Datenbank Identifiants et Référentiels pour l’enseignement supérieur et la recherche IdRef der Agence bibliographique de l’enseignement supérieur (ABES) ist primär auf Personen- und Werk-/Objektsuche ausgerichtet. Die freie Suche in der IdRef bietet dennoch in Bezug auf Elektroakustischer Musik und Computermusik eine durchaus umfangreiche

Verschlagwortung von Inhalten; nicht alle Kategorien sind jedoch im Katalog Rameau existent, sondern es werden auch Schlagworte verlinkter Datenbanken angezeigt. So sind einerseits die Begriffe ‚Musique électroacoustique mixte‘ und ‚Musique mixte‘ sowie ‚Musique mixte acoustique et électronique‘ vorhanden, andererseits werden Begriffe wie ‚Musique électronique‘, ‚Computer Music‘ und ‚Electronic Music‘ aus verlinkten Systemen übernommen, darunter auch die offen zugänglichen Datenbanken des IRCAM (IdRef 2021).

Die Unschärfen in der Kategorisierung, wie sie beispielsweise bei der Instrumentierung zu sehen sind, werden zusätzlich von terminologischen Herausforderungen begleitet, die in vielfältiger Art und Weise an die eingesetzten Technologien gebunden sind. So werden Mixed Music-Aufführungen in den Quellen einerseits mit unspezifischen (technisch orientierten) Überbegriffen bedacht, während andererseits eine große Vielfalt an individuell beschriebenen Technologien in den Dokumentationen und Aufführungsinformationen zu finden sind.

Terminologie und Technologie

Der Begriff ‚Computer‘ wird beispielsweise sowohl im Sinne einer Objektbezeichnung für eine Zusammenstellung von Hardware- und Softwareelementen verwendet, als auch als Benennung einer allgemeinen Recheneinheit. Der Begriff ‚Tonband‘ (‚Tape‘/‚Bande‘) bezeichnet einerseits das physische Trägermaterial, wird andererseits aber auch dafür verwendet, um zu beschreiben, dass eine vorgefertigte Tonspur – unabhängig vom Speichermedium – vorhanden ist (Akkermann 2021). Beide Begriffe, ‚Computer‘ und ‚Tonband‘ verweisen darauf, dass es eine technische Klangwiedergabe gibt, sagen jedoch nichts darüber aus, welchen Ursprung die Klänge haben, welche technische Spezifikation vorliegt (z.B. Anzahl der Spuren) oder in welcher Art und Weise die Klänge wiederzugeben sind. Entgegen der ungenauen Begrifflichkeit bilden die jeweiligen Beschreibungen jedoch gleichzeitig ein Selbstverständnis hinsichtlich der beschriebenen Technologien ab: sie zeigen an, welchen Fokus eine Dokumentation oder Beschreibung setzt und verweisen auch auf den Blick der Erstellenden auf die implementierten Technologien (Akkermann 2021).

Die Bezeichnungen verweisen mitunter auch auf die Geschichte der eingebundenen Technologien, so ist es beispielsweise bis in die 1990er Jahre durchaus üblich, dass bestimmte Programmiersprachen oder Programme nur auf einer bestimmten Hardware ausgeführt werden können. Zeitweise impliziert daher die Erwähnung einer Programmiersprache oder Software zwangsläufig auch die Verwendung einer bestimmten Hardware. Zudem sind komplexere (digitale) technische Geräte nicht frei verfügbar und bis in die 1990er Jahre nur an bestimmten Institutionen zugänglich; auch ein Produktionsort gibt also eventuell Hinweise auf die verwendete Technologie. Dieser enge Zusammenhang von Hardware und Software wird unter anderem von Curtis Roads durch eine 1996 veröffentlichte Liste herausgestellt, in der Roads unter dem Titel „Unit-Generator-based languages“ 19 Programme benennt, die zwischen 1978 und 1992 zur Steuerung von Echtzeit-DSPs entwickelt werden, sowie die jeweiligen Host-Rechner, die grundlegenden Programmiersprachen, die verwendeten DSPs und die individuellen Standorte skizziert (Roads 1996: 807f). Die Zusammenstellung zeigt, dass bis in die 1990er Jahre einzelne Programme an die jeweiligen Prozessoren angepasst werden müssen, was bedeutet, dass spezielle Programme an bestimmte Hardware gebunden waren, und, dass für jede Aufführung mit anderem Equipment neue Versionen der Programme erstellt werden mussten.

Dies kann als eine Erklärung für die Vielfalt der Bezeichnungen gesehen werden, die bei den Beschreibungen der technischen Instrumentierung zu finden ist. Eine andere Erklärung ist sicher auch, dass die schnellen technologischen Entwicklungen in vielen Wiederaufführungen Updates der Technologien zwingend notwendig machen (Akkermann 2020).

Aufführungsinformationen zu Mixed Music Kompositionen der 1980er Jahre

Wie groß die daraus resultierende Bandbreite an Bezeichnungen ist, kann exemplarisch an Aufführungsinformationen von Mixed Music Kompositionen gezeigt werden, die in den 1980er Jahren und den 2000er Jahren am Institut de Recherche et Coordination Acoustique/Musique IRCAM in Paris erarbeitet und (wieder-)aufgeführt wurden.

Die Aufführungsinformationen entstammen Programmheften aus den Spielzeiten 1979/80 bis 1990/1991 und 1999/2000 bis 2010/11 aus der Mediathek des IRCAM, sowie den IRCAM-Datenbanken Brahms und Sidney. Da es keine zentrale Datenbank für Aufführungen am IRCAM gibt, wurden die Daten manuell im Zeitraum 2015/2020 erhoben und in Listen zusammengefügt. Im ersten Schritt wurden die erhobenen Informationen systematisiert und in Überkategorien zusammengefasst, wie beispielsweise Informationen zur aufgeführten Komposition, Informationen zur Organisation der Aufführung (u.a. Ort, Zeit, Anlass), an der Aufführung beteiligte Personen(-gruppen), akustische Instrumentierung der Aufführung, und technische Instrumentierung der Aufführung.

Bei dieser ersten Zusammenstellung der insgesamt 397 Mixed Music Aufführungen, die für diese zwei Dekaden im Umfeld des IRCAM herausgearbeitet werden konnten, wird eine große Vielfalt in den Bezeichnungen der technischen Instrumentierung sichtbar. Während 58 Begriffe nur einmal verwendet werden, sind die meist genannten Beschreibungen: ‚électronique‘ (235), weitere vier Mal mit dem Zusatz ‚de chambre‘ und zehn Mal mit ‚en temps réel‘. Demgegenüber stehen einmal ‚electroacoustic‘ bzw. drei Mal ‚elektronik‘, ‚Bande‘ (12), ‚Midi‘ (17), ‚Ordinateur‘ (8) und ‚Synthetiseur‘ (18). In dieser Zusammenstellung wird berücksichtigt, was in den Beschreibungen zu den Kompositionen in den jeweiligen Aufführungsunterlagen verzeichnet ist. Es fällt auf, dass die Beschreibungen der Technologien von sehr detaillierten Angaben, beispielsweise der Name eines verwendeten Programms, z.B. ‚Max‘ oder die Bezeichnung eines bestimmten Klangprozessors bis hin zu sehr generalisierenden Angaben wie ‚electronique‘ reichen, wobei die generalisierenden Angaben überwiegen. Zudem sind viele sehr ähnliche Bezeichnungen oder synonyme Bezeichnungen in verschiedenen Sprachen (Französisch, Englisch und Deutsch) zu finden, was mit der hohen Anzahl der unterscheidbaren Beschreibungen erklären kann.

In einem zweiten Schritt wird die Beschreibung einzelner Technologien genauer betrachtet. Hierbei wird untersucht, inwieweit die Beschreibung ähnlicher Technologien bei verschiedenen Aufführungen differieren und wie sich diese technische Instrumentierung zu einer etwas allgemeineren Klassifizierung verhält, beispielsweise einer, die, wie in der Klassifizierung der GND anklängt, an der Art der technischen Klangdarbietung angelehnt ist. Denn, entgegen einer historisch nachvollziehbaren Erwartung, ist nicht der Begriff ‚Tape‘ oder ‚Bande‘ (Tonband), sondern der Begriff ‚Elektronik‘ in der Beschreibung der Aufführungen am häu-

figsten zu finden. Auch unterscheiden sich die Angaben zu den Aufführungen in den Programmheften signifikant von den Angaben, die insbesondere in der Datenbank Sidney zu finden sind. Statt einer Präzisierung der Technologien, wie sie in der Dokumentation stattfindet, werden in der Beschreibung der Aufführungen weiterhin die ursprünglichen Begriffe verwendet, auch wenn diese nicht mehr auf die gleiche Technik verweisen, wie bei den Premieren (Akkermann 2020). Gleichwohl ist auch eine Kategorisierung der zusammengefassten Beschreibungen nicht einfach, so sind auch mit Ausgleich der Übersetzungen weiterhin 42 verschiedene Begrifflichkeiten (einschließlich Varianten) benannt.

Wiederaufführungen und musikpraktische Tendenzen

Die Auswertung der ersten Zusammenstellung der Aufführungsinformationen zeigt, dass zwar eine große Varianz bei den Beschreibungen vorliegt, die Komplexität oder Ungenauigkeit der verzeichneten Technologien sich jedoch nicht (zwingend) negativ auf die Wiederaufführung einer bestimmten Komposition auswirkt. Wird beispielsweise die Frage nach einer Wiederaufführung mit der Größe der Ensembles in Relation gesetzt, so wird deutlich, dass am IRCAM vorrangig Kompositionen in kleiner Besetzung Wiederaufführungen erfahren. Die Sidney-Einträge zeigen aber auch, dass es sich oft um Stücke handelt, deren technische Anlage sich entweder einfach in ein digitales Format umsetzen lässt, oder deren ‚Elektronik‘ eine Klangsynthese per Synthesizer umfasst, was in den Aufführungsdaten selbst oft nicht direkt zu sehen ist. Die originalen Synthesizer bleiben in den Aufführungen meist lange in Benutzung, auch, da sie weniger den technischen Veränderungen unterliegen wie komplexere Set-ups mit vielen technischen Einzelkomponenten. Eine Ausnahme bildet hier beispielsweise Boulezs Komposition *Répons*, die trotz großem Ensemble und komplexem Technologieeinsatz seit den 1980er Jahren regelmäßig aufgeführt wird.

Insgesamt zeigt sich, dass Aufführungsinformationen verschiedene Aspekte anzeigen, die auch als Indikatoren für musikpraktische Ausprägungen gesehen werden können. So ist festzustellen, dass in der Darstellung der Musikstücke eher wenig Akzente auf die Darstellung der implementierten Technologien gelegt wird, obwohl die Technologien gerade aus historischer und analytischer Perspektive eine zentrale Rolle einnehmen. Gleichwohl führen die Inkonsistenzen in den Beschreibungen, große Abweichungen zwischen existierenden Categoriesystemen und auch Änderungen der Bezeichnungen bei Wiederaufführungen oft zu irreführenden Annahmen hinsichtlich der technischen Instrumentierung einer Komposition, was wiederum dazu führen kann, dass Kompositionen in Klassifikationssystemen zu Kategorien zugeordnet werden, die sie inhaltlich nur bedingt ausfüllen. Diese wechselseitige Unschärfe kann positive wie negative Auswirkungen auf mögliche Wiederaufführungen haben: So können einerseits Kompositionen trotz technischer Komplexität ob ihrer Klassifizierung z.B. als ‚Tape‘-Komposition auf den ersten Blick einfach umsetzbar erscheinen und in Programme aufgenommen werden ohne von technischen Herausforderungen abgeschreckt zu werden. Andererseits kann es dazu führen, dass die Kompositionen eher eine Wiederaufführung erfahren, die gut such- und findbar in den gängigen Datenbanken erfasst sind und über Verlinkungen ggf. umfangreiches Zusatzmaterial anbieten. Berücksichtigt man, dass einige technische Informationen verloren gehen, wenn sie nicht innerhalb eines gewissen Zeitraums angepasst oder entsprechende Hardware nachhaltig gesichert werden, kann eine Nichtauffind-

barkeit im Laufe der Zeit auch zu einer Nichtaufführbarkeit führen.

Die hier beschriebene Betrachtung der Bezeichnungen der technischen Instrumentierung ist ein Aspekt, an dem sich zeigen lässt, dass die in den Kategorisierung entstehen Ungenauigkeiten sowohl zu Fehlannahmen hinsichtlich der Quelleninhalte als auch zu (ggf. unbeabsichtigter) Selektion führen können, da jede Kategorisierung auch einen impliziten Blick auf die Quellen festschreibt. Ein systematischer Blick auf aufführungsbezogene Aspekte von Mixed Music kann damit helfen, die Perspektive auf den musik- und kulturhistorischen Kontext weiter auszudifferenzieren.

Fußnoten

1. In der GND ist in der Liste der maßgeblichen Begriffe für die Kompositionsart AH-002 nach RDA 6.14.2 vom 13.10.2020 nur der Titel „Elektronische Musik (nur für Zusammenstellungen)“ hinterlegt (DNB 2020a).

Bibliographie

Akkermann, Miriam (2020): Neue Versionen, neue Urheber. Archiv-Zuwachs durch Technikentwicklung. In Simon Schrör (Hrsg.), *Tipping Points. Interdisziplinäre Zugänge zu neuen Fragen des Urheberrechts*, 241–252. Baden-Baden: Nomos.

Akkermann, Miriam (2021): Vocabulary ruts in Mixed Music – multifarious terms with many ascriptions. *Proceedings of the International Computer Music Conference*. Santiago de Chile 10.5281/zenodo.4161673.

Bircher, Katrin / Barbara Wiermann (2018): Normdaten zu „Werken der Musik“ und ihr Potenzial für die digitale Musikwissenschaft. *BIBLIOTHEK – Forschung und Praxis* Preprint (AR 3218).

Boutard, Guillaume / Françoise-Xavier Féron (2017): La Pratique interprétative des Musiques Mixtes avec Électronique Temps Réel: Positionnement et Méthodologie pour Étudier le Travail des Instrumentistes. In Alain Bonardi (Hrsg.), *Analyser la musique mixte*, 39–60. Paris.

Boutard, Guillaume / Fabrice Marandola (2014): Mixed Music Creative Process Documentation Methodology: Outcomes of the DIP-CoRE Project. *Proceedings of the 9th Conference on Interdisciplinary Musicology – CIM14*. Berlin.

Lemouton, Serge / Samuel Goldszmidt (2016): La préservation des œuvres du répertoire IRCAM: Présentation du modèle Sidney et analyse des dispositifs temps réel, *Journées d'Informatique musicale* hal-01944619.

Roads, Curtis (1996): *The Computer Music Tutorial*. MIT Press.

Dennerlein, Katrin (2020): Panel „Datamodelling History of Drama and (Musical)theater“. *DhD 2020 „Spielräume: Digital Humanities zwischen Modellierung und Interpretation“*.

DNB (2020a): Liste der maßgeblichen Begriffe für die Kompositionsart. <https://wiki.dnb.de/download/attachments/106042227/AH-002.pdf> [letzter Zugriff 15. Juli 2021].

DNB (2020b): Liste der normierten Besetzungsangaben. <https://wiki.dnb.de/download/attachments/106042227/AH-001.pdf> [letzter Zugriff 15. Juli 2021].

IdRef (2021): Bibliographie zu den Suchen "Instruments de musique électronique" <https://www.idref.fr/027235122> sowie "Musique électroacoustique" <https://www.idref.fr/029771471#> [letzter Zugriff 25. November 2021].

IRCAM (1979-1991, 1999-2010): Programmhefte, Paris.

Automatisierte Extraktion und Klassifikation von Variantenschreibungen historischer Berufsbezeichnungen in seriellen Quellen des 16. bis 20. Jahrhunderts

Moeller, Katrin

katrin.moeller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Germany

Goldberg, Jan Michael

jan.goldberg@wiwi.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Germany

Einleitung

Berufsangaben kommen in sehr vielen historischen Quellen vor, besonders in seriellen Quellen wie Kirchenbüchern, Adressbüchern, Mieter-, Einwohner- und Bürgerlisten etc. Für eine Vielzahl von Forschungsgebieten bildet nicht nur die Standardisierung, sondern vor allem eine ordnende Klassifikation dieser Nennungen eine zentrale Voraussetzung zur tatsächlichen Analyse von Berufen nach verschiedenen möglichen Ordnungsprinzipien. Einen Ansatz bildet die „Ontologie der historischen Amts- und Berufsbezeichnungen (OhdAB)“ (Moeller 2019, Moeller 2021), die Amts- und Berufstätigkeit des 16. bis 20. Jahrhunderts klassifiziert. Der Vortrag beschäftigt sich mit der automatisierten Zuordnung von Variantenschreibungen von Amts- und Berufsbezeichnungen zu den Berufsgattungsnamen dieses Klassifikationsansatzes und damit zu einer verknüpften Klassifikation nach Tätigkeitskonzepten und Anforderungsniveaus, die auf der Methodik der Klassifikation der Berufe 2010/2020 aufbaut (Bundesagentur für Arbeit 2021). Dabei bleibt es unberücksichtigt, ob diese Varianten aus fehlerhaften Lese- und Schreibprozessen durch Mensch oder Maschine bzw. aus Variantenschreibungen der Quellen resultieren. Im Mittelpunkt des Beitrags steht der Algorithmus zur Identifizierung von Berufsbezeichnungen in strukturierten Quellen. Entwickelt wurde eine Vorgehensweise des Machine Learnings zur Erkennung von Variantenschreibungen, die im Vortrag vorgestellt wird. Diese besteht aus einem komplexen Workflow eines automatisierten Preprocessings zur Identifizierung bzw. Separierung der eigentlichen Berufsangaben und einer auf einem Algorithmus beruhenden Zuordnung unbekannter Varianten zur Klassifikation. Dieser Algorithmus wurde auf der Basis bereits zur Klassifikation (OhdAB) zugeordneter Varianten entwickelt und trainiert. Am Beispiel eines unbereinigten und stark heterogenen Datensatzes des Vereins für Computergenealogie

(Verein für Computergenealogie 2021) wurde eine Erkennungsrate von 75 Prozent der Berufsangaben ermittelt, wobei nur fünf Prozent fehlerhaften Zuordnungen zu identifizieren sind.

Berufsangaben in genealogischen Quellen, Preprocessing-Workflow

Amts- und Berufsangaben kommen in halbstrukturierter Form in zahlreichen historischen und modernen Quellen vor. Das Auslesen dieser Information aus einer begrenzten und bereits vortexturierten Textmenge bildet besondere Anforderungen an Verfahren des Natural Text Processings ab, da hier andere sprachanalytische Analyseformen zur Anwendung kommen (können) als in Volltexten. Demgegenüber ist die Zahl der Varianten besonders bei dieser Textsorte nicht nur durch Schreibvarianten der Quellen, sondern durch zahlreiche maschinelle Erhebungsverfahren zunehmend auch von OCR- oder HTR-Erkennungsfehlern und vor allem durch Abkürzungen geprägt. Zudem stammen viele maschinenlesbare Massendaten aus der Community der Citizen Sciences, die durch zusätzliche Aufnahme- und Eintragsbesonderheiten gekennzeichnet sind. Der im Beispiel verwendete Testdatensatz repräsentiert einen besonders stark verunreinigten Datensatz des Vereins für Computergenealogie.

Der wesentliche Vorteil dieser Quellen speist sich jedoch aus der bereits entitätsspezifisch strukturierten Datenmenge, da die aufzeichnenden Personen der Vergangenheit spezifische Vorstellungen von Amt, Beruf bzw. „Berufsstand“ wiedergaben. Dennoch gibt es auch in diesen Quellen eine begrenzte Menge weiterer Informationen, die zum Teil historische Auffassungen vom Berufsstand (z. B. Verwandtschaftsbeziehungen von Frauen und Kindern, Familienstand, Renten- und Altenteilbezüge, Kirchen-, Amts- und Ehrenvorstände etc.) zum Teil aber auch weniger reflektierte Aufnahmepraktiken oder fehlerhafte Einträge wiedergeben.

Keineswegs können solche Angaben in historischen Quellen jedoch wie in der Problemerkennung nach Rahm und Do (Rahm / Do 2000: 3f.) lediglich als Einquellenprobleme auf einem Level einzelner Instanzen (Berufsangabe) gekennzeichnet und aus der Analyse ausgeschlossen werden. Wie gezeigt, ist für historische Daten dagegen ein kontextualisierender Begriff des Berufsstandes wichtig, der fehlerhafte Einträge erst nach der Zuordnung und Klassifikation bereinigt. Die Angabe des Rechtsstatus oder Familienstandes kann eine Person in ihrem Stand ebenso adäquat beschreiben, während eine Ortsangabe nur eine in das falsche Datenfeld eingetragene Information repräsentieren kann. Dies berücksichtigt die historische Klassifikation OhdAB, indem sie heutige Berufsgegenstände von anderen Entitäten separiert, beide Formen jedoch ordnet und analysiert. Zur Lösung dieser qualitativen Probleme schlagen Müller und Freytag (Müller / Freytag 2003: S. 10-13) einen vierstufigen Prozess der Datenbereinigung vor. An dessen Beginn steht ein Datenaudit (*data auditing*), in welchem die Daten geparkt und analysiert werden. Dadurch werden syntaktische Anomalien erkannt, die es anschließend zu bearbeiten gilt. Dazu wird in einem zweiten Schritt der Ablauf der Datenbereinigung spezifiziert (*workflow specification*). Dabei kann die Behebung syntaktischer Fehler im Nachhinein wiederum andere Anomalien sichtbar machen. Die nachfolgende Durchführung der Datenbereinigung (*workflow execution*) steht im Konflikt zwischen einer möglichst passenden Korrektur und einer akzeptablen Laufzeit. Manuelle Nacharbeit ist zu vermeiden, da diese Ressourcen binden. Eine nicht-automatisierte Kontrolle findet allerdings in einem vierten Schritt statt (*post-processing and*

controlling). Hierfür wird mit dem Beitrag ein konkreter Workflow zur Extraktion von Berufen und Zuweisung der Berufsklassifikation vorgestellt.

Diese Datenbereinigung und das Preprocessing bleibt selbst bei den strukturierten Angaben komplex, zeigt aber durchaus Verbesserungspotential beim Datenmatching. Spezifische Problemlagen der Berufsbezeichnungen boten neben den mehr oder weniger spezifischen Abkürzungen vor allem die Angaben von mehreren Amts- und Berufsbezeichnungen in verschiedenen Sinnkonstruktionen. So können Berufsangaben immer wieder „paarig“ genannt werden, wie der Beruf des Gold- und Silberschmieds und damit einen gemeinsamen Berufsgattungsnamen repräsentieren oder eben auch Reihungen von verschiedenen Berufsamen enthalten, die nacheinander klassifiziert werden müssen. Gleichzeitig wurden temporale Hinzufügungen, Präzisierungen zum konkreten Berufs- oder Arbeitsort, Firmen- oder Einheitsangaben, Besitzinformationen etc. oder fremdsprachliche Angaben identifiziert. Durch das Konzept des Berufsstandes spielen zudem Angaben zum Familienstand, zur Rolle innerhalb der Familie (arbeitende Ehefrau, Witwe oder Kinder), Rechtsinformationen sowie Standestitel eine wichtige Rolle. Daneben gibt es über die zahlreichen ergänzenden Informationen hinaus immer wieder auch falsch angeordnete Entitäten (Namen oder Ortsangaben, weitere Eigennamen oder auch Quellenbezeichnungen).

Aufgrund der qualitativen Datenanalyse wurde ein Kanon von Separatoren und Trennzeichen ermittelt, der verschiedene Informationsketten „anzeigen“, markiert und je nach Qualität auszeichnet, separiert oder löscht. So wurden bspw. über lokale Präpositionen Ortsangaben, über temporale Präpositionen zeitliche Angaben zum Beruf separiert und normiert. Andererseits bildete ein Vokabular zu verschiedenen Formen von Verwandtschaftsbezeichnungen oder des Familienstandes die Grundlage zur separierten Definition der Familienrolle, die für die Einordnung der eigentlichen Ausübung von Tätigkeiten zentrale Informationen abbildet. Im Vortrag sollen die entsprechenden Möglichkeiten dazu kurz systematisch skizziert und in ihrer Funktionalität dargestellt werden.

Algorithmus zur Variantenzuordnung

Da Berufsangaben Strings im Sinne einer semantischen Zeichenkette darstellen, können String-Matching-Algorithmen zur Erkennung einer unscharfen Übereinstimmung auf sie angewendet werden. Die Ähnlichkeit von Strings kann über verschiedene Maße ausgedrückt werden. In der historischen Linguistik stellt die Levenshtein-Distanz eine geeignete Möglichkeit dar, die infrage kommende Beziehung zwischen Wörtern aufzuzeigen. Die Herausforderung, zwei Schreibvarianten desselben Wortes zu erkennen, ist ähnlich gelagert wie die Erkennung einer möglichen linguistischen Verwandtschaft zwischen zwei Wörtern.

Zunächst sollen möglichst viele Berufsangaben den richtigen Entitäten, im Weiteren „Klassen“, zugeordnet werden (True Positiv = TP). Ein Berufsgattungsnamen stellt dabei eine Klasse dar; die bekannten Schreibweisen (Varianten) wiederum sind die Eigenschaften. Eine Übersicht über die verwendeten Begrifflichkeiten ist, insbesondere für die multiple Verwendung der Klassifizierung / Klassifikation, in Abbildung 1 ersichtlich.

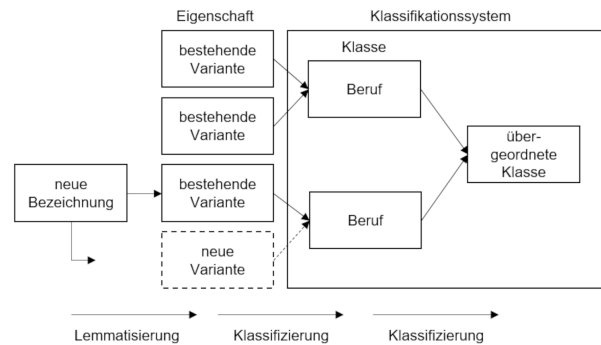


Abb. 1: Begriffe und Zusammenhänge des Algorithmus. [Goldberg / Moeller 2021]

Eine Erhöhung der TP-klassifizierten allein geht jedoch oftmals auch mit der Erhöhung von FP-Klassifizierungen (False Positiv) einher. Aus diesem Grund wird nicht die Anzahl der TP-Klassifizierungen optimiert, sondern das F_1 -Maß als Ausweis dieser falsch zugeordneten Begriffe. Zudem soll die Klassifizierung automatisch geschehen; eine manuelle Überprüfung des Ergebnisses geschieht nicht. Das ist notwendig, um große Datenbestände in einer überschaubaren Zeit klassifizieren zu können. Da der Algorithmus insbesondere auf große Listen von Berufsangaben Anwendung finden soll, ist dessen Effizienz und somit die Laufzeit zu beachten. Der Algorithmus ist in einem Programmcode (Python basiert) umgesetzt worden, der in weiteren Applikationen eingebunden werden kann.

Nach der Bereinigung sind den Berufsangaben trotzdem noch keine Berufsgattungsnamen der OhdAB-Konkordanz zugeordnet. Die notwendige Zuordnung geschieht auf Basis der Eigenschaften der bestehenden Klassen. Darum findet ein Abgleich mit den vorhandenen Varianten der OhdAB statt. Eine Berufsangabe soll der Klasse zugeordnet werden, deren Zugehörigkeit am wahrscheinlichsten ist. Die Ähnlichkeit einer Berufsangabe zu den Eigenschaften (bestehende Varianten) einer Klasse (Beruf) wird dabei als Indikator für die Wahrscheinlichkeit einer korrekten Zuordnung (Normierung/Lemmatisierung¹) genutzt. Diese kann über einen Vergleich der Zeichenketten ermittelt werden. Jedoch muss nicht zwingend eine Lemmatisierung stattfinden: Wenn die Ähnlichkeit zu jeder Klasse so gering ist, dass eine korrekte Zuordnung unwahrscheinlich ist, kann kein Pendant gefunden werden.

Zeichenketten können auf verschiedene Arten verglichen werden. Kirby et al. empfehlen für die weitere Forschung eine Variation von verschiedenen Vergleichsmethoden (Kirby, 2015, S. 58). Dadurch, dass die Variante einer Normschreibweise der Konkordanz zugeordnet ist, ist auch ihre Zuordnung zu einer Berufsgattung der OhdAB eindeutig. Besteht keine Übereinstimmung mit einer Variante, so ist eine teilweise Übereinstimmung zu überprüfen. Daher wurden für die Entwicklung des Algorithmus eine Reihe verschiedener Ansätze ausgetestet, die einerseits eine 1:1 Zuordnung (unter Verzicht von Groß- und Kleinschreibung) sowie verschiedene Variationen der Levenshtein-Distanz und weiterer Iterationsschritte umfassen.

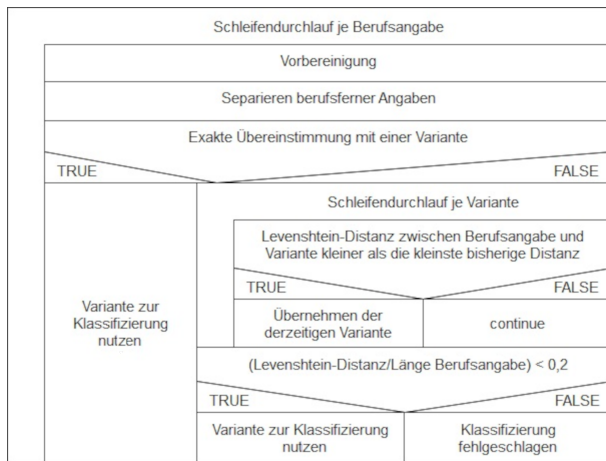


Abb. 2: Algorithmus, dargestellt in einem Nassi-Shneiderman-Diagramm. [Goldberg / Moeller 2021]

Im Vortrag möchten wir unser Vorgehen wie auch die ermittelten Testwerte vorstellen und genau erläutern (was hier aufgrund der verfügbaren Platzmenge nicht stattfinden kann). Zusammengefasst wird das F_1 -Maß optimiert, wenn eine relative Levenshtein-Distanz gewählt wird, Abkürzungen erweitert werden und erlernte neue Varianten im Anschluss nochmal mit allen Daten verglichen werden, die nicht lemmatisiert werden konnten. Der letztgenannte Aspekt des teil-maschinellen Lernens führt dazu, dass neue Varianten stetig hinzukommen können. Der Algorithmus wird auf die Testdaten (220.000 Berufsangaben) angewendet. 64 Prozent der beruflichen Bezeichnungen können direkt lemmatisiert und einer bestehenden Variante zugeordnet werden. Bei den weiteren wird eine Ähnlichkeitsanalyse durchgeführt, die in vier Prozent der Fälle ein Ergebnis erbringt. Insgesamt hat die Ähnlichkeitsanalyse nur eine vergleichsweise geringe Auswirkung. Relevanter ist der Einsatz der Bereinigung: Wird diese gänzlich ausgelassen, so können nur 58 Prozent der Angaben direkt in den Varianten erkannt werden.

	Direkt gefunden	Ähnlichkeitsanalyse	Nicht gefunden	Leere Bezeichnungen
mit Bereinigung (229.699 Angaben)				
Anzahl	147.781	9.674	68.955	3.259
Anteil	64,35 %	4,21 %	30,02 %	1,42 %
ohne Bereinigung (229.669 Angaben)				
Anzahl	131.064	9.160	86.344	3.101
Anteil	58,07 %	3,89 %	37,59 %	1,35 %

Tab. 1: Vergleich des Effektes der Bereinigung auf die Erkennung. [Goldberg / Moeller 2021]

Die durch die Ähnlichkeitsanalyse neu zugeordneten Berufsangaben können in der Variantenliste ergänzt werden. Dieses kann geschehen, indem die neuen Treffer direkt nach Erkennung in die Menge der Varianten eingehen oder aber alle nicht erkannten Bezeichnungen anschließend mit allen Treffern abgeglichen werden. Letzteres ist mit mehreren Iterationen denkbar. Hierbei zeigt sich, dass die nachfolgende, iterative Verarbeitung ein besseres

Ergebnis in Bezug auf das F_1 -Maß ergibt als die kontinuierliche Ergänzung (siehe Tabelle 2). Dabei ist der Lerneffekt größer, je mehr Berufsangaben verarbeitet werden, da die Chance steigt, dass eine ähnliche Bezeichnung auftritt. Bei einem Durchlauf mit jeder zehnten Datei wird noch keine zusätzliche Erkennung erreicht. Allerdings werden auch bei einer Verarbeitung aller Daten nur weitere 0,01 Prozent der Berufsangaben lemmatisiert. Dieses ist darauf zurückzuführen, dass bereits sehr viele Schreibversionen in den zugrundeliegenden Varianten der OhdAB abgedeckt sind (momentan ca. 200.000). Bei einer willkürlichen Halbierung der ursprünglichen Varianten steigt der Anteil der so zusätzlich erkannten Angaben deutlich um rund 9 Prozent (von 3,10 Prozent auf 12,01 Prozent). Werden diese lemmatisierten Varianten in einem zweiten Durchlauf zur Gesamtzahl der Varianten ergänzt, können weitere Berufsbezeichnungen lemmatisiert werden. Die TP-Rate jedoch ist etwas niedriger. Eine hohe FP-Rate in der ersten Ähnlichkeitserkennung führt tendenziell zu einer Fortführung von Fehlern.

	Direkt gefunden	Ähnlichkeitsanalyse	Nicht gefunden	Leere Bezeichnungen
mit Bereinigung (229.699 Angaben)				
Anzahl	147.781	9.674	68.955	3.259
Anteil	64,35 %	4,21 %	30,02 %	1,42 %
ohne Bereinigung (229.669 Angaben)				
Anzahl	131.064	9.160	86.344	3.101
Anteil	58,07 %	3,89 %	37,59 %	1,35 %

Tab. 2: Vergleich der Klassifikation unter Halbierung der zugrundeliegenden Berufsvarianten der OhdAB. [Goldberg / Moeller 2021]

Durch den Algorithmus – und dessen programmtechnische Umsetzung – wird in der Folge eine automatisierte Lösung zur Lemmatisierung deutschsprachiger Berufsangaben geboten. Mittels Variation der Ähnlichkeitsanalyse konnte zwar formal kein Optimierungsproblem gelöst werden; es hat sich aber gezeigt welche Faktoren das F_1 -Maß verschlechtern und welche es verbessern. Zudem ist es durch den Algorithmus möglich, berufsferne Angaben von der eigentlichen Bezeichnung des Berufs zu separieren.

Fußnoten

1. Der Begriff der Lemmatisierung wird hier als Zuweisung einer Normschreibung zu verschiedenen Schreibvarianten des gleichen Berufsgattungsnamens verstanden.

Bibliographie

- Bundesagentur für Arbeit** (2021): *Klassifikationen der Berufe - Statistik der Bundesagentur für Arbeit*. Nürnberg. [online].
- Bundesagentur für Arbeit** (2011): *Klassifikation der Berufe*. Nürnberg 2010. Bd 1 (2011): *Systematischer und alphabetischer Teil mit Erläuterungen*.
- Church of Jesus Christ of Latter-day Saints** (2019): *The GEDCOM Standard*. Release 5.5.1. 2019.

Theresa Cosca / Alissa Emmel (2010): "Revising the Standard Occupational Classification system for 2010". In: *Monthly labor review* 133, 32–41. PDF. [online] 320603628.

Jyldyz Djumaliev / Antonio Lima / Cath Sleeman (2018): *Classifying Occupations According to Their Skill Requirements in Job Advertisements*. [online].

Hyukjun Gweon / Matthias Schonlau / Lars Kaczmirek / Michael Blohm / Stefan Steiner (2017): "Three Methods for Occupation Coding Based on Statistical Learning". In: *Journal of Official Statistics* 33, H. 1, 101–122. DOI: 10.1515/jos-2017-0006 130422746.

Jan Michael Goldberg / Katrin Moeller (erscheint 2022): "Automatisierte Identifikation und Lemmatisierung historischer Berufsbezeichnungen in deutschsprachigen Datenbeständen", in: *Zeitschrift für digitale Geisteswissenschaften* (im Druck).

J. Tuomas Harviainen / Bo-Christer Björk (2018): "Genealogy, GEDCOM, and popularity implications". In: *Informaatio tutkimus* 37, H. 3, S. 4–14. Artikel vom 29.10.2018. DOI: 10.23978/inf.76066 366701630

Graham Kirby / Jamie Carson / Fraser Dunlop / Chris Dibben / Alan Dearle / Lee Williamson / Eilidh Garrett / Alice Reid: "Automatic Methods for Coding Historical Occupation Descriptions to Standard Classification". In: *Population Reconstruction*, Hg. von Gerrit Bloothoof / Peter Christen / Kees Mandemakers / Marijn Schraagen. Cham / Heidelberg u.a. 2015, S. 43–60.

Thomas Krause (2012): *Entwurf und Implementierung einer effizienten Dublettenerkennung für große Adressbestände*. Köln. URN: urn:nbn:de:hbz:832-epub-3667.

Marco H. D. van Leeuwen / Ineke Maas / Andrew Miles (2002): *HISCO. Historical International Standard Classification of Occupations*, Leuven.

Vladimir Iosifovič Levenštejn (1966): "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals". In: *Soviet Physics- Doklady* 10, 707–710. 129482234.

Katrin Moeller (2019): "Standards für die Geschichtswissenschaft! Zu differenzierten Funktionen von Normdaten, Standards und Klassifikationen für die Geisteswissenschaften am Beispiel von Berufsklassifikationen". In: *Aufklärungsforschung digital. Konzepte, Methoden, Perspektiven*. Hg. von Jana Kittelmann und Anne Purschwitz, Halle, 17–43.

Katrin Moeller (2021): "Ontologie historischer, deutschsprachiger Berufs- und Amtsbezeichnungen". In: *Websites des Historischen Datenzentrums Sachsen-Anhaltgeschichte.uni-halle.de*. 13.07.2021. [online].

Heiko Müller / Johann-Christoph Freytag: *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. Berlin 2003. PDF. [online] 496492772.

Wiebke Paulus / Britta Matthes (2013): "Klassifikation der Berufe 2010 - Struktur, Codierung und Umsteigeschlüssel". In: *FDZ-Methodenreport*. Hg. von Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung. Nürnberg. [online].

Michael Piotrowski (2012): "Natural Language Processing for Historical Texts". In: *Synthesis Lectures on Human Language Technologies* 5, H. 2, S. 1–157. 616519060

Erhard Rahm / Hong Hai Do (2000): "Data Cleaning: Problems and Current Approaches". In: *Bulletin of the Technical Committee on Data Engineering* 23, H. 4, S. 3–13. URN: urn:nbn:de:bsz:15-qucosa2-329680.

Back 'em up Computerspiele als Objekte kulturellen Erbes

Schneider, Sophie

schneiso@hu-berlin.de
HU Berlin, Germany

Zweifellos sind digitale Spiele Objekte kulturellen Erbes, deren Erforschung innovative Perspektiven auf gesellschaftliche und medientechnologische Entwicklungen schafft (Arndt et al. 2020: 346; Jett et al. 2016: 505; Lee et al. 2015a: 2621f.; Lee et al. 2015b: 9; McDonough et al. 2010: 100). Flüchtig als Massen- oder Unterhaltungsmedium deklariert, verbirgt sich hinter dem Computerspiel womöglich ein größeres wissenschaftliches Potenzial, als es ihm häufig zugesprochen wird (Loebel/Hahn 2020; Seadle 2008: 6). Lautete das Motto der 7. Jahrestagung des DHD-Verbands bereits „Spielräume“, so sucht man nach Beiträgen, die inhaltlich digitale Spiele per se in den Fokus nehmen, weitestgehend vergeblich (vgl. Schöch 2020). Dabei könnten die Digital Humanities als interdisziplinäres „großes Zelt“ der verschiedenen digital-geisteswissenschaftlichen Fächer von einer differenzierten Auseinandersetzung mit Computerspielen durchaus profitieren. Die zentrale Bereitstellung von Ressourcen in Form digital verfügbar gemachter Computerspiele als Datengrundlage etwa, oder die Übernahme von Verfahren, die neue Perspektiven auf einen digitalen Text, ein digitales Bild-, Ton- oder Videodokument schaffen, wären wegweisend. Der vorliegende Beitrag hinterfragt daher die Rolle, die Computerspiele aktuell in den DH einnehmen - nicht als Ergebnis oder Methode (z.B. Gamification), sondern als Gegenstand digitaler Forschungen. *Wie können digitale Spiele als Ressourcen erschlossen und leichter zugänglich gemacht werden?* Welche Infrastrukturen und Tools werden benötigt? Welche neuen Forschungsfragen und Aufgabenfelder lassen sich hieraus für die Digital Humanities generieren?

Computer- bzw. Videospiele¹ sind interaktive, multimediale und -dimensionale Objekte, bei welchen insbesondere die interdependente Einheit aus Hard- und Software, der digitale Charakter („digital born“) und der schnelle Wandel des Mediums berücksichtigt werden muss. An dem Erhalt und der Bereitstellung von Videospielbeständen sind weltweit Bibliotheken, Archive und Museen (z.B. die Mitgliedsinstitutionen der European Federation of Games Archives, Museums and Preservation Projects, vgl. EFGAMP 2017) sowie vereinzelt spezialisierte Arbeitsgruppen und zeitlich begrenzte Projekte (z.B. die Game Research Group (GAMER) der University of Washington, vgl. GAMER o.J., oder das Projekt Preserving Virtual Worlds, vgl. McDonough et al. 2010) beteiligt. Digitale Spiele als „nicht-textuelle Materialien“ stellen die Gedächtniseinrichtungen jedoch vor neuartige Herausforderungen (vgl. Deutsche Forschungsgemeinschaft 2018; Kommission Zukunft der Informationsinfrastruktur 2011), die zunächst in die Bereiche Standardisierung von und Vernetzung mit Metadaten, Erschließung, Präsentation sowie Erhalt und Nachnutzung zusammengefasst wurden.² Der Beitrag wertet die bisherigen Bemühungen einer Annäherung an Lösungsansätze in jenen Aufgabenbereichen für das Medium Computerspiel aus.

Standardisierung und Vernetzung

Existierende Metadatenschemata wie FRBR oder CIDOC CRM erlauben keine optimale Darstellung videospieldspezifischer Attribute und Relationen und ihre Anwendung ist jeweils nur in konkreten Nutzungsszenarien sinnvoll, nicht jedoch in Gestalt eines universalen Standards (Hoffmann 2019; Jett et al. 2016; Lee et al. 2015b). Dies lässt sich beispielsweise in der begrenzten Anzahl an zur Verfügung stehenden Elementen begründen, mit welchen insbesondere die verschiedenen Ebenen der Granularität (z.B. in Bezug auf verschiedene Ausgaben eines Spiels) nicht abbildbar sind. Sowohl Forschende als auch Nutzende sind allerdings besonders an den Unterschieden zwischen verschiedenen Versionen von Computerspielen (darunter auch Modifikationen, Updates und herunterladbare Inhalte) oder Plattformen sowie der Kompatibilität von einzelnen Hard- und Softwarekomponenten interessiert (Arndt et al. 2020; Kaltman et al. 2016; Lee et al. 2015a; McDonough et al. 2010: 98-104). Insofern könnte eine Standardisierung mehr Transparenz und Nachnutzbarkeit verschaffen, selbst wenn sich die konkreten Anforderungen der einzelnen Einrichtungen und Zielgruppen aufgrund unterschiedlicher Zielsetzungen in der Beschreibung mit Metadaten voneinander unterscheiden. In Bezug auf digital distribuierte Spiele bzw. Spielinhalte entwickeln sich zudem neue Informationsbedarfe zur rechtlichen Lage und den Zugriffsmöglichkeiten/-bedingungen: "[i]nformation regarding limited install activation; persistent online authentication; software tampering; and rights to copy, share, and resell the game [...]" (Lee et al. 2015a: 2618f.).

Die entsprechenden Zielgruppen interessieren sich oftmals auch für zugehörige oder ähnliche Spiele, diese Relationen müssen sich ebenfalls in der Metadatenstruktur wiederfinden (Lee et al. 2015a; Lee et al. 2015b; Ryan et al. 2015b). Der Einsatz von Semantic Web-Technologien wie Linked Open Data (LOD) erscheint besonders vielversprechend, da die Metadatensätze nicht statisch festgelegt, sondern erweiter- und anpassbar sind und mit anderen Ontologien (z.B. SKOS-, OWL-Vokabulare) verknüpft werden können (Kaltman et al. 2016). Eine Implementierung sollte Funktionen wie das Definieren einer Beziehung zwischen zwei Objekten mittels URI/IRI und verschiedene Zugriffsmöglichkeiten auf die Daten (APIs, standardisierte Abfragesprachen/SPARQL-Endpoints, Export und ggf. Import von RDF/XML-Dateien) anbieten. Geeignet wären hierfür beispielsweise "Triple Stores", die das Speichern und Abrufen von RDF-Tripeln erleichtern sollen.

Erschließung

Die Möglichkeiten der inhaltsbasierten Erschließung von Videospielen und vergleichbaren interaktiven Objekten gehen jedoch weit über die reine Beschreibung mit Metadaten hinaus. Das Identifizieren verwandter Objekte kann mittels automatisierter und ähnlichkeitsbasierter Erschließungsmethoden („analogical reasoning“), z.B. über Clusteringverfahren, erfolgen. Ryan et al. entwerfen das Konzept einer Game Relatedness in Anlehnung an die Semantic Relatedness aus dem Natural Language Processing. Unter Zuhilfenahme der Latent Semantic Analysis (LSA) überprüfen sie die Hypothese, dass Menschen Spiele bevorzugen, die sich ähnlich bzw. die miteinander verwandt ("related") sind (Ryan et al. 2015b). Die Analyse von Zhang et al. ist im Vergleich deutlich detaillierter und nimmt einzelne Momente in Videospielen in den Fokus. Dafür wird noch ein weiterer Schritt notwendig: das Crawling, d.h. in diesem Fall das automatische (Ab-)Spielen von Videospielen sowie das regelmäßige Abspeichern von Momenten

als Screenshots (inkl. Metadaten). Nach dem Zerlegen eines Videospiels in die diversen Momente werden die Bilder mithilfe von Deep Neural Networks in ein Vector Space Model überführt. Damit soll das Auffinden „ähnlicher“, hier: zeitlich (fester Schwellenwert) oder räumlich (gleicher Raum oder gleiches Level) nah beieinander liegender, Momente im Videospiel erleichtert werden (Zhang et al. 2018).

Es reicht nicht aus, nur die Spiele an sich zu erschließen und zu erhalten, auch deren Kontext muss zu einem späteren Zeitpunkt noch rekonstruierbar sein (Lange 2020; McDonough et al. 2010: 99-101). Nur so kann ein langfristiger Zugang zu den Technologien und zum Gameplay gesichert werden. Eine besondere Herausforderung stellen sogenannte Massively Multiplayer Online Games (MMOGs) dar, in denen die Kommunikation und Interaktion mit anderen Spielern zu einem späteren Zeitpunkt nur schwer nachzuvollziehen ist.

Folksonomies und Social Tagging könnten erfolgversprechende Mechanismen für die Einbindung der entsprechenden Communities in die Erschließung sein und bedeuten damit zugleich eine Entlastung im Bereich anderer Erschließungsmethoden (Lew et al. 2006: 13f.; Szuban 2018). Gleichzeitig stellt sich hierbei stets die Frage nach der Qualität und dem Umfang nutzergetriebener Erschließung. Eine Kombination aus hierarchischen und kollaborativen Erschließungsmethoden scheint derzeit am erfolgversprechendsten zu sein.

Präsentation

Vor dem Hintergrund eines "visualisitic turn" (Sachs-Hombach 2014: 97f.) erfahren visuelle Suchfunktionen bzw. -schnittstellen im Multimedia Information Retrieval (MIR) einen Bedeutungszuwachs (Lew et al. 2006). Zhang et al. stellen einen Prototyp³ für eine visuelle Suchmaschine vor, mit dem eine Suche anhand von Bildern (beispielsweise Screenshots von Momenten in einem Videospiel) oder auch anhand des Memory States durchgeführt werden kann. Während man darin mit einem konkreten visuellen Beispiel suchen kann („Query by Example“), erlaubt die Anwendung GameSage⁴ die textuelle Beschreibung des gesuchten Spiels bzw. der Idee eines Spiels (Ryan et al. 2015c). Eine wichtige Funktion für visuelle Suchmaschinen ist das Relevance Feedback, welches es Suchenden ermöglicht, die Relevanz der gefundenen Ergebnisse zu beurteilen. Dieses Feedback fließt im Anschluss in die Gesamtberechnung des Rankings mit ein (Lew et al. 2006; Rochio/Salton 1965; Zhang et al. 2018).

Das Auffinden von Videospielen ist jedoch nicht auf die zielgerichtete Suche beschränkt, auch das Browsing und Visualisierungen können die Exploration von Objekten und Inhalten unterstützen. Ein wesentlicher Aspekt bei der Visualisierung ist das Hineinzoomen in bzw. das Herauszoomen aus Sammlungsinhalten(n), sodass sich Nutzende auf und zwischen verschiedenen Abstraktionsebenen frei bewegen können (Brüggemann et al. 2020; Shneiderman 1996). Das Ideal einer räumlichen wie zeitlichen Flexibilität lässt sich auch auf die Erkundung von Videospielbeständen übertragen: "Freeze framing, fast forwarding, rewinding, zooming in and out, altering playback speed, captioning, annotating and subtitling all add plasticity and interpretative opportunity." (Newman 2019: 13) Das Clustering (hier: von Bildern) ermöglicht eine natürlichere Betrachtung des Bestandes als simple Ranking-Listen (Lew et al. 2006: 9). Drei Beispiele für "large-scale visualizations of nearly 12,000 digital games" (Ryan et al. 2015a: 1), in denen Clustering-Algorithmen auf ein LSA-Model angewendet werden, sind GameGlobs⁵, Game-

Tree⁶ und GameSpace⁷. Während alle drei Visualisierungen auf den gleichen Daten beruhen, sorgt die Anwendung unterschiedlicher Clusteringverfahren und Visualisierungsansätze für verschiedenartige Einstiege (Ryan et al. 2015a).

Darüber hinaus können ganze Forschungsumgebungen oder Tools wie das Game and Interactive Software Scholarship Toolkit (GISST) in digitale Sammlungen integriert werden (Kaltman et al. 2017). Eine Einbettung von Spielausschnitten bzw. zugehörigen Zitierhinweisen⁸ oder die Bereitstellung der verschiedenen Bearbeitungs- und Analysefunktionen des GISST für digital verfügbare Ressourcen (Videospiele) wäre hierfür denkbar. Allgemein lässt sich ein hoher Bedarf an einer umfangreicheren Auswahl vergleichbarer Tools verzeichnen (Kaltman et al. 2017: 4; Zhang et al. 2018: 10). Roeder und Rettinghaus widmen sich der Frage, welche Herausforderungen sogenannte Diskmags an digitale Editionen stellen. Dadurch wird einerseits die Diversität der vorhandenen kontextuellen Materialien zum Objekt Videospiel deutlich sowie andererseits die Notwendigkeit, diese nicht nur zu erschließen, sondern auch in einer Form zu präsentieren, die ihrer zunehmend interaktiven und multimedialen Natur gerecht wird (Roeder/Rettinghaus 2020).

Erhalt und Nachnutzung

Um Computerspiele über einen längeren Zeitraum zu bewahren, werden verschiedene Strategien verfolgt (Lange 2020; McDonough et al. 2010: 58-88; Seadle 2008). Die naheliegendste Option besteht hierbei darin, "die historische Hardware so lange wie möglich am Leben zu erhalten" (Lange 2020: 127). Da viele der verwendeten Speichermedien mit der Zeit zunehmend Datenverluste aufweisen, die Hardware nur über eine begrenzte Haltbarkeit verfügt und auch das Fehlen relevanter Systemkomponenten ein Abspielen ggf. verhindern kann, stellt diese Strategie keine dauerhafte Lösung dar. Die Bewahrung und damit auch die Selektion der für relevant befundenen Werke erfolgte bislang größtenteils über die Gaming-Community selbst, sodass hierdurch ebenfalls keine Nachhaltigkeit garantiert werden kann. Eine andere Möglichkeit ist die Migration, d.h. der Quellcode der Werke wird an aktuell verbreitete und damit besser zugängliche Systeme angepasst. Die am meisten genutzte Strategie vor dem Kontext der Erhaltung von Computerspielen ist die Emulation. Dabei handelt es sich grob gesagt um eine Simulation von Hardware mittels Software. Gleichzeitig sollte jedoch nicht vergessen werden, dass auch Emulatoren und vergleichbare Software einem zeitlichen Verfall unterliegen und sich in der Entwicklung und Archivierung nicht als besonders nachhaltig erwiesen haben.

Die Referenzier- bzw. Zitierbarkeit von Videospielen spielt ebenfalls eine große Rolle für deren Nachnutzung (Arndt et al. 2020; Kaltman et al. 2017; Kaltman et al. 2021). Wissenschaftler*innen zitieren Videospiele größtenteils nicht nach existierenden Standards und die bislang angewendeten Praktiken sowie erarbeiteten Empfehlungen sind sehr heterogen. Eine Herausforderung (z.B. für den Einsatz persistenter Identifikatoren) besteht darin, dass der benötigte Detailgrad der Referenz vom individuellen Forschungsinteresse abhängig ist und somit idealerweise eine eindeutige Zitierfähigkeit auf verschiedenen Abstraktionsebenen geschaffen werden müsste. Zusätzlich besteht die Möglichkeit, einzelne Momente aus Videospielen zu zitieren (Newman 2019).

Einer direkten Nachnutzung von digitalen Spielen - sowohl für den privaten Gebrauch als auch für wissenschaftliche Zwecke - steht derzeit häufig das Urheberrecht bzw. Copyright im Weg, welches diese als Werke geistiger Schöpfung schützen soll

(Lange 2020; McDonough et al. 2010: 52-57; Newman 2019; Seadle 2008). Die Schwierigkeit, Videospiele definitorisch eindeutig in eine der bereits existierenden Medienkategorien einzuordnen, stellte dabei bislang ein Problem dar, denn geltende Ausnahmeregelungen konnten nicht auf diese Objekte bezogen werden. Positive Entwicklungen zeichnen sich beispielsweise vor dem Hintergrund der Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes (19/27426) ab. Gemäß Artikel 6 DSM-RL soll es „Einrichtungen des Kulturerbes [erlaubt sein], Vervielfältigungen zum Zweck der Bestandserhaltung vorzunehmen“ und das „unabhängig von Format oder Medium“ (Deutscher Bundestag 2021: 98f.). Dies lässt sich jedoch nicht in gleicher Weise auf den Verleih oder die Exposition von digitalen Werken oder Ausschnitten übertragen, sodass die Legalität einer Weiterverwendung über technische (Kopierschutz) und rechtliche Hürden (z.B. Endnutzervereinbarungen oder lizenzierte Inhalte) hinweg oftmals unklar bleibt.

Ausblick

Computerspiele verbinden beispielhaft die beiden Schwerpunkte Technik und Kultur, ohne die es keine Digital Humanities gäbe. Als Forschungsobjekte dürfen sie daher gerade vor diesem disziplinären Hintergrund nicht außen vor gelassen werden. Zum jetzigen Zeitpunkt ist das Interesse an (Computer-)Spielen auch in den digitalen Geisteswissenschaften spürbar gestiegen und Netzwerke werden initiiert (vgl. DHd 2014). Die Bearbeitung gemeinsamer bzw. sich überschneidender Fragestellungen⁹ schafft dabei nicht nur einen nährhaften Boden für Kollaboration und Vernetzung, sondern ermöglicht zudem die extrinsische Reflexion etablierter Arbeitsweisen auf beiden Seiten. Das digitale Spiel kann im Sinne der vorangegangenen Ausführungen als Forschungsdatum verstanden werden, dessen Erhalt auch für die Nachvollziehbarkeit sowie Reproduzierbarkeit wissenschaftlicher Studien sichergestellt werden muss. Hierfür scheint eine ausführliche Thematisierung von digitalen Spielen im Rahmen der Nationalen Forschungsdateninfrastruktur (NFDI) sinnvoll, insbesondere innerhalb der Konsortien NFDI4Culture und 4Memory. Die darin formulierten Ziele und Aufgabenfelder wie die Standardisierung oder langfristige Bewahrung und Nachnutzbarmachung von Forschungsdaten sowie der Auf- bzw. Ausbau der dafür benötigten Infrastrukturen sind mit den hier dargelegten Aufgabenbereichen für Computerspiele auffällig deckungsgleich (NFDI4Culture o.J.; 4Memory 2020).

Es wurden einige Ausgangspunkte aufgezeigt, deren Ausdifferenzierung einen ausführlicheren (geistes-)wissenschaftlichen Diskurs zum digitalen Spiel nach sich ziehen könnte. Die kurze Haltbarkeit sowie enge industrielle Anbindung stellen das Computerspiel als ein Objekt kulturellen Erbes vor eine intensive Bewährungsprobe. Gerade in Deutschland scheint es bereits an zentralen, digitalen Infrastrukturen, Ressourcen- und Informationsangeboten sowie groß angelegten Forschungs- und Förderinitiativen hierfür zu fehlen. Daher besteht Handlungsbedarf seitens der Digital Humanities sowie einzelner Gedächtnisinstitutionen, sich der Entwicklung derartiger Angebote anzunehmen und den Austausch in diese Richtung weiter anzustoßen.

Fußnoten

1. Die Begriffe haben sich im Laufe der Zeit in ihrer Bedeutung angenähert und werden heute überwiegend synonym verwendet, vgl. Koubek 2020.
2. Die hier vorgestellten Bereiche sind weder als holistisch noch als disjunkt zu betrachten, sie sollen lediglich einen ersten Anhaltspunkt konstituieren.
3. Vgl. https://drive.google.com/file/d/1eGkx1mh_Nry1hH-P2S4j0p4HVL2pMCzTy/view [letzter Zugriff 07. Juni 2021].
4. Vgl. <http://gamecip-projects.soe.ucsc.edu/gamesage> [letzter Zugriff 08. Juni 2021].
5. Vgl. <http://gamecip-projects.soe.ucsc.edu/gameglobs/viz> [letzter Zugriff 08. Juni 2021].
6. Die Anwendung konnte nicht aufgefunden werden (Stand: 08.06.2021), vgl. daher Ryan et al. 2015a.
7. Vgl. <http://gamecip-projects.soe.ucsc.edu/gamespace/> [letzter Zugriff 08. Juni 2021].
8. Für eine interaktive Demonstration der Unterstützung von Videospielzitationen in wissenschaftlichen Texten unter Verwendung des GISST vgl. Kaltman et al. 2021, Absatz 8.
9. Exemplarisch für die Annäherung an derartige interdisziplinäre Fragestellungen können die Aktivitäten und Beiträge des *Arbeitskreises Geschichtswissenschaft und Spiele* herangezogen werden, vgl. Arbeitskreis Geschichtswissenschaft und Spiele 2021.

Bibliographie

- Arbeitskreis Geschichtswissenschaft und Spiele** (2021): „gespielt | Blog des Arbeitskreises Geschichtswissenschaft und Digitale Spiele“ [online] <https://gespielt.hypotheses.org/> [letzter Zugriff 26. November 2021].
- Arndt, Tracy / Freybe, Konstantin / Lahmann, André / Roth, Martin** (2020): „Reference Evil -Bibliographische Herausforderungen bei Videospielen“, in: *Bibliotheksdienst* 54(5): 345-362 <https://doi.org/10.1515/bd-2020-0045>.
- Brüggemann, Viktoria / Bludau, Mark-Jan / Dörk, Marian** (2020): „The Fold: Rethinking Interactivity in Data Visualization“, in *DHQ: Digital Humanities Quarterly* 14(3): <http://www.digitalhumanities.org/dhq/vol/14/3/000487/000487.html> [letzter Zugriff 08. Juni 2021].
- Deutsche Forschungsgemeinschaft** (2018): „Erschließung und Digitalisierung“, in: *DFG Förderung von Informationsinfrastrukturen für die Wissenschaft. Ein Positionspapier der Deutschen Forschungsgemeinschaft* 20-27 https://www.dfg.de/download/pdf/foerderung/programme/lis/positionspapier_informationsinfrastrukturen.pdf [letzter Zugriff 08. Juni 2021].
- Deutscher Bundestag** (2021): „Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes“ <https://dserver.bundestag.de/bt-d/19/274/1927426.pdf> [letzter Zugriff 08. Juni 2021].
- DHd** (2014): „AG Spiele“ [online] <https://dig-hum.de/ag-spiele> [letzter Zugriff 26. November 2021].
- EFGAMP** (2017): „Our digital gaming heritage. Our responsibility“ [online] <https://efgamp.eu/> [letzter Zugriff 26. November 2021].
- 4Memory** (2020): „Our objectives: LINKAGE“ [online] <https://4memory.de/linkage/> [letzter Zugriff 26. November 2021].
- GAMER** (o.J.): „Welcome To the UW Game Research Group“ [online] <https://gamer.ischool.uw.edu/> [letzter Zugriff 26. November 2021].
- Hoffmann, Tracy** (2019): „Developing a Mediated Vocabulary for Video Game Research“, in: *Proceedings of the Doctoral Symposium on Research on Online Databases in History (RODBH 2019)* 26-36.
- Jett, Jacob / Sacchi, Simone / Lee, Jin Ha / Clarke, Rachel Ivy** (2016): „A Conceptual Model for Video Games and Interactive Media“, in: *Journal of the Association for Information Science and Technology* 67(3): 505-517 <https://doi.org/10.1002/asi.23409>.
- Kaltman, Eric / Wardrip-fruin, Noah / Mastroni, Mitch / Lo-wood, Henry / De groat, Greta / Edwards, Glynn / Barrett, Marcia / Caldwell, Christy** (2016): „Implementing Controlled Vocabularies for Computer Game Platforms and Media Formats in SKOS“, in: *Journal of Library Metadata* 16: 1-22 <https://doi.org/10.1080/19386389.2016.1167494>.
- Kaltman, Eric / Osborn, Joseph / Wardrip-Fruin, Noah / Mateas, Michael** (2017): „Game and Interactive Software Scholarship Toolkit (GISST)“, in: *Proceedings of the 12th International Conference on the Foundations of Digital Games (FDG '17)* 1-4.
- Kaltman, Eric / Osborn, Joseph / Wardrip-Fruin, Noah** (2021): „From the Presupposition of Doom to the Manifestation of Code: Using Emulated Citation in the Study of Games and Cultural Software“, in: *Digital Humanities Quarterly* 15(1). <http://digitalhumanities.org/dhq/vol/15/1/000501/000501.html> [letzter Zugriff 26. November 2021].
- Kommission Zukunft der Informationsinfrastruktur** (2011): „AG Nichttextuelle Materialien“, in: *Gesamtkonzept für die Informationsinfrastruktur in Deutschland* B39-B53 https://www.hof.uni-halle.de/web/dateien/KII_Gesamtkonzept_2011.pdf [letzter Zugriff 08. Juni 2021].
- Koubek, Jochen** (2020): „Gamebegriffe“, in: Zimmermann, Olaf / Falk, Felix (eds.) *Handbuch Gameskultur*. Berlin: Deutscher Kulturrat e.V. 34-38.
- Lange, Andreas** (2020) „Bewahrung“, in: Zimmermann, Olaf / Falk, Felix (eds.) *Handbuch Gameskultur*. Berlin: Deutscher Kulturrat e.V. 125-130.
- Lee, Jin Ha / Clarke, Rachel Ivy / Perti, Andrew** (2015a): „Empirical Evaluation of Metadata for Video Games and Interactive Media“, in: *Journal of the Association for Information Science and Technology* 66(12): 2609-2625 <https://doi.org/10.1002/asi.23357>.
- Lee, Jin Ha / Clarke, Rachel Ivy / Sacchi, Simone / Jett, Jacob** (2015b): „Relationships among Video Games: Existing Standards and New Definitions“, in: *Proceedings of the American Society for Information Science and Technology* 51(1): 1-11 <https://doi.org/10.1002/meet.2014.14505101035>.
- Lew, Michael S. / Sebe, Nicu / Djeraba, Chabane / Jain, Ramesh** (2006): „Content-Based Multimedia Information Retrieval: State of the Art and Challenges“, in: *ACM Transactions on Multimedia Computing, Communications and Applications* 2(1): 1-19 <https://doi.org/10.1145/1126004.1126005>.
- Loebel, Jens-Martin / Hahn, Carolin** (2020): „#5: Ein Besuch bei Jochen Koubek, Computerspielwissenschaft, Universität Bayreuth“, in: *Digitale Wissenschaft* [Audio-Podcast] <https://open.spotify.com/episode/1UWQp8o2XXoM6xHuKKMr4C> [letzter Zugriff 04. Juni 2021].
- McDonough, Jerome / Olendorf, Robert / Kirschenbaum, Matthew / Kraus, Kari / Reside, Doug / Donahue, Rachel / Phelps, Andrew / Egert, Christopher / Lowood, Henry / Rojo, Susan** (2010): „Software Preservation and the Law“

52-57, "Preservation Strategies" 58-88 und "Packaging Virtual Worlds" 98-104, in: *Preserving Virtual Worlds Final Report* <http://hdl.handle.net/2142/17097> [letzter Zugriff 08. Juni 2021].

Newman, James (2019): "Saving (and Re-Saving) Videogames: Rethinking Emulation for Preservation, Exhibition and Interpretation", in: *The International Journal of Creative Media Research* 1: 1-18 <https://doi.org/10.33008/IJCMR.2019.08>.

NFDI4Culture (o.J.): "Task Areas" [online] <https://nfdi4culture.de/what-we-do/task-areas.html> [letzter Zugriff 26. November 2021].

Rocchio, J. J. / Salton, Gerard M. (1965): "Information search optimization and interactive retrieval techniques", in: *Proceedings of the November 30--December 1, 1965, fall joint computer conference, part I (AFIPS '65 Fall, part I)* 293-305 <https://doi.org/10.1145/1463891.1463926>.

Roeder, Torsten / Rettinghaus, Klaus (2020): „Game On! Digitale Archäologie und Edition zu(m) Spielen“, in: *DHd2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts* 138-141 <http://doi.org/10.5281/zenodo.3666690>.

Ryan, James Owen / Kaltman, Eric / Fisher, Andrew Max / Hong, Timothy / Owen-Milner, Taylor / Mateas, Michael / Wardrip-Fruin, Noah (2015a): "Large-Scale Interactive Visualizations of Nearly 12,000 Digital Games", in: *Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015)* 1-2.

Ryan, James Owen / Kaltman, Eric / Hong, Timothy / Mateas, Michael / Wardrip-Fruin, Noah (2015b): "People Tend to Like Related Games", in: *Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015)* 1-5.

Ryan, James Owen / Kaltman, Eric / Mateas, Michael / Wardrip-Fruin, Noah (2015c): "Tools for Videogame Discovery Built Using Latent Semantic Analysis", in: *Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015)* 1-2.

Sachs-Hombach, Klaus (2014): „Bilder als wahrnehmungsnahe Zeichen“, in: Sachs-Hombach (ed.) *Das Bild als kommunikatives Medium*. Elemente einer allgemeinen Bildwissenschaft. 3. Aufl., Köln: Halem 73-98.

Schöch, Christof (ed.) (2020): "DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts" <http://doi.org/10.5281/zenodo.3666690> [letzter Zugriff 14.07.2021].

Seadle, Michael (2008): "The digital library in 100 years: damage control", in: *Library Hi Tech* 26(1): 5-10 <https://doi.org/10.1108/07378830810857744>.

Shneiderman, Ben (1996): "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations", in: *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)* 336-343.

Szuban, Peter (2018): "Reconstituting Vocabularies: User Generated Databases, Social Tagging, and Folksonomies in Giantbomb's Videogame Wiki Database", in: *iJournal* 4(1): 41-49.

Zhang, Xiaoxuan / Holtz, Misha / Zhan, Zeping / Smith, Adam M. (2018): "Crawling, Indexing, and Retrieving Moments in Videogames", in: *Proceedings of the 13th International Conference on the Foundations of Digital Games (FDG 2018)* 1-10 <https://doi.org/10.1145/3235765.3235786>.

Best of Both Worlds Zur Kombination algorithmischer und manueller Verfahren bei der Erschließung großer Handschriftenkorpora

Nantke, Julia

julia.nantke@uni-hamburg.de
Universität Hamburg, Germany

Bläß, Sandra

sandra.blaess@uni-hamburg.de
Universität Hamburg, Germany

Flueh, Marie

marie.flueh@uni-hamburg.de
Universität Hamburg, Germany

Maus, David

david.maus@sub.uni-hamburg.de
Staats- und Universitätsbibliothek Hamburg, Germany

Einleitung

Für historische Zeiträume vor der Etablierung elektronischer Kommunikationsformen bilden Briefe das zentrale Medium des Austauschs über räumliche Distanzen hinweg. Aus heutiger Sicht sind Briefe deshalb wichtige historische Quellen, die persönliche Beziehungen ebenso dokumentieren wie gesellschaftlich relevante Orte, Ereignisse, Themen etc. Dies gilt umso mehr für ein solch umfangreiches Korrespondenznetz wie jenes des Ehepaars Ida und Richard Dehm. Die Dehms bildeten um 1900 das Zentrum eines europaweiten Netzwerks von Künstler:innen und Kulturschaffenden. An der in ca. 35.000 Briefen überlieferten Korrespondenz waren viele der zentralen Akteur:innen des damaligen Kunst- und Kulturbetriebs beteiligt.

Begreift man Erinnerung mit Jan Assmann als kreativen Gestaltungsprozess (vgl. Assmann 1999: 16), so lässt sich zunächst grundsätzlich feststellen, dass digitale Verfahren der Erschließung und Präsentation eine andere „Formung“ (Assmann 1999: 32) kultureller Vergangenheit ermöglichen als analoge, maßgeblich gedruckte Formate. In Bezug auf das Briefnetzwerk der Dehms ermöglicht die Digitalisierung durch neue Verfahren der computergestützten Erschließung und Präsentation eine Konzeptualisierung und Darstellung des Briefnetzwerks als kulturhistorischer Zusammenhang. Diese Forschungsperspektive rekurriert auf die Fähigkeit insbesondere schriftlicher Alltagsdokumente, Dialoge, Gedanken und Diskurse als „Zeitinseln“ (Assmann 1988: 12) zu transportieren und für die Nachwelt zu konservieren. Im Zuge einer kuratierenden Erschließung lassen sich diese Zeitinseln entsprechend für die wissenschaftliche und kulturhistorische Rezeption in der Gegenwart repräsentieren. Als „immutable mobiles“ (Latour 1990: 26) zeugen die Briefe zudem nicht nur von

zeithistorischen Ereignissen und gesellschaftlichen Entwicklungen des europäischen kulturellen Lebens um 1900, sondern in ihnen materialisieren sich auch die mit dem Medium Brief in dieser Zeit verknüpften kulturellen Praktiken und Kommunikationsformen. Nicht zuletzt zeigt außerdem der Fall des zu Lebzeiten weltberühmten und nach seinem Tod 1920 rasant dekanonisierten Dichters Richard Dehmel eindrücklich, wie stark literarische Moden und personell-institutionelle Konstellationen von kulturbetrieblichen, gesellschaftlichen und politischen Dynamiken abhängen und sich im Laufe der Zeit verändern.

Ziel des Projekts *Dehmel digital* ist es, die im Dehmel-Archiv der Staats- und Universitätsbibliothek Hamburg (SUB) archivierten Briefe in eine digitale Repräsentation zu überführen, welche die in den Briefen gespeicherten persönlichen, kulturellen und gesellschaftlichen Dynamiken erfahrbar und erforschbar macht. Kulturhistorische Zusammenhänge werden anhand ihres konkreten textuellen Niederschlags in den brieflichen Quellen erschlossen und repräsentiert (vgl. dazu auch Baßler 2005: 176f.; Baillet 2011). Indem die Briefe im Projekt als digitale Quellen rekontextualisiert und publiziert werden, erhalten sie gleichzeitig einen neuen „Ort der kanonisierten Erinnerung“, von dem aus sie vergegenwärtigt und erinnert werden und so zum kulturellen Gedächtnis beitragen können (vgl. Assmann 1999: 31).

Im Sinne des kulturellen Gedächtnisses ist das Briefkorpus vor allem als Ganzes interessant und relevant: Denn um nicht nur Richard Dehmels subjektive Wahrnehmung zum Kulturdiskurs der Jahrhundertwende zu erfassen, ist es unerlässlich, sich nicht auf dessen eigene Zeitzeugnisse zu beschränken, sondern auf das gesamte Gruppengedächtnis seines Netzwerks zurückzugreifen. Ein konzeptueller Wandel digitaler Editionsformate zugunsten der Dokumentation personeller und institutioneller Zusammenhänge deutet sich bereits an, steckt aber editionspraktisch noch in den Anfängen (Nutt-Kofoth 2020; Nantke 2019). Die Plattform *Briefe und Texte aus dem intellektuellen Berlin um 1800*, die *digitale Quelledition Der Sturm* sowie die Edition *Jean Paul – Sämtliche Briefe digital* sind teilweise noch in der Entwicklung befindliche Editionsprojekte, deren Materialauswahl und -präsentation anstelle von Einzelautor:innen und deren Werken auf kommunikative und institutionelle Netzwerke ausgerichtet sind. Sie bilden in ihrer Anlage Vorbilder für das Projekt *Dehmel digital*.

Zentral für die Etablierung solcher Editionsformate ist nicht zuletzt, dass den neuen digitalen Möglichkeiten der umfanglicheren *Repräsentation* personeller, institutioneller und medialer Netzwerke, die in den genannten Beispielen erprobt werden, auch entsprechende Verfahren der computergestützten *Erschließung* zur Seite gestellt werden. Dabei ist es entscheidend, Möglichkeiten und Grenzen einer computationellen Erschließung großer Handschriftenkorpora zu reflektieren, die dem erhöhten Materialumfang, den es zu erschließen gilt, gerecht werden. Es gilt zu diskutieren, in welchem Verhältnis Erschließungsumfang und textkritische Prüfung stehen sollen und wie sich die aktuellen digitalen Möglichkeiten nutzen lassen, um computergestützt auch große Datenmengen zu bewältigen, Zusammenhänge darzustellen und auf diese Weise neue Orte der Erinnerung zu etablieren. Welche computationellen Verfahren lassen sich an welchen Schnittstellen miteinander kombinieren und mit welchen Limitationen ist dabei zu rechnen? Wie kann also in Anbetracht begrenzter personeller und zeitlicher Ressourcen eine gute Mitte zwischen quantitativ-statistischen Verfahren und den qualitativ-philologischen Anforderungen einer digitalen Repräsentation gefunden werden? Für das Projekt *Dehmel digital* gilt, dass die Erfassung und adäquate Darstellung dieser großen Menge an Quellen nur im Rahmen einer Kombination algorithmischer quantitativer Verfahren und manueller Praktiken umsetzbar ist. Wir verfolgen deshalb

den Ansatz, die Qualitäten und arbeitspraktischen Vorteile beider Welten gewinnbringend zu verbinden.

Der Workflow

Der Prozess der Überführung der handschriftlichen Originale in maschinenlesbare Repräsentationen untergliedert sich in eine Reihe aufeinander aufbauender Transformationen, die miteinander zusammenhängende Abstraktionsschichten vom Ursprungsmaterial produzieren (vgl. Abb. 1).

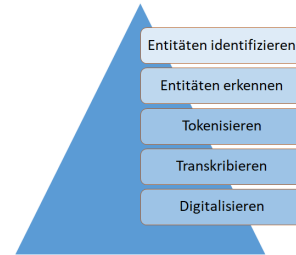


Abb. 1: Abstraktionsschichten des Workflows

Dabei besteht der Anspruch, ein Verfahren zu etablieren, welches zum einen die Erschließung eines möglichst großen Teils der Briefe ermöglicht, aber zum anderen einer editionsphilologisch validen Repräsentation der Dokumente verpflichtet bleibt. Entsprechend dieser doppelten Zielsetzung greifen in dem im Rahmen des Projekts entwickelten Workflow manuelle Arbeitsschritte und algorithmisch getriebene Prozesse ineinander. Dabei kombinieren wir mehrere bereits innerhalb der Digital Humanities etablierte Verfahren, die wir für den Einsatz in einer Edition modifizieren (vgl. Abb. 2).

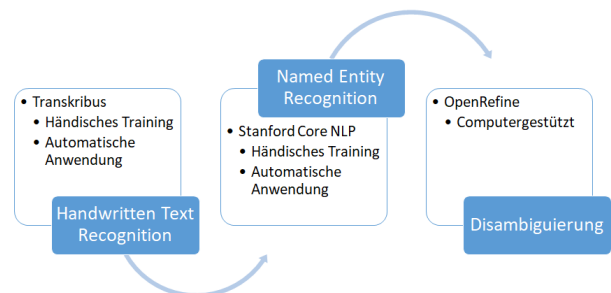


Abb. 2: Die einzelnen Arbeitsschritte

Nachdem die Originalbriefe in der SUB in hochauflösende Bilddigitalisate übersetzt und mit grundlegenden Metadaten angereichert wurden, transkribieren wir zunächst einige Briefe einer:ines Schreibenden manuell in *Transkribus* (vgl. <https://readcoop.eu/de/transkribus/>), bis genug Daten als Basis für das Training eines HTR-Modells erzeugt wurden. Im Rahmen dieses manuellen Schritts werden zudem in strukturierter Form weitere Metadaten zu den Briefen erfasst. Nach erfolgreichem Training eines HTR-Modells in *Transkribus* erfolgt die weitere Transkription im iterativen Wechsel zwischen automatisierter Transkription

und manueller Nachkorrektur, wobei die korrigierten Briefe wiederum als Trainingsdaten in den HTR-Workflow eingespeist werden. Die Transkriptionen werden als PAGE XML aus *Transkribus* exportiert.

Eine XML-Verarbeitungskette wandelt diese Transkriptionen in ein TEI-basiertes Format um und wendet von uns trainierte Modelle des *Stanford Named Entity Recognizers* (vgl. Manning et al. 2014) an, um Personen, Orte, Institutionen und Werke zu taggen. Da sich die klassischen Entitäten in Privatbriefen aus der Zeit um 1900 anders darstellen als in Gebrauchstexten, anhand derer die meisten Classifier trainiert wurden, nutzen wir Teile unserer Transkriptionen sowie die Daten weiterer digitaler Briefeditionen für das Training eigener Classifier. Ergänzend zum überwachten maschinellen Lernen werden Listen implementiert, um sicherzustellen, dass bestimmte, regelmäßig wiederkehrende Entitäten zuverlässig gefunden werden. Die sukzessive Erweiterung des Trainingsmaterials führt zunächst zu besseren Erkennungsraten innerhalb unseres Korpus, ist perspektivisch aber auch als generische Möglichkeit der automatisierten Briefannotation gedacht, die im Sinne einer digitalen, kollaborativen, wissenschaftlichen Korrespondenzanalyse auch für andere Projekte zugänglich gemacht werden soll. Erkannte Entitäten werden *inline* unter Beibehaltung der Bezüge zu Layout und Struktur des Dokuments eingebracht. Hierfür wird eine vereinfachte Implementierung der *Separated Markup API for XML* (vgl. Verwer 2020) verwendet, die klassifizierte Tokens der Named Entity Recognition und textstrukturelles Markup integriert. Die so erkannten Entitäten werden als Basis für die teilautomatisierte Generierung von Personen-, Orts-, Institutionen- und Werkregistern genutzt. Erst dieser Schritt einer basalen inhaltlichen Erschließung ermöglicht letztlich einen grundlegenden Überblick über sowie gezielte Einblicke in das umfangreiche Korpus anhand zentraler Entitäten und damit eine Nutzung als editorisch aufbereitete historische Quelle.

Um die wiederum anhand maschineller Lernverfahren teilautomatisiert ermittelten Entitäten in stabile Registereinträge zu überführen, wird ein maschinengestützter Abgleich (data reconciliation) mit Normdatensystemen und lokalen Wissensbasen durchgeführt. Die mit *OpenRefine* (vgl. <https://openrefine.org/>) computergestützt disambiguierten Entitäten-Typen bilden wiederum die Basis für manuell zu erstellende Makrokommentare zu den zentralen Akteur:innen, Institutionen und Werken des Korpus sowie für Netzwerkvisualisierungen, die wiederum mit den Briefen verlinkt sind und einen vom grafisch visualisierten Netzwerk ausgehenden Einstieg in die textnahe Lektüre der Briefe ermöglichen. Darüber hinaus werden die disambiguierten Entitäten mit Normdaten zu Personen, Orten, Werken und Organisationen verknüpft und die Briefe des Dehmel-Korpus via API in den Webservice *correspSearch* (<https://correspsearch.net/de/start.html>) eingebunden, mit dessen Hilfe Verzeichnisse verschiedener Briefeditionen nach Absender, Empfänger, Schreibort und -datum durchsucht werden können. Auf diese Weise wird das um die Dehmels bestehende Korrespondenznetz in bereits etablierte Netzwerkzusammenhänge integriert.

Ergebnis dieser – im Vortrag anhand von konkreten Beispielen genauer darzustellenden – Reihe von Datentransformationen sind also transkribierte, annotierte und mit Metadaten angereicherte XML-Dateien sowie stabile, disambiguierte und mit Normdaten verknüpfte Entitätenregister, die gemeinsam den Input für das Webportal von *Dehmel digital* bilden.

Die Sicherung und Pflege der Daten übernimmt die SUB Hamburg, sodass die nachhaltige Verfügbarkeit und Nutzbarkeit unserer Ergebnisse gewährleistet sind.

Repräsentation: Nutzungsszenarien und Gestaltungsfragen im Hinblick auf ein ‚digitales Gedächtnis‘

Die eingangs beschriebene Relevanz des Korpus als Teil eines digital repräsentierten Gedächtnisses europäischer Kulturgeschichte ist mit der Frage einer adäquaten Repräsentation verknüpft. Hierbei sind Überlegungen im Hinblick auf die Kontextualisierung der Inhalte relevant, die bereits durch die Anlage des Erschließungsworkflows vorstrukturiert werden: Im Projekt *Dehmel digital* steht anstelle der Äußerungen von Einzelpersonen die Korrespondenz als Netzwerkzusammenhang im Fokus; die im Dehmel-Archiv konservierten einzelnen Zeitinseln sollen verstärkt in einen Kontext zueinander und zu den Inhalten anderer digitaler Editionen gebracht werden können. Die materielle Erschließung, semantische Annotation und inhaltliche Kommentierung erfolgen deshalb aus der Perspektive einer möglichst breiten, dezentralen Dokumentation des Korrespondenznetzes als historisches Zeugnis eines kollektiven Gedächtnisses.

Dementsprechend ist es das Ziel des im Rahmen von *Dehmel digital* entwickelten Webportals, möglichst vielfältige Nutzungsszenarien von der skalierbaren Lektüre (vgl. Weitin 2017) bis hin zu maschinell gestützten Auswertungen anhand eigener Forschungsfragen zu ermöglichen. Hierbei sind unterschiedliche Nutzer:innengruppen von kulturinteressierten Museumsbesucher:innen des Dehmelhauses (vgl. <https://www.dehmelhaus.de/aktuell.html>) bis hin zur (digital arbeitenden) geisteswissenschaftlichen Community mitgedacht. Latour zufolge erreicht Wissen viele Menschen an verschiedenen Orten am wirkungsvollsten, wenn es in mobiler, beständiger, präsentierbarer, lesbarer und kombinierbarer Form vorliegt (vgl. Latour 1990: 23–26). Eine digitale Präsentation auf einem Webportal erweist sich insbesondere in Bezug auf die Darstellbarkeit von Netzwerken und die Mobilität der Präsentation im Allgemeinen als besonders effektiv, da sie dauerhaft von überall aus kostenlos mit eigenen Geräten abgerufen werden kann, nicht an ephemere Materialien gebunden ist und verschiedene Aufbereitungen je nach Zielgruppe und Nutzungsinteresse miteinander kombiniert werden können (vgl. dazu grundsätzlich bezogen auf digitale Repräsentationen auch Sahle 2016: 30): Auf dem Portal selbst unterstützen facetiierte Suchen, Netzwerk- sowie Kartenvisualisierungen und Makrokommentare die strukturierte Rezeption sowie eigenständige Recherchen im Korpus. Sie bilden moderierte Einstiege in das umfangreiche Material, die sich bei Bedarf an individuelle Interessen flexibel anpassen lassen (vgl. Spoerhase 2015: 640–643). Über das Portal hinaus stehen die produzierten Daten der Nachnutzung in anderen Forschungsszenarien offen. In diesem Sinne werden nicht nur die Faksimiles und die erzeugten Transkriptionen in verschiedenen Formaten (TEI, Plaintext, PDF), sondern ebenfalls die für den beschriebenen Workflow entwickelten HTR- und NER-Modelle sowie unsere Routinen zum Download zur Verfügung gestellt.

Die Digitalisierung des kulturellen Gedächtnisses ermöglicht in der Kombination manueller und algorithmischer Arbeitsprozesse zum einen überhaupt erst die quantitativ-qualitative Erschließung des Dehmelschen Korrespondenznetzes. Zum anderen ist die digitale Repräsentation ebenfalls die Bedingung für die vielfältigen und skalierbaren Rezeptionsszenarien desselben als kulturelles Artefakt – und zwar nicht als Zeugnis einer isolierten Vergangenheit, sondern einer, die durch Diskussion und individuelle Aneignung

nung der Quellen mit der Gegenwart verbunden werden kann (vgl. Assmann 1988: 13).

Der Beitrag stellt den hier beschriebenen Workflow anhand von konkreten Beispielen dar und gibt anhand der ersten öffentlichen Beta-Version Einblicke in die geplante und bereits prototypisch implementierte Umsetzung auf dem Portal. Dabei wird es ebenfalls darum gehen, anhand der Beispiele den Umgang mit dem Spannungsfeld zwischen quantitativer Erschließung und klassischer philologischer Arbeit zu diskutieren.

Bibliographie

Assmann, Jan (1988): "Kollektives Gedächtnis und kulturelle Identität". In: Ders.; Hölscher, Tonio (Hrsg.): *Kultur und Gedächtnis*. Frankfurt am Main, S. 9–19.

Assmann, Jan (1999): "Kollektives und kulturelles Gedächtnis. Zur Phänomenologie und Funktion von Gegen-Erinnerung". In: Borsdorf, Ulrich; Grütter, Heinrich Theodor (Hrsg.): *Orte der Erinnerung. Denkmal, Gedenkstätte, Museum*. Frankfurt am Main/New York, S.13–32.

Baillet, Anne (2011): "Einleitung". In: Dies. (Hrsg.): *Netzwerke des Wissens. Das intellektuelle Berlin um 1800*. Berlin 2011, S. 11–23.

Baßler, Moritz (2005): *Die kulturpoetische Funktion und das Archiv. Eine literaturwissenschaftliche Text-Kontext-Theorie*. Tübingen.

Briefe und Texte aus dem intellektuellen Berlin um 1800 (o.D.): Hrsg. v. Anne Baillet, Humboldt Universität Berlin. URL: <http://www.berliner-intellektuelle.eu/>.

Der Sturm. Digitale Quellenedition zur Geschichte der internationalen Avantgarde (2018): Hrsgg. von Marjam Trautmann und Torsten Schrade, Mainz, Akademie der Wissenschaften und der Literatur. URL: <https://sturm-edition.de>.

Jean Paul. Sämtliche Briefe digital (2018): Hrsgg. im Auftrag der Berlin-Brandenburgischen Akademie der Wissenschaften von Markus Bernauer, Norbert Miller und Frederike Neuber. URL: <https://www.jeanpaul-edition.de/start.html>.

Latour, Bruno (1990): "Drawing things together". In: Michael E. Lynch und Steve Woolgar (Hrsg.): *Representations in Scientific Practice*. Cambridge, S. 19–68.

Manning, Christopher / Surdeanu, Mihau / Bauer, John / Finkel, Jenny / Bethard, Steven J. / McClosky, David (2014): "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, S.55–60.

Nantke, Julia (2019): "Konzepte digitaler (Re-)Präsentationen von Literatur zwischen Pluralisierung und Standardisierung". In: Martin Huber, Sybille Krämer und Claus Pias (Hrsg.): *Forschungsinfrastrukturen in den digitalen Geisteswissenschaften. Wie verändern digitale Infrastrukturen die Praxis der Geisteswissenschaften?* Frankfurt a. M., S. 58–76. urn:nbn:de:hebis:30:3-526104

Nutt-Kofoth, Rüdiger (2020): "Der Brief als Forschungsfeld - Editionswissenschaft". In: Marie Isabel Matthews-Schlinzig, Jörg Schuster, Gesa Steinbrink und Jochen Strobel (Hrsg.): *Handbuch Brief. Von der Frühen Neuzeit bis zur Gegenwart*. Berlin/Boston, S. 81–96.

Sahle, Patrick (2016): "What is a Scholarly Digital Edition?" In: Matthew James Driscoll und Elena Pierazzo (hrsg.): *Digital Scholarly Editing: Theory and Practice*. Cambridge, S. 19–39. DOI: <http://dx.doi.org/10.11647/OBP.0095.02>.

Spoerhase, Carlos (2015): "Gegen Denken? Über die Praxis der Philologie". In: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 89, S. 637–646.

Verwer, Nico (2020): "Plain text processing in structured documents". In: *Proceedings of Declarative Amsterdam 2020*. DOI: 10.1075/da.2020.verwer.plain-text-processing.

Weitin, Thomas (2017): "Scalable Reading". In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47, S. 1–6.

o.V.: *corresp Search*. URL: <https://correspsearch.net/de/star-t.html> [13.Juli 2021].

o.V.: Dehmelhaus Stiftung Hamburg. URL: <https://www.dehmelhaus.de/aktuell.html> [29. November 2021].

o.V.: *OpenRefine*. URL: <https://openrefine.org/> [13. Juli 2021].

o.V.: *Transkribus. KI-gestützte Handschriftenerkennung*. URL: <https://readcoop.eu/de/transkribus/> [13. Juli 2021].

Data Cleaning als digitale Quellenkritik VD17 und das Genre der katholischen Dissertation im Alten Reich

Heßbrüggen-Walter, Stefan

early.modern.thought.online@gmail.com
HSE University, Russian Federation

Die forschende Nachnutzung von Metadaten analoger und digitaler Artefakte ist ein wesentlicher Bestandteil der Arbeit am kulturellen Gedächtnis. Die digitalen Geisteswissenschaften, insbesondere die digitale Literaturwissenschaft, haben hierzu schon beachtliche Beiträge geleistet, insbesondere in der Betrachtung der Geschichte literarischer Gattungen (Fischer 2018, Gittel 2021). Auch in der Wissenschafts- und Geistesgeschichte ist die Relevanz von Metadaten mittlerweile erkannt worden (Sangiacomo 2020, Scholz 2021). Historische Arbeit an und mit Metadaten, so die erste hier zu verteidigende Teilthese, erfordert aber, wie die Bearbeitung jeder anderer digitalen Quelle auch, „Quellenkritik“ (entsprechende Überlegungen in den Geschichtswissenschaften konzentrieren sich eher auf Archivalien und digital entstandenes Material, siehe Friederich 2018). Denn anders als mancher Kritiker meint, bleibt es Utopie, „in der elektronisch schaltbaren Präsenz des Wissens zugleich seine absolute Verfügbarkeit“ herzustellen (Jochum 1998, 29f). Metadaten des kulturellen Erbes sind vielmehr menschengemachte und deswegen außerordentlich fragile Gebilde.

Dies gilt erst recht, wenn mehrere Institutionen über einen längeren Zeitraum einen komplexen Bestand an Schrifttum zu verzeichnen haben. Meine Fallstudie betrifft einen zahlenmäßig unbedeutenden Ausschnitt der deutschen Nationalbibliografie für Drucke des 17. Jahrhunderts, VD17 (Anonym 2020), nämlich Metadaten zu philosophischen Dissertationen, die im Erfassungszeitraum von VD17 (1601-1700) an katholischen Institutionen angefertigt wurden.¹

Ich hoffe zu zeigen, und dies ist meine zweite Teilthese, dass die angesprochene Quellenkritik der Metadaten solcher Dissertationen bereits auf der Ebene der „Arbeitsvorbereitung“ stattzufinden hat, also in jenem Projektschritt, den wir in den digitalen Geisteswissenschaften englisch als *data cleaning* bezeichnen. Ein deut-

scher Ausdruck hat sich hierfür bislang nicht eingebürgert (CrankyPhilosoph 2021). Kurz könnte man beide Teilthesen auch so zusammenfassen: *Data cleaning* ermöglicht nicht nur Forschung, *data cleaning* erfordert auch Forschung oder kann selbst zu Forschungseinsichten führen.

Es wird gelegentlich die Auffassung vertreten, dass die Produzenten von Metadaten, im vorliegenden Fall also Bibliotheken, selbst dafür Sorge zu tragen hätten, dass ihre Aufnahmen den Anforderungen genügen, die von Forschenden v. a. in den Digital Humanities an solche Datensätze gestellt werden (Király / Brase 2021, 358-359). Inwiefern ein solcher Anspruch realistisch ist, wäre wohl zuerst aus bibliothekswissenschaftlicher Sicht zu klären. Wenn man ihn erhebt, muss man sich jedoch im Lichte des folgenden darüber im Klaren sein, dass dann Erschließungsleistungen in der Katalogisierung bspw. Alter Drucke selbst schon Forschungsleistungen darstellen, diese nicht bloß ermöglichen.

Die hauptsächliche Forschungsfrage, um die es im folgenden geht, lautet: wie viele katholische philosophische Dissertationen sind in VD17 verzeichnet? Diese auf den ersten Blick vielleicht nicht besonders dringlich erscheinende Frage gilt es zu beantworten, wenn man beispielsweise die Relevanz des Genres Dissertation im katholischen Raum in Beziehung setzen will zur Rolle der Gattung in der protestantischen Universität. Drei Anmerkungen sind zum Verständnis ihrer Voraussetzungen hilfreich. Erstens sind die „örtlichen Kriterien“ für VD17, also das von der Bibliografie abzudeckende Territorium notorisch vage (Stäcker 2004, 214f). Zweitens haben nicht alle deutschen Bibliotheken an diesem Projekt mitgewirkt, es steht also anzunehmen, dass nicht alle zwischen 1601 und 1700 in diesem vagen Raum erschienenen katholischen Dissertationen in der Bibliografie verzeichnet sind. Drittens zählt VD17 kleinere Abweichungen im Druck als eigenständige Editionen, so dass das Zählen von „Werken“ („geistigen Gegenständen“, Heßbrüggen-Walter 2020) hier eigene Herausforderungen birgt, um die es zunächst noch nicht gehen soll. In operationalisierbarer Form würde unsere Forschungsfrage also lauten: wie viele VD17-Zitiernummern beziehen sich auf eine an einer katholischen Bildungseinrichtung angefertigte Dissertation? Anhand eines kurzen Beispiels werde ich zudem erläutern, welche Herausforderungen bei der schlüssigen Erfassung von mit Dissertationen verbundenen Körperschaften zu bewältigen sind.

Man kann sich berechtigterweise auf den Standpunkt stellen, dass die Beantwortung der Frage nach der Zahl katholischer Dissertationen in VD17 mit *data cleaning* alleine nicht zu leisten ist, selbst wenn die auf diesen Arbeitsschritt folgende Auswertung der aufbereiteten Daten in einer Zeile Programmcode bestehen würde. Das stimmt auch. Um jedoch die Daten in einer Form zur Verfügung zu stellen, die eine solche Auswertung möglich macht, sind weitere unter das *data cleaning* zu subsumierende Arbeitsschritte notwendig, die ihrerseits, wie nun zu zeigen ist, Forschungsleistungen darstellen bzw. voraussetzen.

Zunächst kurz zum Begriff der Dissertation und der Relevanz dieses Genres nicht nur für die philosophiegeschichtliche Forschung im engeren Sinne, sondern auch für die Erforschung frühneuzeitlicher Bildungsgeschichte im weitesten Sinne. Wesentliches Merkmal dieser Texte ist ihre Bindung an eine Einrichtung höherer Bildung – neben Universitäten auch akademische Gymnasien ohne Promotionsrecht oder an Klöster gebundene Ordensstudien. Eine genauere Erfassung dieser Dissertationen und der mit ihnen verbundenen Personen und Institutionen ist nicht nur eine Vorbedingung ihrer in weiten Teilen noch ausstehenden philosophiegeschichtlichen Erschließung (siehe jedoch für Dillingen Leinsle 2006), sondern schon allein auf der Ebene der Prosopographie von Belang. Denn nicht nur haben Geistliche und Mönche an katholischen Universitäten studiert, auch die Ordensstudien selbst

haben Dissertationen hervorgebracht. Damit sind Dissertationen wertvolle Datenquellen für die Identifizierung konkreter Individuen in monastischen Gemeinschaften und für deren Praxis philosophischen Unterrichts.

Philosophische Dissertationen sind als Genre in den AAD *Gattungsgenres* (Anonym o. J.) enthalten, die für die Formalerschließung in VD17 zugrundegelegt wurden (Anonym 2020). Daneben enthalten die *Gattungsgenres* auch die Kategorie „Ordensliteratur“ mit 22 Unterkategorien. Da die akademische Lehre im katholischen Deutschland des 17. Jahrhunderts Ordensgemeinschaften und den mit diesen verbundenen Institutionen oblag, kann man zunächst davon ausgehen, dass katholische philosophische Dissertationen in der Schnittmenge beider Kategorien zu finden sind.

Rahmenbedingungen

VD17-Daten wurden über die zur Verfügung gestellte SRU-Schnittstelle im MODS-Format abgefragt und in nach Suchbegriff unterschiedenen XML-Dateien gespeichert. Für unsere Analyse sind zwei im Rahmen von MODS als Namen bzw. Namens-Ids erfasste Felder in den Katalogisaten von Belang: ‚name-corporate‘ mit dem Attribut ‚oth‘ erfasst die für die Dissertation verantwortliche Bildungseinrichtung. Und das sogenannte *statement of responsibility* erfasst weitere für Dissertationen einschlägige Informationen. Die beschriebenen Tags wurden mit Hilfe eines Jupyter-Notebooks ausgelesen und weiterverarbeitet.²

Data Cleaning I: protestantische Dissertationen über katholische Orden

Die Hoffnung, dass die Schnittmenge der Kategorien ‚philosophische Ordensliteratur‘ und ‚philosophische Dissertation‘ alleine Arbeiten enthält, die an katholischen Institutionen entstanden sind, erfüllte sich nicht. Von den 1245 Titeln, die beide Kriterien erfüllen, sind vielmehr neun an protestantischen Institutionen, nämlich den Universitäten Wittenberg, Leipzig, Jena, Halle und Gießen sowie den Gymnasien in Ulm und Bayreuth entstanden. Es handelt sich bei diesen Arbeiten allerdings nicht um Beiträge katholischer Gastprofessoren, sondern vielmehr um Arbeiten, die aus protestantischer Sicht über katholische Orden verfasst worden sind. Sie behandeln zum Beispiel die reservatio mentalis der Jesuiten oder das Leben des Heiligen Norbert von Xanten, Erzbischof von Magdeburg und Stifter des Prämonstratenserordens. Damit verbleiben 1236 katholische Dissertationen. Die Kategorie ‚Ordensliteratur‘ ist also nicht eindeutig bestimmt: es kann sich hier sowohl um Literatur handeln, die aus einem Orden hervorgeht, als auch um Literatur, die über einen Orden oder dessen Mitglieder verfasst wurde.

Data Cleaning II: Augsburg

VD17 enthält jedoch auch 30 Datensätze zu Dissertationen katholischer Bildungseinrichtungen, bei denen die Klassifikation als Ordensliteratur fehlt. Diese Datensätze wurden identifiziert, indem alle mit einer als katholisch ausgewiesenen Institution in Zusammenhang stehenden philosophischen Dissertationen in einem zweiten Anlauf darauf hin durchsucht wurden, ob für sie auch

nicht als Ordensliteratur ausgezeichnete Titel in VD17 verzeichnet sind.

Dazu mussten in den im ersten Arbeitsschritt die in den schon erfassten 1235 Datensätzen verzeichneten Bildungsinstitutionen eindeutig identifiziert werden. Körperschaften werden in VD17, anders als ein Teil der Autoren, jedoch nicht mit GND-Identifikatoren versehen und haben auch keine vereinheitlichte Ansetzungsform. Hier mussten also Daten bereinigt und ergänzt werden. Dies warf besondere Schwierigkeiten auf, wenn in einer Stadt sowohl eine katholische wie eine protestantische Bildungseinrichtung vorhanden sind. Dies soll am Beispiel Augsburgs erklärt werden.

Für das „Jesuitenkolleg (Augsburg)“ (GND 4222329-5) enthält VD17 insgesamt 14 Ansetzungsformen für 37 Dissertationen. Acht dieser Ansetzungsformen sind eindeutig identifizierbar, denn sie enthalten das Patrozinium der Lehranstalt oder nehmen eindeutig Bezug auf ihren Träger, den Jesuitenorden: Lyceum S. Salvatoris Augsburg, Lyceum S. Salvatoris Augustae Vindel, S. Salvatoris Lyceum Augustanum, Lyceum S. Salvatoris Augustanum, S. Salvatoris Lyceum Augustanum [sic!], Gymnasium S. Salvatoris Augsburg, Gymnasium ad S. Salvatoris Augsburg, Gymnasium ad S. Salvatoris. Sie erfassen 56,8% der an dieser Institution verfassten und in VD17 überlieferten Dissertationen.

Neben dem dem Erlöser (*salvator*) gewidmeten Jesuitenkolleg existierte in Augsburg im 17. Jahrhundert jedoch auch ein protestantisches Gymnasium. Die Ansetzungsformen „Gymnasium Augsburg“ und „Lyceum Augustanum“ sind also nicht eindeutig. Das protestantische Gymnasium der Stadt befand sich zu seiner Gründung in den Räumlichkeiten eines vor der Reformation der Hl. Anna geweihten Klosters und leitete daraus seinen Namen ab (GND 2012843-5). Jedoch finden sich auch von Jesuiten veranstaltete Dissertationen, deren Ansetzung in VD17 ebenfalls auf ein der Hl. Anna geweihtes Kloster Bezug nehmen, und zwar sechs: Sankt Anna, Augsburg, Augustae Vindelicorum ad S. Annam, Kloster St. Anna, Augsburg, S. Anna, Augsburg, Gymnasium St. Anna, Augsburg, Collegium Philosophicum ad S. Annam, Augsburg. Der dem Kloster gewidmete Wikipedia-Artikel gibt näheren Aufschluss: das protestantische Gymnasium bei St. Anna bezog schon 1613 ein neues Gebäude, behielt aber den alten Namen bei. Zugleich zog das dem Erlöser gewidmete Jesuitenkolleg 1635 in die Räumlichkeiten des Klosters St. Anna (Wikipedia 2021) und nutzte also anscheinend dessen Bezeichnung auch für die eigene Institution. Diese streng genommen irreführenden Zuschreibungen machen 43,2% der jesuitischen Dissertationen Augsburgs aus. Die Dissertation mit der VD17-Zitiernummer 23:241965E ist nur als am – ja streng genommen nicht existenten ‚Gymnasium Augsburg‘ entstandene – ‚Dissertation:phil.‘ ausgezeichnet. Ohne die beschriebenen Klärungen hätte sie damit nicht als katholische Dissertation identifiziert werden können.

setzung dieses Papers. Es verbleiben 60 Katalogeinträge, für die zwar keine Institution eingetragen ist, diese aber aus den vorliegenden Metadaten maschinell oder durch Inspektion bspw. der Schlüsselseiten festgestellt werden können. Wesentliches Hilfsmittel war hier das bereits erwähnte MODS-Tag *statement of responsibility*, das die bei der Katalogisierung für wesentlich erachteten Elemente des Untertitels der Dissertation, darunter zumindest gelegentlich eben auch die verantwortliche Körperschaft, erfasst. Die Suche nach 26 Zeichenketten, die jeweils eine Körperschaft eindeutig identifizieren, erlaubte die Klärung der institutionellen Verantwortlichkeit von 46 der 60 ungeklärten Dissertationen. Für die weiteren 14 Titel war eine manuelle Inspektion der Schlüsselseiten erforderlich. Es erweist sich in diesem Zusammenhang als misslich, dass in VD17, anders als noch in VD16, bewusst auf die vollständige, aber eben „umständliche“ (Stäcker 2004, 214) Erfassung der Titelseiten verzichtet wurde.

Zusammenfassung und Ausblick

Die im Rahmen des *data cleaning* der Katalogisate katholischer Dissertationen erzielten quellenkritischen Forschungsergebnisse betreffen zunächst den im VD17 verwendeten Gattungsbegriff „Ordensliteratur“: es ist nicht deutlich, ob dieser nur die literarischen Hervorbringungen von Ordensgemeinschaften oder auch Schriften über Ordensgemeinschaften erfassen soll. Der inklusive Gebrauch der Kategorie in VD17 ist nirgends dokumentiert und für den unvoreingenommenen Betrachter auch nicht sofort offensichtlich. Die Ansetzung von Institutionen schließt nicht den Gebrauch von Normdaten ein und ist somit zunächst nicht nachvollziehbar und unzuverlässig. Zwar hatte dies zumindest in unserem Fall keinen Einfluss auf die Zählung katholischer Dissertationen, wirkt aber dennoch auf den unbefangenen Nutzer irreführend.

Eingangs war die Frage aufgeworfen worden, wieviele katholische Dissertationen in VD17 verzeichnet sind. In der Kombination der Gattungsbegriffe für Ordensliteratur und philosophische Dissertation wurden zunächst 1245 einschlägige Katalogisate identifiziert. Davon waren neun jedoch protestantischen Institutionen zuzuordnen. Diesen 1236 Dissertationen sind 30 Titel zuzuschlagen, die zwar an einer katholischen Institution entstanden sind, aber nicht als Ordensliteratur ausgezeichnet wurden. Ein Titel, der als Dissertation erfasst wurde, aber keine Hochschulschrift darstellt, ist abzuziehen. Damit kommen wir auf 1265 in VD17 verzeichnete katholische Dissertationen. Die Frage, wie viele katholische Dissertationen insgesamt in deutschen Bibliotheken und darüber hinaus vorhanden sind, würde weitere Recherchen in Verbundkatalogen und Altbestandsbibliotheken ohne VD17-Katalogisierung voraussetzen.

1950 Wörter

Data Cleaning III: Keine Institution

In 68 Katalogisaten von Dissertationen finden wir keine Angabe einer Institution, die für die Dissertation verantwortlich zeichnen würde, obwohl die Titel sowohl als Ordensliteratur wie auch als philosophische Dissertation klassifiziert worden sind. In einem Fall handelt es sich tatsächlich nicht um eine Dissertation im formalen Sinne einer Hochschulschrift (VD17 14:697769A). In sieben Fällen ist auf dem Titelblatt zwar der Dissertationscharakter ersichtlich, aber keine verantwortliche Institution angegeben. Sie könnte vermutlich unter Zuhilfenahme von Druckort und Autor erschlossen werden, doch der Status solcher erschlossener Metadaten wäre eigens zu bedenken und liegt außerhalb der Ziel-

Fußnoten

1. Diese Beschränkung erschien auch aus rechtlichen Gründen angezeigt, weil größere Ausschnitte von Datenbanken urheberrechtlichem Schutz unterliegen können. VD17 macht keine Angaben zu den rechtlichen Rahmenbedingungen einer etwaigen Nachnutzung der zur Verfügung gestellten Daten.
2. Daten und das Python-Notebook sind unter <https://gitlab.com/shessbru/catholic-dissertations-in-vd17/-/tree/main> einsehbar.

Bibliographie

Anonym (o. J.): AAD Gattungsgenres, URL: <http://uri.gbv.de/terminology/aadgenres/>

Anonym (2020): VD17 - Das Verzeichnis der im deutschen Sprachraum erschienenen Drucke. URL: <http://archive.md/ngh9x>

CrankyPhilosoph (2021): „Dumme Frage in den Saal: wie heißt ‚data cleaning‘ eigentlich auf deutsch? #DH“. Tweet. @FrueheNeuzeit (Blog), URL: <https://twitter.com/FrueheNeuzeit/status/1414988624874070023>.

Fischer, Frank / Jäschke, Robert (2018): „Liebe und Tod in der Deutschen Nationalbibliothek: Der DNB-Katalog als Forschungsobjekt der digitalen Literaturwissenschaft.“ in: *DHD2018: „Kritik der digitalen Vernunft“, Digital Humanities im deutschsprachigen Raum*, Feb. 2018, Köln, Deutschland, 261-266. URL: <https://hal.archives-ouvertes.fr/hal-01787558>

Friederich, Christine (2018): Tagungsbericht: "HT 2018: Quo vadis Quellenkritik? Digitale Perspektiven", 25. 09. 2018–28. 09. 2018 Münster, in: *H-Soz-Kult*, 23. 11. 2018, URL: <http://www.h-sozkult.de/conferencereport/id/tagungsberichte-7977>.

Gittel, Benjamin (2021): „An Institutional Perspective on Genres: Generic Subtitles in German Literature from 1500-2020“, in: *Journal of Cultural Analytics* 10.22148/001c.22086.

Heßbrüggen-Walter, Stefan (2013): "Tatsachen im semantischen Web: Nanopublikationen in den digitalen Geisteswissenschaften?", in: Haber, Peter, Pfanzelter, Eva (eds.): *Historyblogosphere*, München: Oldenbourg Wissenschaftsverlag, 149-160. 10.1524/9783486755732.149

Heßbrüggen-Walter, Stefan (2020): "Positivismus der geistigen Gegenstände: Carnap und die Digital Humanities", in: *DHD 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHD 2020)*, Paderborn: Zenodo. 10.5281/zenodo.4621804

Jochum, Uwe (1998): „Die Bibliothek als locus communis“, in: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte*, 14-30 10.1007/BF03375514.

Király, Péter / Brase, Jan (2021): "Qualitätsmanagement", in: Markus Putnings, Heike Neuroth, Janna Neumann (eds.): *Praxishandbuch Forschungsdatenmanagement*, Berlin, Boston: De Gruyter Saur, 2021, 357-380. 10.1515/9783110657807-020

Leinsle, Ulrich G. (2006): *Dilinganae disputationes: der Lehrinhalt der gedruckten Disputationen an der Philosophischen Fakultät der Universität Dillingen 1555 - 1648*. (Jesuitica 11). Regensburg: Schnell Steiner.

Sangiacomo, Andrea / Beers, Daan (2020): „Divide et Impera: Modeling the Relationship between Canonical and Noncanonical Authors in the Early Modern Natural Philosophy Network“, in: *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, 365-413 10.1086/710178.

Scholz, Luca (2021): „A Distant Reading of Legal Dissertations from German Universities in the Seventeenth Century“, in: *The Historical Journal*, 1-31, 10.1017/S0018246X2100011X.

Stäcker, Thomas (2004): „VD 17 – mehr als eine Zwischenbilanz“, in: *Zeitschrift für Bibliothekswesen und Bibliographie* 51: 213-21.

Wikipedia (2021): „Karmelitenkloster Augsburg“, in: *Wikipedia*, 25. April 2021. https://de.wikipedia.org/w/index.php?title=Karmelitenkloster_Augsburg&oldid=211306498.

Der CLARIAH-DE Tutorial Finder

Eine Suchumgebung für Lehr- und Schulungsmaterialien in den Digital Humanities

Werthmann, Antonina

werthmann@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache

Gradl, Tobias

tobias.gradl@uni-bamberg.de
Universität Bamberg

Einleitung

Zu digitalen Methoden, Forschungsdaten, Ressourcen und Diensten im Bereich der Digital Humanities sind zahlreiche Lehr- und Schulungsmaterialien vorhanden, die nicht nur Forschenden, sondern auch Lehrenden und Studierenden einen guten Einstieg in praktische Anwendungsfälle bieten. Als Beispiele finden sich Sammlungen und Plattformen wie TeLeMaCo¹, DARIAH-Campus² und linguisticsweb.org³, die Materialien zu unterschiedlichen Themen der Digital Humanities zusammenstellen. Zudem gibt es eine Vielzahl von einzeln entwickelten Anleitungen zu wichtigen Grundlagen aus dem Bereich der digitalen Editionen, des maschinellen Lernens und des Natural Language Processing (NLP), wie „Digitale Textedition mit TEI“⁴, „NLP Based Analysis of Literary Texts“⁵ oder „TopicsExplorer“⁶. Diese sind zumeist über verschiedene Plattformen oder die Webseiten zugehöriger Projekte zu finden. Die Formate der Materialien reichen von einfachen Dokumentationen über aufgezeichnete Vorträge und Präsentationsfolien bis hin zu didaktisch-methodisch ausgearbeiteten Lerneinheiten und Modulen zum Selbstlernen sowie direkt einsetzbaren Unterrichtsmaterialien. Sie variieren in ihrem Schwierigkeitsgrad und können sowohl für einen niederschweligen Einstieg ohne Vorkenntnisse als auch für eine Vertiefung unterschiedlichster Themen genutzt werden. Die Angebotslandschaft der Lehr- und Schulungsmaterialien ist nicht nur äußerst vielfältig und heterogen, sondern sie zeichnet sich auch dadurch aus, dass sie im Rahmen von unterschiedlichen nationalen und internationalen Projekten, Initiativen und Vorhaben entwickelt, zusammengestellt und der wissenschaftlichen Community zur Verfügung gestellt wird.

Hintergrund und Ziele

Um eine bessere Erreichbarkeit und Zugänglichkeit zu bestehenden sowie neuen Angeboten von Lehr- und Schulungsmaterialien im Bereich der Digital Humanities zu ermöglichen, sollten diese in einem zentralen Verzeichnis zur Verfügung gestellt werden. Im Rahmen des CLARIAH-DE Projekts⁷ wurde – zu

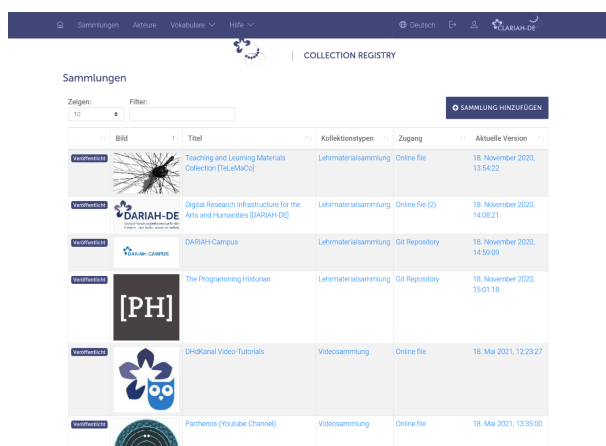


Abb. 3: In die CR des Tutorial Finder eingetragene Sammlungen

Anschließend werden mithilfe der DME die Daten der Sammlung unter Beachtung ihres disziplinären Kontexts, ihrer Struktur und Komplexität modelliert, Verknüpfungen erstellt und in Form von Zugriffstellen bereitgestellt. Die Generische Suche kann danach auf Datenmodelle und Mappings der DME zugreifen, Informationen zu den Daten weiterverarbeiten, transformieren und integrieren. Hierdurch entstehen Brücken zwischen heterogenen Sammlungen – unabhängig von dort jeweils angebotenen Datenformaten. Volltexte und Metadaten aus TeLeMaCo können so beispielsweise gemeinsam mit den Transkripten aus dem Digital Medieval Webinar Repository (DMWR)¹⁴ durchsucht und ausgewertet werden.

Einmal kontextuell erschlossen und modelliert können Datenmodelle auf weitere integrative Formate abgebildet werden. Im Fall des Tutorial Finders wurde zunächst ein – um die Fähigkeit Volltext aufzunehmen – erweitertes Schema von DataCite¹⁵ gewählt. Über die grundlegende Suchfunktionalität hinausgehende Darstellungen in Form von Wissensgraphen werden derzeit evaluiert und weitergeführt.

Suchfunktionalität des Tutorial Finders

Nachdem eine Sammlung eingetragen ist und ihre Inhalte modelliert sind, kann sie über die Suchfunktionalität des Tutorial Finders durchsucht werden. Hierfür stehen eine *einfache* und eine *erweiterte Suche* zur Verfügung. Die *einfache Suche* (siehe Abbildung 4) erlaubt eine Volltextsuche sowohl auf der Ebene der Primär- als auch der Metadaten – unabhängig von der Existenz von Mappings zwischen den betrachteten Sammlungsdaten. Sie kann auf sämtlichen verfügbaren Sammlungen mit Lehr- und Schulungsmaterialien oder auf einer benutzerdefinierten Teilmenge hieraus ausgeführt werden.

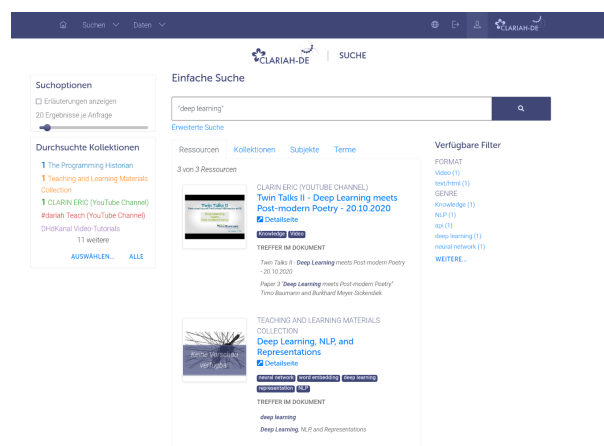


Abb. 4: Suchergebnisse des Tutorial Finders zu „deep learning“ mittels einfacher Suche

Die *erweiterte Suche* (siehe Abbildung 5) bietet neben der Volltextsuche eine gezielte Suche, bei der spezifische Kriterien zur Einschränkung der Suchergebnisse angewendet werden können, die auf dem Schema von DataCite basieren. So können Metadatenfelder des Datenmodells beispielsweise herangezogen werden, um Suchergebnisse nach Orten oder Genres zu filtern.

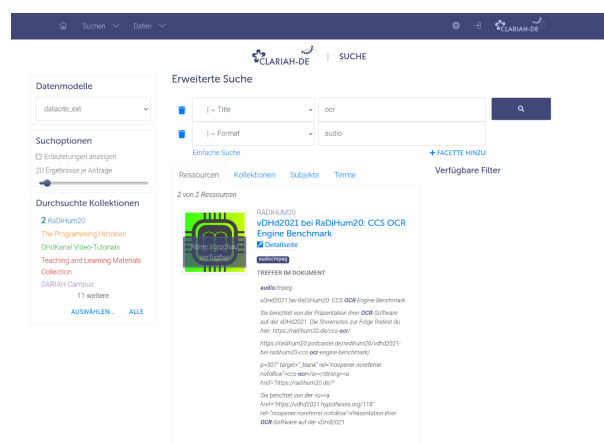


Abb. 5: Gezielte Suche nach Audiodateien zu „ocr“ mittels erweiterter Suche

Wählt man ein Objekt der Ergebnisliste aus, gelangt man zu Informationen über die Zugehörigkeit zu einer Sammlung (siehe Abbildung 6). Insbesondere wird auch die URL präsentiert, unter der das Objekt in seinem ursprünglichen Sammlungskontext zu finden ist. Verfügbare Metainformationen zum Objekt können angesehen und heruntergeladen werden.

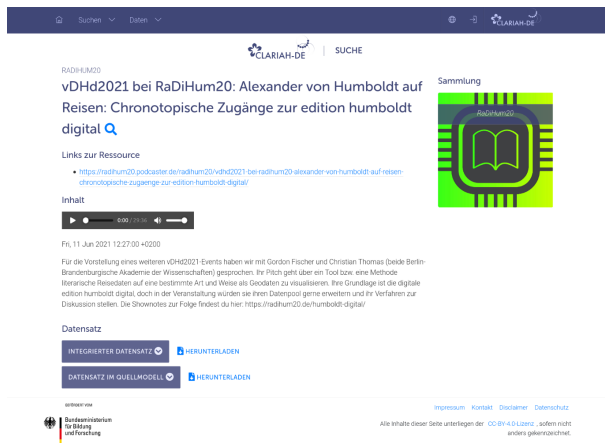


Abb. 6: Detailansicht zu einem Suchergebnis im Tutorial Finder

Dabei hostet der Tutorial Finder die Lehr- und Schulungsmaterialien nicht selbst, sondern ermöglicht bzw. erleichtert lediglich den Zugang auf diese über Metadaten und Volltexte. Es sind daher die zugrunde liegenden Lizenzen für die Nutzung in den unterschiedlichen Lehr- und Lernkontexten zu beachten.

Angebot und Erweiterung des Tutorial Finders

Aktueller Stand

Der Tutorial Finder ist ein offenes Angebot und befindet sich im stetigen Ausbau. Er wurde zunächst als Dienst für CLARIAH-DE konzipiert, der eine übergreifende Suche nach den Materialien aus den zwei vorangegangenen Projekten – CLARIN-D¹⁶ und DARIAH-DE¹⁷ – wie TeLeMaCo von CLARIN oder DARIAH-Campus ermöglichen sollte. Im Laufe der Entwicklung und Erweiterung wurden in den Tutorial Finder Sammlungen und Materialien aufgenommen, die aus anderen für die Community wichtigen nationalen und internationalen Projekten, Vorhaben und Initiativen, z. B. The Programming Historian¹⁸ oder linguisticsweb.org, stammen. Weiterhin finden sich darin einzeln außerhalb jeglicher Sammlungen entwickelte Tutorials und Anleitungen, die jedoch methodisch aufbereitet sind und eine wertvolle Lehr- und Lernquelle zu aktuellen Themen der (Digitalen) Geisteswissenschaften und benachbarter Disziplinen bieten.

Erweiterung des Angebots

Die Sammlung des Tutorial Finders wird nicht nur im Rahmen von Text+ erweitert. Auch die Community der Digital Humanities kann hierzu einen Beitrag leisten und Vorschläge zur Angebots-erweiterung des Tutorial Finders machen: Interessierte, die eine Sammlung von Lehr- und Schulungsmaterialien oder einzelne Tutorials einem möglichst breiten Nutzerkreis zur Verfügung stellen wollen, können diese über den CLARIAH-DE Helpdesk¹⁹ vorschlagen oder sie als authentifizierte Nutzer*innen über den Sammlungseditor der CR (siehe Abbildung 2) im Tutorial Finder selbst anlegen. Sofern erforderlich, kann über den Helpdesk jederzeit auch Unterstützung angefordert werden. So können die

Nutzer*innen das Spektrum der angebotenen Materialien unkompliziert erweitern und über den Tutorial Finder auffindbar, durchsuchbar und zugänglich machen.

Neben der thematischen Relevanz für den Bereich der Digital Humanities bestehen einige grundlegende Voraussetzungen für die Aufnahme einer Sammlung von Lehr- und Schulungsmaterialien in den Tutorial Finder:

- Die Materialien einer Sammlung müssen unter einem **freien Lizenzierungsmodell** zur Verfügung stehen.
- Die Sammlung muss **öffentlich zugänglich** sein.
- Die Sammlung muss **aktuell gehalten** werden.
- Es muss eine **Ansprechperson benannt** werden.

Lizenzauswahl für Lehr- und Schulungsmaterialien

Die für die Lehr- und Schulungsmaterialien am häufigsten genutzten Standardlizenzen sind die Creative-Commons-(CC)-Lizenzen²⁰. Sie werden in den Geistes-, Kultur- und Sozialwissenschaften und in Bildungseinrichtungen allgemein für die Lizenzierung von Texten, Bildern und Videos oder Multimediawerken genutzt. Zu den Vorteilen der CC-Lizenzen gehört, dass sie im Vergleich zu anderen Open-Content-Lizenzen verhältnismäßig einfach zu verstehen sowie international anerkannt und anwendbar sind. Den Nutzenden ermöglichen sie, schnell zu erkennen, ob und unter welchen Bedingungen sie das Material verwenden können. Den Autor*innen bieten sie eine Auswahl von modular aufgebauten Standard-Lizenzen, mit denen sie ihre Materialien nach eigenen Bedürfnissen lizenzieren können (vgl. Kamocki & Ketzan, 2014 für mehr Details zu CC-Lizenzen). Für umfassende Informationen zu verschiedenen Lizenzthemen und zur Auswahl einer passenden Lizenz steht die *CLARIN ERIC Legal Information Platform*²¹ zur Verfügung.

Zugänglichkeit der Sammlung mit Lehr- und Schulungsmaterialien

Die Infrastruktur des Tutorial Finders ermöglicht eine Abfrage von Angeboten einer Sammlung über eine standardisierte Schnittstelle, z.B. OAI-PMH. Die Bereitstellung einer solchen standardisierten Schnittstelle ist jedoch keine Voraussetzung. Wichtig ist nur, dass die Inhalte der Sammlung online zugreifbar sind und so in den Tutorial Finder eingebunden werden können. So verfügt beispielsweise die TeLeMaCo-Sammlung über keine dedizierte Zugriffsschnittstelle. Die Daten wurden in diesem Fall direkt von der Webseite in der DME mithilfe eines iterativen Web-Crawling erschlossen, modelliert und extrahiert und so in den Tutorial Finder übernommen. Ein anderes Beispiel bildet die Anbindung von Sammlungen, die ihre Materialien in einem öffentlich zugänglichen Git-basierten Repository²² zusammenstellen und verwalten, wie The *Programming Historian* oder *DARIAH-Campus*. In diesem Fall kann das Git-Protokoll als Schnittstelle angebunden werden.

Aktualität der Sammlung mit Lehr- und Schulungsmaterialien

Da der Tutorial Finder lediglich den Zugriff auf Sammlungen mit Schulungs- und Lehrmaterialien ermöglicht, die bereits öffentlich zugänglich sind, und sie selbst weder hostet noch pflegt, wird die regelmäßige Überprüfung und Aktualisierung der Samm-

lungsinhalte von den jeweiligen Autor*innen bzw. den dafür zuständigen Personen vollzogen.

Klärung der zuständigen Ansprechpersonen

Bei der Beschreibung einer Sammlung im Sammlungseditor der CR (siehe Abbildung 2) werden die Autor*innen bzw. die dafür zuständigen Ansprechpersonen ggf. auch mit der Angabe der Affiliation genannt. Diese Angaben sind notwendig, um z. B. Fragen zu Hintergründen der Sammlung, Zugriffsmöglichkeiten oder allgemeinen und technischen Voraussetzungen für die Aufnahmen der Sammlung in den Tutorial Finder zu klären.

Zusammenfassung und Ausblick

In diesem Beitrag wurde der Tutorial Finder vorgestellt, der als Angebot des CLARIAH-DE Projekts zum zentralen Durchsuchen von vielfältigen Lehr- und Schulungsmaterialien zu digitalen Methoden, Forschungsdaten, Ressourcen und Diensten im Bereich der Digital Humanities entwickelt wurde. Er ersetzt nicht die bereits vorhandenen Sammlungen von Lehr- und Schulungsmaterialien, da der Tutorial Finder die Materialien nicht selbst hostet, sondern die Suche in verteilten Inhalten ermöglicht, die Sichtbarkeit von Materialien erhöht und ihre Nachnutzbarkeit steigert. Der Tutorial Finder ist für verschiedene Nutzergruppen – Forschende, Lehrende und Studierende aus dem Bereich der Digital Humanities – relevant und kann unter ihrer aktiven Teilnahme erweitert werden. Der Dienst wird derzeit für die weitere inhaltliche und technische Entwicklung in die neu geschaffenen Strukturen der Text+²³ Forschungsdateninfrastruktur überführt und weiterentwickelt – insbesondere mit Blick auf die organisatorische und fachliche Einbindung in die NFDI Konsortien von Text+ und NFDI4Culture²⁴

Fußnoten

1. <https://telemaco.clarin-d.uni-saarland.de/hub/> [letzter Zugriff auf alle Verweise in diesem Dokument am 15. Juli 2021]
2. <https://campus.dariah.eu/>
3. <http://linguisticsweb.org/doku.php?id=start>
4. <https://de.dariah.eu/tei-tutorial>
5. <https://dariah-de.github.io/DARIAH-DKPro-Wrapper/tutorial.html>
6. <https://github.com/DARIAH-DE/TopicsExplorer>
7. CLARIAH-DE war ein vom Bundesministerium für Bildung und Forschung (BMBF) gefördertes, zweijähriges (2019–2021) Verbundprojekt, in dessen Rahmen die Ressourcen der Forschungsinfrastrukturen CLARIN-D und DARIAH-DE zusammengeführt wurden.
8. <https://teaching.clariah.de/search/>
9. <https://dfa.de.dariah.eu/doc/search/>
10. <https://dfa.de.dariah.eu/doc/colreg/>
11. <https://dfa.de.dariah.eu/doc/dme/>
12. Für mehr Informationen zu den Komponenten der GS siehe auch Gradl (2019) sowie Henrich und Gradl (2021).
13. <https://github.com/DARIAH-DE/DCDDM>
14. <https://zenodo.org/communities/dmwr/>
15. <https://schema.datacite.org>
16. CLARIN-D ist der deutsche Beitrag zur *Common Language Resources and Technology Infrastructure* (CLARIN), die als Eu-

ropean Research Infrastructure Consortium (ERIC) ein Netzwerk aus 20 Ländern neben weiteren Partnern bildet, URL: <https://www.clarin-d.net/de>

17. DARIAH-DE ist der deutsche Partner der *Digital Research Infrastructure for the Arts and Humanities* (DARIAH-EU), die mit 19 Ländern und weiteren Partnern ein ERIC auf europäischer Ebene ist, URL: <https://de.dariah.eu>

18. <https://programminghistorian.org/>

19. <https://www.clariah.de/support>; die Überführung in Text+ Strukturen ist in Arbeit

20. <https://creativecommons.org>

21. <https://www.clarin.eu/content/legal-information-platform>

22. Für die ausführliche Anleitung zur Nachnutzung Git-basierter Sammlungen im Rahmen der Infrastrukturdienste von CLARIAH-DE siehe Gradl & Jegan (2021).

23. <https://www.text-plus.org/>

24. <https://nfdi4culture.de/>

Bibliographie

Annisius, Marie / Bock, Sina / Gradl, Tobias / Schopf, Juliane / Stegmeier, Jörn / Werthmann, Antonina (2021): „CLARIAH-DE Work Package 3: Skills Training and Promotion of Junior Researchers“ in: *CLARIAH-DE 2. Vollversammlung / 2nd General Assembly*. Göttingen: Zenodo. DOI: 10.5281/zenodo.4578414

Eckart, Thomas / Werthmann, Antonina / Buddenbohm, Stefan / Sambale, Heidemarie / Helfer, Felix / Jegan, Robin (2021): „CLARIAH-DE Work Package 4 Technical Integration and Coordination of Technical Developments“ in: *CLARIAH-DE 2. Vollversammlung / 2nd General Assembly*. Göttingen: Zenodo. DOI: 10.5281/zenodo.4572610

Gradl, Tobias (2019): „Dokumentation der Datenförderungsarchitektur“ in: *DARIAH-DE Working Papers Nr. 39*. Göttingen: DARIAH-DE, URN: urn:nbn:de:gbv:7-dariah-2019-11-7

Gradl, Tobias / Robin Jegan (2021): „Nachnutzung Git-basierter Sammlungen im Rahmen der Infrastrukturdienste von CLARIAH-DE“ in: *DARIAH-DE Working Papers Nr. 42*. Göttingen: DARIAH-DE, URN: urn:nbn:de:gbv:7-dariah-2021-2-5

Henrich, Andreas / Tobias Gradl (2021): „Integration von Forschungsdaten. Wie können Forschungsinfrastrukturen helfen?“ in: *Innovation in der Bauwirtschaft Innovation in the Building Industry*. Berlin, Boston: De Gruyter, 749-762. DOI: 10.1515/9783110538915-039

Kamocki, Paweł / Erik Ketzan (2014): „Creative commons and language resources: general issues and what's new in CC 4.0“. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3224/file/Kamocki_Ketzan_Creative_Commons_and_Language_Resources_2014.pdf

Der Einsatz von Computer Vision-Methoden für Filme

Eine Fallanalyse für die Kriminalfilm-Reihe Tatort

Schmidt, Thomas

thomas.schmidt@ur.de

Lehrstuhl für Medieninformatik, Universität Regensburg

Kurek, Sarah

sarah.kurek@stud.uni-regensburg.de

Lehrstuhl für Medieninformatik, Universität Regensburg

Einleitung

Quantitative Methoden haben in den Filmwissenschaften eine lange Tradition, die bis auf die prädigitale Ära zurückreichen (Salt 1974; Vonderau 2020). Zahlreiche Projekte in den digitalen Filmwissenschaften setzen mittlerweile computergestützte Methoden ein, um quantitative Analysen durchzuführen oder qualitative Arbeiten zu unterstützen. Anwendungsbereiche sind unter anderem die Analyse von Farben (Burghardt et al. 2016; 2018; Kurzhals et al. 2016; Flueckiger 2017; Pause / Walkowski 2018; Masson et al. 2020;), Annotationsmöglichkeiten (Kuhn et al. 2015; Halter et al. 2019; Schmidt / Halbhuber 2020; Schmidt et al. 2020a) oder Schnittlängen und -typen (DeLong 2015; Baxter et al. 2017). Häufig werden dabei die Texte von Filmen (Skripte, Untertitel) analysiert (Hoyt et al. 2014; Holobut et al. 2016; Byszuk 2020; Holobut / Rybicki 2020). Durch Entwicklungen im Bereich der Computer Vision (computergestützte Bilderkennung/Bildanalyse) (CV) bieten sich jedoch neue Möglichkeiten für Digital Humanities (DH)-Projekte, die mit Videos arbeiten, die bereits erfolgreich für Filme, Internetvideos oder Theateraufführungen eingesetzt werden (Zaharieva et al. 2012; Howanitz et al. 2019; Pusturien et al. 2020; Schmidt et al. 2021c; Schmidt / Wolff 2021). Wir präsentieren im folgenden Beitrag eine explorative Studie zum Einsatz einer Auswahl an CV-Methoden, die wir für potentiell wertvoll für den Bereich der Spielfilm-Analyse einschätzen. Wir orientieren uns dabei am explorativen Forschungsansatz definiert von Wulff (1998) für die Filmanalyse.

Als Fallstudie wird die deutschsprachige Kriminalfilm-Reihe „Tatort“ gewählt. Mit ca. 9 Millionen Zuschauern handelt es sich um eine der beliebtesten Fernsehformate in Deutschland.¹ Aufgrund seiner national hohen kulturellen Bedeutung ist der Tatort ein häufiger Untersuchungsgegenstand in der Filmanalyse (Buhl 2013) und wird zur Analyse gesellschaftspolitischer Themen wie Migration (Ortner 2007), Verhältnis von Ost- und Westdeutschland (Welke 2005), des Zusammenhangs von Emotionen und Geschlechtern (Finger et al. 2010) oder zur Analyse von Online-Texten herangezogen (Schmidt et al. 2021d). Der Fokus unserer Analysen liegt auf gruppenbasierten Vergleichen. Als Gruppen differenzieren wir zwischen unterschiedlichen Städten/ErmittlerInnen-Teams. So spielen die einzelnen Folgen des Tatorts in unterschiedlichen Städten mit unterschiedlichen Hauptfiguren. Diese Gruppen transportieren teilweise unterschiedliche

Stimmungen, Lokalkolorit sowie Geschlechts- und Altersrepräsentationen in den Figurenkonstellationen.

Die Ziele dieses Beitrags sind (1) Nutzen und Limitationen der angewandten Methoden zu analysieren und (2) explorativ festzustellen, ob die Methoden besondere Charakteristiken von, in diesem Fall, Filmgruppen aufzeigen. Als CV-Methoden werden Objekt-, Emotions-, Geschlechts-, Alters- und Ortserkennung untersucht und frei verfügbare state-of-the-art-Modelle verwendet. Die genannten Methoden wurden ausgewählt, weil sie als gewinnbringend für Forschungsideen auf dem vorliegenden Korpus angesehen werden und bereits in ähnlichen Settings exploriert wurden (Schmidt et al. 2021c).

Korpus

Als Korpus werden 13 Folgen des Tatorts genutzt. Alle Filme (je ca. 90 Minuten) liegen im mp4-Format mit einer Auflösung von 960x540 Pixeln und 25 Frames pro Sekunde vor. Alle angewandten CV-Methoden nutzen Bild-Dateien weswegen wir 1 Frame für jede Sekunde eines Films extrahieren und als Korpusgrundlage verwenden. Abbildung 1 fasst die wichtigsten Metadaten der Filme zusammen.

ID	Folge	Titel	Ausstrahlungsdatum	Extrahierte Frames
N1	1085	Ein Tag wie jeder andere	24.02.19	5 255
SW1	1087	Für immer und dich	10.03.19	5 368
M1	1096	Die ewige Welle	26.05.19	5 342
CH 1	1099	Ausgezählt	16.06.19	5 306
CH 2	1106	Der Elefant im Raum	27.10.19	5 174
M2	1114	One Way Ticket	26.12.19	5 320
M3	1118	Unklare Lage	26.01.20	5 340
SW2	1121	Ich hab im Traum geweinet	23.02.20	5 373
N2	1122	Die Nacht gehört dir	01.03.20	5 267
M4	1135	Lass den Mond am Himmel stehn	07.06.20	5 249
SW3	1138	Rebland	27.09.20	5 344
M5	1146	In der Familie (Teil 1)	29.11.20	5 338
M6	1147	In der Familie (Teil 2)	06.12.20	5 325

Abb. 1: Metadaten des ausgewählten Tatort-Korpus.

Wir differenzieren zwischen den folgenden Standorten/ErmittlerInnen-Teams, die im Folgenden für gruppenbasierte Vergleiche genutzt werden: Luzern in der Schweiz (im Folgenden abgekürzt als CH; insgesamt 2 Filme), München (M; 6 Filme), Nürnberg (N; 2 Filme) und Schwarzwald (SW; 3 Filme). Die 4 Gruppen unterscheiden sich bezüglich des Kolorits (ländlich vs städtisch) und in der Alters- und Geschlechtsausprägung. Aufgrund der ungleichen Menge an Filmen pro Gruppe werden im Folgenden primär Werte normalisiert an der Länge (pro Sekunde, gewählte Frames für das Korpus) betrachtet.

Objekterkennung

Für die Objekterkennung wird Detectron2 von Facebook AI Research verwendet, was als state-of-the-art-Lösung gilt (Wu et al. 2019). Das Modell basiert auf einem vortrainierten maskierten RCCN-Modell und wurde auf dem COCO-Datensatz (Lin et al. 2015) trainiert. Es kann 80 Objektklassen wie Fahrzeuge oder

Tiere vorhersagen. Als Schwellenwert für die Erkennung wird eine Wahrscheinlichkeit von 50% gewählt. Dies ist ein eher niedriger Wert, ermöglicht uns aber breitere Explorationen der Methode. Diese und alle anderen Methoden wurden in *Python* mit verschiedenen SDKs implementiert.

Die Abbildungen 2-3 illustrieren die je 10 häufigsten Objekte pro Tatort-Gruppe und insgesamt. Wir interpretieren die Ergebnisse rein deskriptiv.

CH			M			N		
Objekt	#	%	Objekt	#	%	Objekt	#	%
Person	21 295	91,53	Person	61 944	90,26	Person	21 000	89,26
Stuhl	2 258	14,26	Stuhl	8 671	13,37	Stuhl	4 493	18,26
Tasse	1 640	10,24	Auto	5 369	7,42	Buch	3 516	7,86
Krawatte	1 439	10,78	Buch	4 296	4,69	Auto	3 001	9,48
Buch	1 269	4,83	Flasche	3 692	5,28	Tasse	1 241	7,70
Auto	1 188	6,16	Tasse	3 652	6,87	Krawatte	1 007	8,19
Laptop	718	5,62	Esstisch	2 511	6,10	Esstisch	986	7,13
Fernseher	714	6,14	Fernseher	2 265	5,38	Mobiltelefon	759	5,89
Esstisch	665	5,44	Topfpflanze	1 841	4,20	Flasche	636	3,58
Flasche	629	3,30	Krawatte	1 696	4,84	Fernseher	509	4,56

Abb. 2: Verteilung von erkannten Objekten in der CH, M und N-Gruppe. # ist die absolute Zahl. % der Anteil an den gewählten Frames des jeweiligen Gruppenkorpus.

SW			Gesamt		
Objekt	#	%	Objekt	#	%
Person	21 295	91,53	Person	138 354	91,36
Stuhl	2 258	14,26	Stuhl	22 764	16,28
Tasse	1 640	10,24	Buch	17 194	6,54
Krawatte	1 439	10,78	Auto	12 067	7,76
Buch	1 269	4,83	Tasse	9 465	8,52
Auto	1 188	6,16	Flasche	7 867	5,68
Laptop	718	5,62	Esstisch	6 092	6,67
Fernseher	714	6,14	Krawatte	4 768	5,93
Esstisch	665	5,44	Topfpflanze	4 421	4,57
Flasche	629	3,30	Fernseher	4 420	5,25

Abb. 3: Verteilung von erkannten Objekten in der SW-Gruppe und insgesamt. # ist die absolute Zahl. % der Anteil an den gewählten Frames des jeweiligen Gruppenkorpus.

Zu den häufigsten erkannten „Objekten“ gehören (trivialerweise) Personen. Ansonsten wird vor allem Innenarchitektur (Stühle, Tassen, Bücher) erkannt aber auch Bildschirme wie Laptops und Fernseher. Diese Erkennungen basieren vor allem auf Szenen der Recherchen der einzelnen Teams; Bildschirme werden dabei häufig als Fernseher erkannt (siehe Abbildung 4).

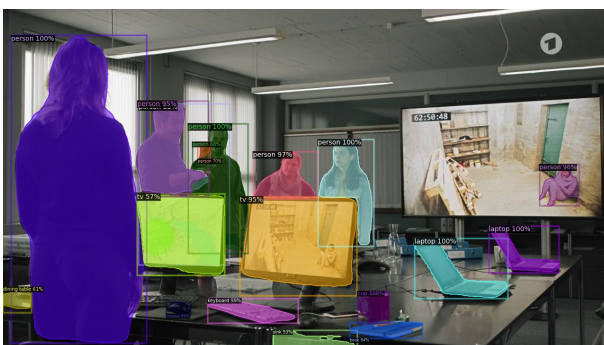


Abb. 4: Frame mit als Fernseher erkannten Computerbildschirmen (CH1).

Verhältnismäßig wenig Objekte weisen auf Außenszenen hin wie zum Beispiel Autos (Abbildung 5).



Abb. 5: Frame mit als Auto erkannten Objekten (M5)

Die Unterschiede der einzelnen Tatort-Städte unseres Korpus sind eher gering und die Verteilung sehr homogen. Vereinzelt können Objekthäufungen aufgrund von sehr speziellen inhaltlichen Unterschieden einzelner Folgen festgestellt werden. Die SW-Episoden beispielsweise weisen, im Unterschied zu den anderen Gruppen, vermehrt Betten auf, da eine der Folgen (SW2) in einem Hotel spielt (siehe Abbildung 6).

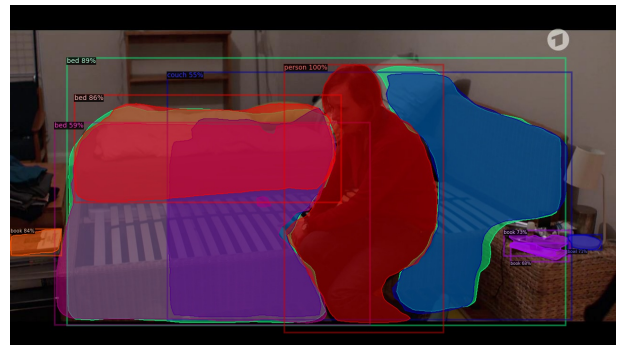


Abb. 6: Frame mit als Bett erkannten Objekten (SW2).

Es wurde keine systematische Evaluation durchgeführt, jedoch konnte man in der explorativen Analyse feststellen, dass die meisten Ergebnisse korrekt sind. Falsche Zuweisungen sind jedoch nicht selten, wenngleich die Fehlinterpretation häufig nachvollziehbar ist (siehe Abbildung 7).

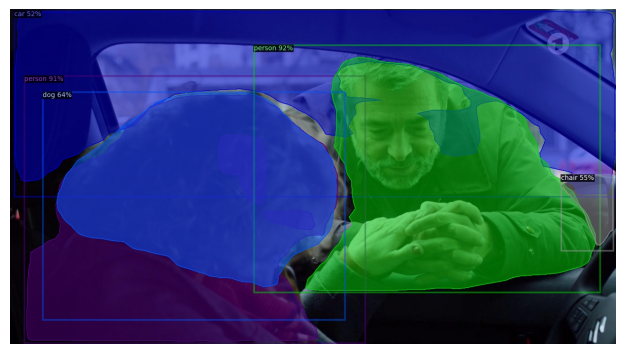


Abb. 7: Als „Hund“ erkannte Haare einer Figur (CH2).

Figurenanalyse

Unter Figurenanalyse bezeichnen wir im Folgenden alle Methoden, die die Gesichter der Figuren als Analyseelement benutzen: Emotions-, Alters- und Geschlechtererkennung.

Emotionserkennung

Gesichtsbasierte Emotionserkennung ist eine etablierte Methode in der Mensch-Maschine-Interaktion mit zahlreichen Anwendungsbeispielen (Halbhuber et al. 2019; Hartl et al. 2019; Schmidt et al. 2020c). Für die Emotionserkennung wird das Python-Modul *FER* (Goodfellow et al. 2013) genutzt. Das Modell führt erste eine Gesichtserkennung durch (Zhang et al. 2016) und dann eine Emotionsprädiktion über ein convolutional neural network (CNN), das auf über 35 000 vorannotierten Bildern trainiert wurde. Das Modell gibt Wahrscheinlichkeitswerte für die sieben Klassen *Wut*, *Ekel*, *Furcht*, *Freude*, *Trauer*, *Überraschung* und *Neutral* zwischen 0 und 1 aus, die sich insgesamt zu 1 summieren. Zur Bestimmung der Gesamtemotion eines Frames werden die jeweiligen Werte für die Kategorien summiert und der Durchschnitt gebildet. Für die film- oder gruppenbasierten Analysen werden Mittelwerte über alle Frames hinweg gebildet.

Abbildung 8 illustriert die wichtigsten statistischen Werte dieser Auswertung.

Emotion	Wert	CH	M	N	SW	Gesamt
Wut	M	0,17	0,19	0,18	0,19	0,18
	Max	0,95	0,97	0,96	0,97	0,97
Ekel	M	0,01	0,01	0,01	0,01	0,01
	Max	0,73	0,76	0,77	0,87	0,87
Furcht	M	0,10	0,10	0,10	0,10	0,10
	Max	0,78	0,86	0,82	0,84	0,86
Freude	M	0,11	0,11	0,12	0,16	0,12
	Max	1,00	1,00	1,00	1,00	1,00
Neutral	M	0,29	0,25	0,25	0,21	0,25
	Max	0,99	0,98	0,97	0,99	0,99
Trauer	M	0,30	0,31	0,31	0,31	0,31
	Max	0,96	0,99	0,97	0,97	0,99
Überraschung	M	0,03	0,04	0,03	0,03	0,03
	Max	0,99	1,00	0,90	0,99	0,06

Abb. 8: Deskriptive Statistik für die Emotionserkennung. Minimalwerte sind stets 0. Höchster M (Durchschnitt) pro Emotion für die Gruppen ist hervorgehoben.

Die häufigsten Emotionsklassen sind Trauer(M=0,31) (siehe Abbildung 9), Neutral (M=0,25) und Wut (M=0,18). Dies ist eine passende Verteilung für die Grundstimmung von Kriminalfilmen. Überraschung und Ekel werden eher selten vorhergesagt. Die Tendenz zu negativem Sentiment und Emotionen findet man bei der Annotation und Prädiktion von anderen narrativen Erzählformen auch (Schmidt / Burghardt 2018; Schmidt 2019; Schmidt et al. 2019; 2021a; 2021b)

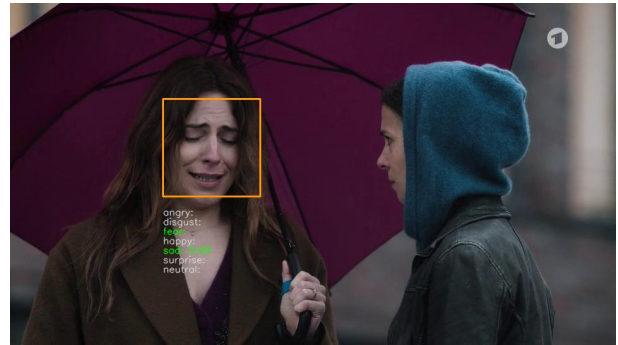


Abb. 9: Frame mit dem höchsten Wert für Trauer (0,99) im Gesamtkorpus (M5)

Deskriptiv betrachtet sind die Ergebnisse der einzelnen Episodengruppen erneut sehr homogen. Für Output mit kontinuierlichen Werten überprüfen wir die Unterschiede aber noch mittels Signifikanztests. Wir verwenden einen *Welch-ANOVA-Test* (alle Voraussetzungen für den Test sind erfüllt (Field 2009)) und finden signifikante Unterschiede gemäß eines Signifikanzniveaus von $p < 0,05$ (Abbildung 10).

	p-Wert	F-Wert	η^2
Wut	<0,01	13,02	<0,01
Ekel	<0,01	13,12	<0,01
Angst	<0,01	25,06	<0,01
Freude	<0,01	93,52	0,01
Neutral	<0,01	127,40	0,01
Trauer	<0,01	4,31	<0,01
Überraschung	<0,01	35,43	<0,01

Abb. 10: Ergebnisse des Welch-ANOVA-Signifikanztest für die Episodengruppen (Emotionen).

Die Effekte der Unterschiede bestätigen jedoch die deskriptive Interpretation, da sie laut Cohen (1988) als sehr gering einzustufen sind ($\eta^2 < 0,01$ = schwacher, $< 0,06$ moderater und $< 0,14$ = starker Effekt). Auch Post-Hoc-Tests unter den einzelnen Gruppen weisen zwar signifikante Unterschiede auf, sind jedoch geringfügig.

Die explorative Evaluation des Materials zeigt, dass die Modelle für extreme Emotionsausprägungen nachvollziehbare Ergebnisse produzieren (siehe auch Abbildung 11), ein Hauptproblem jedoch ist, dass Fehler in der vorangestellten Gesichtserkennung vorkommen. So hat das Modell große Probleme mit der Erkennung von Gesichtern, die nicht frontal in die Kamera blicken. Dies liegt der Tatsache zu Grunde, dass die Modelle primär mit Bildern trainiert werden bei denen die Personen frontal in die Kamera blicken (Goodfellow et al. 2013).

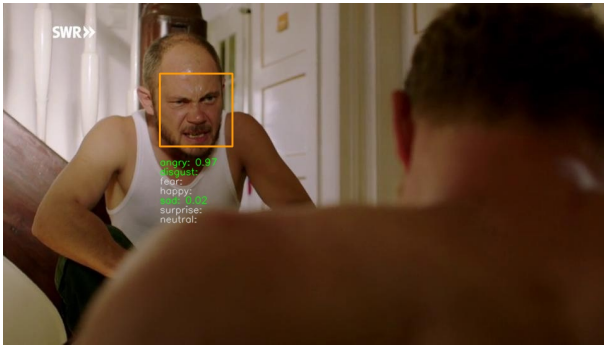


Abb. 11: Frame mit dem höchsten Wert für Wut (0,97) im Gesamtkorpus (SW1).



Abb. 14: Frame mit der Person mit dem geringsten Alterswert von 10,86 (SW 3).

Alters- und Geschlechtserkennung

Die Vorhersage des Alters und des Geschlechts von Figuren wird mit dem Modul *py-agernder*² durchgeführt. Es basiert auf einem CNN, das auf dem IMDB-Wiki-Datensatz, bestehend aus über 500 000 annotierten Gesichtern (Rothe et al. 2018), trainiert wurde und erzielt in Evaluationen sehr gute Ergebnisse (Agustsson et al. 2017). Die Altersprädiktion gibt einen Wert zwischen 0 und 100 aus, der das Alter kennzeichnet. Die Geschlechtsprädiktion einen Wert zwischen 0 und 1 für den gilt, <0,5 eher männlich und >0,5 eher weiblich. Für beide Verfahren wurde für jeden Frame der Mittelwert aller erkannten Gesichter für einen Gesamtwert gebildet. In Abbildung 12 werden die Ergebnisse für beide Methoden gesamt und pro Tatortgruppe zusammengefasst.

	Wert	CH	M	N	SW	Gesamt
Alter	M	42,98	42,10	41,15	39,08	41,47
	Max	71,87	75,46	69,50	73,11	75,46
Geschlecht	M	0,38	0,34	0,41	0,44	0,38
	Max	0,99	0,99	0,99	0,99	0,99

Abb. 12: Deskriptive Statistik für die Alters- und Geschlechtserkennung. Maximal- und Minimalwerte von M werden pro Episodengruppe hervorgehoben.

Gemäß der Alterserkennung liegt der Altersdurchschnitt bei 41,47 Jahren. Die dominanten und häufig in Frames gezeigten ErmittlerInnen der ausgewählten Folgen sind jedoch überwiegend Ende 40 und Anfang 50. Die älteste Person im Gesamtkorpus wird mit 72 Jahren identifiziert (Abbildung 13), die jüngste ist ein Kind mit 10 Jahren (Abbildung 14).

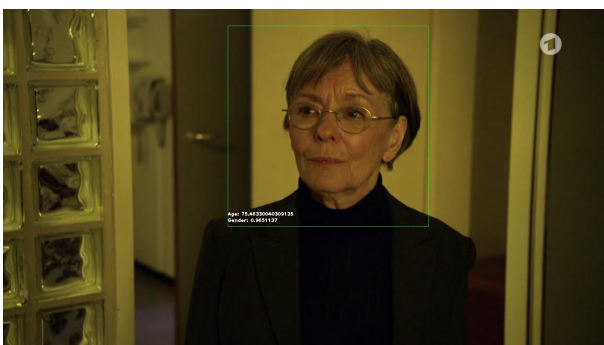


Abb. 13: Frame mit der Person mit dem höchsten Alterswert von 75,46 (M2).

Rein deskriptiv sind die Unterschiede zwischen den Gruppen gering. Ein Welch-ANOVA-Test weist erneut auf signifikante Unterschiede mit einem geringen Effekt hin (Abbildung 15).

	p-Wert	F-Wert	η^2
Alter	<0,001	358,43	0,02
Geschlecht	<0,001	124,69	0,02

Abb. 15: Ergebnisse des Welch-ANOVA-Signifikanztest für die Episodengruppen (Geschlecht/Alter).

Tatsächlich zeigen Post-Hoc-Tests, dass die Hauptunterschiede mit einem mittleren Effekt zwischen den Folgen aus der Schweiz (CH) und aus dem Schwarzwald (SW) bestehen, welche gleichzeitig den höchsten, respektive geringsten Altersunterschied haben. Bei Betrachtung der Filme wird klar, dass dies vor allen daran liegt, dass in den SW-Folgen viele Kinder und Jugendliche mitspielen (Abbildung 14). Ein Problem der Altersanalyse, das wir bei unseren Explorationen identifizieren konnten, ist jedoch in diesem Zusammenhang, dass Kinder und Jugendliche meist überschätzt werden (Abbildung 15). Grund hierfür ist auch wieder die Trainingsgrundlage des Modells, die primär aus Personen im Erwachsenenalter besteht.

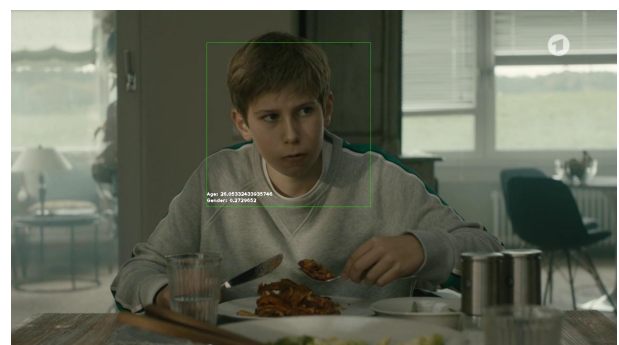


Abb. 16: Frame mit Kind, dessen Alter auf 26 Jahre überschätzt wird (M4)

Mit einem Mittelwert von 0,38 ist eine vermehrte Repräsentation männlicher Gesichter festzustellen (Abbildung 12). Dies entspricht auch der realen Figuren-Belegung der Serie, die, obschon sie gemischte ErmittlerInnen-Paare enthält, vor allem in den Nebenfiguren von männlichen Charakteren dominiert wird. Abbildung 17 und 18 zeigen die jeweils höchsten Ausprägungen des Korpus.



Abb. 17: Frame mit dem „männlichsten“ Gesicht (Minimalwert für Geschlecht: 0,002) (M2).



Abb. 19: Beispiel für falsche Geschlechtererkennung: Das Gesicht wird als männlich (0,29) identifiziert (M6).

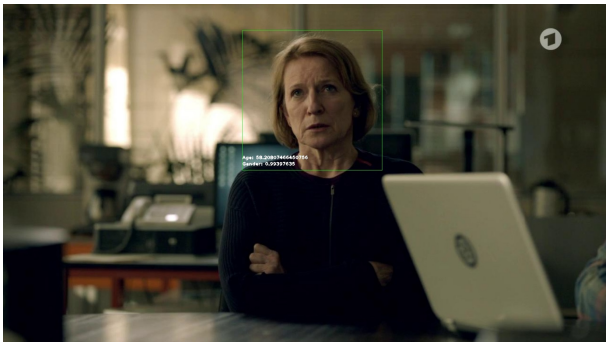


Abb. 18: Frame mit dem „weiblichsten“ Gesicht (Minimalwert für Geschlecht: 0,99) (N1).

Ein Welch-ANOVA-Test zeigt wiederum signifikante Werte mit schwachen Effekten auf (Abbildung 15). Post-Hoc-Tests zeigen, dass der signifikante Unterschied aufgrund der stärkeren Differenzen der Episoden aus München (M) mit den Episoden aus Schwarzwald (SW) und Luzern (CH) zustande kommt. In der Tat sind die beiden letztgenannten Gruppen jene, die ErmittlerInnen-Gruppen bestehend aus Mann und Frau haben und damit höhere Werte Richtung weiblicher Gesichter zeigen. Der erhöhte Wert bezüglich männlicher Ausprägung beim Münchner-Tatort ist zum einen konform mit der Dominanz an männlichen Ermittlern und wird bei der Einzelfolgen-Analyse deutlich, da eine Folge in einem Männergefängnis spielt.

Insgesamt wirkt die Geschlechtsprädiktion plausibel. Ähnlich zur Emotionserkennung ist ein Problem mangelnde korrekte Gesichtserkennung aufgrund nicht-frontaler Kamerawinkel und schwammige, dunkle Einstellungen (Abbildung 19).

Ortserkennung

Als Ortserkennung bezeichnen wir die Methodik den groben Schauplatz eines Bildes zu erkennen. Dabei ist nicht der geographische Ort gemeint, sondern die abstrakte Umgebung, also zum Beispiel, ob ein Bild in einem Zimmer spielt oder draußen. Wir verwenden den Trainingsdatensatz *Places365*³, der aus über 1,8 Millionen annotierten Bildern besteht. Für unsere Prädiktion nutzen wir ein vortrainiertes CNN und präparieren die Frames in einer Vorverarbeitung für das CNN (Zhou et al. 2017). Anstatt den 365 Teilklassen fokussieren wir uns jedoch auf die Hauptkategorien *innen*, *draußen-künstlich*, *draußen-natürlich* und *draußen-gemischt*. Jeden Frame weisen wir die Kategorie zu, die das Modell mit der höchsten Wahrscheinlichkeit vorhersagt.

	Wert	CH	M	N	SW	Gesamt
innen	#	9 363	26 403	9 247	12 765	57 778
	%	89,34	82,73	87,88	79,36	83,74
draußen-künstlich	#	809	3 650	982	2 014	7 455
	%	7,72	11,44	9,33	12,52	10,80
draußen-natürlich	#	207	1 265	189	735	1 372
	%	1,98	3,96	1,80	4,57	3,47
draußen-gemischt	#	101	596	104	571	2 396
	%	0,96	1,87	0,99	3,55	1,99

Abb. 20: Häufigkeitsverteilung für die Ortserkennung. # ist die absolute Zahl. % der Anteil an den gewählten Frames des jeweiligen Gruppenkorpus.

Unabhängig von der Episodengruppe wird der größte Anteil der Frames als innen kategorisiert (Abbildung 20), was der Realität der Filme entspricht in denen meist Ermittlungen und Recherchen in Zimmern stattfinden (Abbildung 21).



Abb. 21: Beispiel für Frame, das als „innen“ erkannt wurde (CH1).

Ein Chi-Quadrat-Signifikanz-Test weist dennoch auf signifikante Unterschiede zwischen den Gruppen hin ($\chi^2 = 809,23$; $p < 0,001$; $\varphi = 0,06$). In der Tat weisen die Episoden aus dem Schwarzwald als einer eher ländlichen Gegend den höchsten Anteil an Frames der Kategorie „draußen“ auf (Abbildung 22). Die CH-Gruppe, die von uns auch als eher ländlich postuliert wurde, kann dies hier aber nicht bestätigen, was jedoch inhaltlich plausibel ist, da beispielsweise eine Folge komplett in den Räumen eines Schiffes spielt.



Abb. 22: Beispiel für einen Frame, der als gemischt-draußen erkannt wurde (SW1).

Methodenreflexion und Ausblick

Durch die Durchführung der hier vorgestellten Fallstudie, konnten wir die CV-Methoden gewinnbringend explorieren. Wir konnten signifikante Unterschiede zwischen Episodengruppen identifizieren, die jedoch meist geringe Effekte aufzeigten. Die meisten Charakteristika, die wir so herausarbeiten konnten, bestätigten Annahmen. Neue Auffälligkeiten der Filmgruppen konnten jedoch nicht entdeckt werden. Die Gründe hierfür sind vielseitig: Das Korpus ist, da es sich um die gleiche Serie nur mit variierenden Figuren handelt, bezüglich des Genres, des Settings und der Besetzung eventuell zu homogen um gruppenbasierte Unterschiede in Signifikanztests deutlich zu machen. Vergleiche von Filmgruppen, die sich klarer voneinander unterscheiden (z.B. unterschiedliche Filmgenres), könnten deutlichere Effekte generieren.

Trotzdem haben wir vielversprechende Forschungsideen, die mit den präsentierten Methoden in einer Art *Distant Viewing*-Ansatz (Arnold / Tilton 2019) mit ausreichend Filmmaterial un-

tersucht werden können, z.B. für den Bereich Gender Studies Korrelationen zwischen Emotionen und Geschlechtern oder Repräsentationsanalysen der Geschlechter (ähnlich zu Schmidt et al. 2020b). Die momentane Methodenauswahl ist auch noch rein bildfokussiert, wenngleich andere Kanäle z.B. der Audio-Kanal auch Potential für die Analyse haben. In der Tat werden in ersten Projekten in den DH der Einsatz von multimodalen Methoden oder dem Audio-Kanal bereits untersucht (Ortloff et al. 2019; Schmidt et al. 2019; Schmidt / Wolff 2021).

Die Exploration der Methoden für die vorliegende Fallstudie haben jedoch auch Probleme in der Exaktheit und Leistung offenbart, z.B. Probleme in der Gesichtserkennung. Systematische Evaluationen sind notwendig, um das Ausmaß der Problematik einschätzen zu können. Auch sind die Klassifikationstaxonomien, beispielsweise der Objekterkennung und Ortserkennung, eventuell nicht passend für die Interessen von FilmwissenschaftlerInnen. Wir planen momentan größere Annotationsstudien, um (1) die Leistung von state-of-the-art-Standard-Modellen exakt zu evaluieren und (2) Trainingsmaterial für die Domänenadaption an eine spezielle Filmdomäne zu erstellen. Für die Annotationsstudien sollen studentische Hilfskräfte größere Mengen eines Querschnitts von Filmframes aus Filmen unterschiedlicher Epochen und Genres annotieren.

Fußnoten

1. <https://de.statista.com/statistik/daten/studie/377327/umfrage/fernsehzuschauer-der-krimireihe-tatort/>
2. <https://pypi.org/project/py-agender/>
3. <https://github.com/CSAILVision/places365>

Bibliographie

- Agustsson, Eiríkur / Timofte, Radu / Escalera, Sergio / Baro, Xavier / Guyon, Isabelle / Rothe, Rasmus (2017): "Apparent and Real Age Estimation in Still Images with Deep Residual Regressors on Appa-Real Database", in: *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 87–94. DOI: 10.1109/FG.2017.20
- Arnold, Taylor / Tilton, Lauren (2019): "Distant viewing: Analyzing large visual corpora", in: *Digital Scholarship in the Humanities*. DOI: 10.1093/digitalsh/fqz013
- Baxter, Mike / Khitrova, Daria / Tsivian, Yuri (2017): "Exploring cutting structure in film, with applications to the films of D. W. Griffith, Mack Sennett, and Charlie Chaplin", in: *Digital Scholarship in the Humanities*, 32(1):1–16. DOI: 10.1093/llc/fqv035
- Buhl, Hendrik (2013): "Tatort: gesellschaftspolitische Themen in der Krimireihe", in: *Alltag, Medien und Kultur*. Band 14. UVK, Konstanz.
- Burghardt, Manuel / Kao, Michael / Walkowski, Niels-Oliver (2018): "Scalable MovieBarcodes—An Exploratory Interface for the Analysis of Movies.", in: *IEEE VIS Workshop on Visualization for the Digital Humanities* (Vol. 2).
- Burghardt, Manuel / Kao, Michael / Wolff, Christian (2016): "Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie Analysis", in: *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków: 753–755.

- Byszuk, Joanna** (2020): "The Voices of Doctor Who – How Stylometry Can be Useful in Revealing New Information About TV Series", in: *Digital Humanities Quarterly*, 014(4).
- Cohen, Jacob** (1988): *Statistical power analysis for the behavioral sciences*. Academic press.
- DeLong, Jordan** (2015): "Horseshoes, handgrenades, and model fitting: The lognormal distribution is a pretty good model for shot-length distribution of Hollywood films", in: *Literary and Linguistic Computing*, 30(1):129–136. DOI: 10.1093/lilc/fqt030
- Field, Andy P.** (2009): *Discovering statistics using SPSS: And sex, drugs and rock „n“ roll* (3rd ed). SAGE Publications.
- Finger, Juliane / Unz, Dagmar C. / Schwab, Frank** (2010): "Crime Scene Investigation: The Chief Inspectors' Display Rules", in: *Sex Roles*, 62(11-12):798–809.
- Flueckiger, Barabara** (2017): "A Digital Humanities Approach to Film Colors", in: *The Moving Image: The Journal of the Association of Moving Image Archivists*, 17(2): 71–94. JSTOR. DOI: 10.5749/movingimage.17.2.0071
- Goodfellow, Ian J. et al.** (2013): *Challenges in Representation Learning: A report on three machine learning contests*. arXiv:1307.0414 [cs, stat]. <http://arxiv.org/abs/1307.0414> [14.06.2021]
- Halbhuber, David / Fehle, Jakob / Kalus, Alexander / Seitz, Konstantin / Kocur, Martin / Schmidt, Thomas / Wolff, Christian** (2019): "The Mood Game - How to use the player's affective state in a shoot'em up avoiding frustration and boredom", in: Alt, Florian / Bulling, Andreas / Döring, Tanja (eds.), *Mensch und Computer 2019 - Tagungsband*. New York: ACM. DOI: 10.1145/3340764.3345369
- Halter, Gaudenz / Ballester-Ripoll, Rafael / Flueckiger, Barabara / Pajarola, Renato** (2019): "VIAN: A Visual Annotation Tool for Film Analysis", in: *Computer Graphics Forum*, 38(3): 119–129. DOI: 10.1111/cgf.13676
- Hartl, Philipp / Fischer, Thomas / Hilzenthaler, Andreas / Kocur, Martin / Schmidt, Thomas** (2019): "AudienceAR - Utilising Augmented Reality and Emotion Tracking to Address Fear of Speech", in: Alt, Florian / Bulling, Andreas / Döring, Tanja (eds.), *Mensch und Computer 2019 - Tagungsband*. New York: ACM. DOI: 10.1145/3340764.3345380
- Holobut, Agata / Rybicki, Jan / Wozniak, Monika** (2016): "Stylometry on the Silver Screen: Authorial and Translational Signals in Film Dialogue", in: *Book of Abstracts of the International Digital Humanities Conference (DH) (2016)*.
- Holobut, Agata / Rybicki, Jan** (2020): "The Stylometry of Film Dialogue: Pros and Pitfalls", in: *Digital Humanities Quarterly*, 014(4).
- Howanitz, Gernot / Bermeitinger, Bernhard / Radisch, Erik / Sebastian Gassner / Rehbein, Malte / Handschuh, Siegfried** (2019): "Deep Watching - Towards New Methods of Analyzing Visual Media in Cultural Studies", in: *Book of Abstracts of the International Digital Humanities Conference (DH) (2019)*.
- Hoyt, Eric / Ponto, Kevin / Roy, Carrie** (2014): "Visualizing and Analyzing the Hollywood Screenplay with ScripThreads", in: *Digital Humanities Quarterly*, 008(4).
- Kuhn, Virginia / Craig, Alan / Simeone, Michael / Satheesan, Simeone P. / Marini, Luigi** (2015): "The VAT: Enhanced video analysis", in: *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, 1–4. DOI: 10.1145/2792745.2792756
- Kurzhals, Kuno / John, Markus / Heimerl, Florian / Kuznetsov, Paul / Weiskopf, Daniel** (2016): "Visual Movie Analytics", in: *IEEE Transactions on Multimedia*, 18(11): 2149–2160. DOI: 10.1109/TMM.2016.2614184
- Lin, Tsung-Yi / Maire, Michael / Belongie, Serge / Bourdev, Lubomir / Girshick, Ross / Hays, James / Perona, Pietro / Ramanan, Deva / Zitnick, C. Lawrence / Dollár, Piotr** (2015): "Microsoft COCO: Common Objects in Context". arXiv:1405.0312 [cs]. <http://arxiv.org/abs/1405.0312> [14.06.2021]
- Masson, Eef / Olesen, Christian G. / Noord, Nanne van / Fossati, Giovanna** (2020): "Exploring Digitised Moving Image Collections: The SEMIA Project, Visual Analysis and the Turn to Abstraction.", in: *Digital Humanities Quarterly*, 014(4).
- Ortloff, Anna-Marie / Güntner, Lydia / Windl, Maximiliane / Schmidt, Thomas / Kocur, Martin / Wolff, Christian** (2019): "SentiBooks: Enhancing Audiobooks via Affective Computing and Smart Light Bulbs", in: Alt, Florian / Bulling, Andreas / Döring, Tanja (eds.), *Mensch und Computer 2019 - Tagungsband*. New York: ACM. DOI: 10.1145/3340764.3345368
- Ortner, Christina** (2007): "Tatort: Migration. Das Thema Einwanderung in der Krimireihe Tatort", in: *Medien & Kommunikationswissenschaft*, 55(1):5–23.
- Pause, Johannes / Walkowski, Niels-Oliver** (2018): "Everything is illuminated. Zur numerischen Analyse von Farbigkeit in Filmen", in: *Zeitschrift für digitale Geisteswissenschaften*.
- Pustu-Iren, Kader / Sittel, Julian / Mauer, Roman / Bulgakowa, Oksana / Ewerth, Ralph** (2020): "Automated Visual Content Analysis for Film Studies: Current Status and Challenges", in: *Digital Humanities Quarterly*, 014(4).
- Rothe, Rasmus / Timofte, Radu / Van Gool, Luc** (2018): "Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks", in: *International Journal of Computer Vision*, 126(2):144–157. DOI: 10.1007/s11263-016-0940-3
- Salt, Barry** (1974): "Statistical style analysis of motion pictures.", in: *Film Quarterly*, 28(1): 13–22.
- Schmidt, Thomas** (2019): "Distant Reading Sentiments and Emotions in Historic German Plays", in: *Abstract Booklet, DH_Budapest_2019*. Budapest, Hungary, 57–60.
- Schmidt, Thomas / Burghardt, Manuel** (2018): "An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing", in: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, 139–149.
- Schmidt, Thomas / Halbhuber, David** (2020): "Live Sentiment Annotation of Movies via Arduino and a Slider", in: *Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*. Late Breaking Poster.
- Schmidt, Thomas / Wolff, Christian** (2021): "Exploring Multimodal Sentiment Analysis in Plays: A Case Study for a Theater Recording of Emilia Galotti", in: *Proceedings of the Conference on Computational Humanities Research 2021 (CHR 2021)*. Amsterdam, the Netherlands.
- Schmidt, Thomas / Burghardt, Manuel / Wolff, Christian** (2019): "Towards Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti", in: *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*. Copenhagen, Denmark, 405–414.
- Schmidt, Thomas / Engl, Isabella / Halbhuber, David / Wolff, Christian** (2020a): "Comparing Live Sentiment Annotation of Movies via Arduino and a Slider with Textual Annotation of Subtitles", in: *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*, 212–223.
- Schmidt, Thomas / Engl, Isabella / Herzog, Juliane / Judisch, Lisa** (2020b): "Towards an Analysis of Gender in Video Game

Culture: Exploring Gender-specific Vocabulary in Video Game Magazines", in: *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*. Riga, Latvia.

Schmidt, Thomas / Schlindwein, Miriam / Lichtner, Katharina / Wolff, Christian (2020c): "Investigating the Relationship Between Emotion Recognition Software and Usability Metrics", in: *i-com*, 19(2): 139-151. DOI: 10.1515/icom-2020-0009

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021a): "Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language", in: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 67-79.

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021b): "Using Deep Neural Networks for Emotion Analysis of 18th and 19th century German Plays", in: *Fabrikation von Erkenntnis: Experimente in den Digital Humanities* (vDHD Sonderband). Melusina Press. DOI:10.26298/melusina.8f8w-y749-udlf

Schmidt, Thomas / El-Keilany, Alina / Eger, Johannes / Kurek, Sarah (2021c): "Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies", in: *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*. Krasnoyarsk, Russia.

Schmidt, Thomas / Grünler, Johanna / Schönwerth, Nicole / Wolff, Christian (2021d): "Towards the Analysis of Fan Fictions in German Language: Exploration of a Corpus from the Platform Archive of Our Own", in: *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*. Krasnoyarsk, Russia.

Vonderau, Patrick (2020): "Quantitative Werkzeuge", in: Hagener, Malte / Pantenburg, Volker (eds.): *Handbuch Filmanalyse*, Springer Fachmedien, 399-413. DOI: 10.1007/978-3-658-13339-9_28

Welke, Tina (2005): "Die Tatortfolge 'Quartet in Leipzig' als gesamtdeutscher Tatort: Analyse einer inszenierten deutsch-deutschen Annäherung", in: *Verl. für Gesprächsforschung*, Radolfzell.

Wu, Yuxin / Kirillov, Alexander / Massa, Francisco / Lo, Wan-Yem / Girshick, Ross (2019): *Detectron2* <<https://github.com/facebookresearch/detectron2>> [14.06.2021]

Wulff, Hans J. (1998): "Semiotik der Filmanalyse: Ein Beitrag zur Methodologie und Kritik filmischer Werkanalyse", in: *Kodikas/Code*, 21(1-2): 19-36.

Zaharieva, Maia / Breiteneder, Christian (2012): "Recurring Element Detection in Movies". in Schoeffmann, Klaus et al. (eds.): *Advances in Multimedia Modeling*, Springer, 222-232. DOI: 10.1007/978-3-642-27355-1_22

Zhang, Kaipeng / Zhang, Zhanpeng / Li, Zhifeng / Qiao, Yu (2016): "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks", in: *IEEE Signal Processing Letters*, 23(10): 1499-1503. DOI: 10.1109/LSP.2016.2603342

Zhou, Bolei, et al. (2017): "Places: A 10 million image database for scene recognition", in: *IEEE transactions on pattern analysis and machine intelligence* 40.6: 1452-1464.

Der SSH Open Marketplace Kontextualisiertes Praxiswissen für die Digital Humanities

Zarei, Alireza

alireza.zarei@gwdg.de
GWDG Göttingen

Seung-Bin, Yim

Seung-Bin.Yim@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Đurčo, Matej

Matej.Durco@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Illmayer, Klaus

Klaus.Illmayer@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Barbot, Laure

laure.barbot@dariah.eu
DARIAH-EU

Fischer, Frank

frank.fischer@dariah.eu
DARIAH-EU; Higher School of Economics, Moskau

Gray, Edward

edward.gray@dariah.eu
DARIAH-EU

1. Was ist der SSH Open Marketplace?

Die internationale Digital Humanities-Gemeinde als "community of practice" hat bereits früh damit begonnen, Kataloge mit forschungsrelevanten Tools aufzubauen, um damit einen essentiellen Teil ihrer Praxis zu kartografieren. Diese Tool-Directories "are frequently referred to as an important component of digital humanities infrastructure" (Dombrowski 2020).

Auf der Ebene einzelner Organisationen, die Übersichten ihrer eigenen und für ihre Stakeholder relevanten Tools anbieten, funktioniert das auch problemlos: DARIAH-DE etwa listet eigene Werkzeuge und Dienste auf (<https://de.dariah.eu/en/dienste-und-werkzeuge>), CLARIN-NL bietet einen Überblick über Werkzeuge der CLARIN-Infrastruktur (<http://www.clarin.nl/node/404>), DH Austria unterhält eine fokussierte, aber vielfältigere Liste (<https://dha.acdh.oeaw.ac.at/en/know-more>), es gibt eine französische Liste von Werkzeugen zur Korpusexploration (<http://explorati-ondecopus.corpusecrits.huma-num.fr/>) und die Special Interest Group (SIG) der ADHO zu Digital Literary Stylistics betreibt ei-

nen einfachen Google-Spreadsheet, zu dem alle und jede*r beitragen können (<https://dls.hypotheses.org/774>).

Bei Projekten, die darauf abzielen, die gesamte Tool-Landschaft zu kartieren, ist es mit Listen nicht mehr getan. Datenbanken kommen zum Einsatz, womit auch die Kosten und der Wartungsaufwand steigen. Als Beispiele dienen das kanadische TAPoR (Text Analysis Portal for Research), das DiRT Directory (Digital Research Tools, vgl. Dombrowski 2014) oder TERESA (Tools E-Registry for E-Social science, Arts and Humanities). Während es einen Konsens darüber gibt, dass diese Directories von Nutzen sind, sind fast alle diese Projekte an einem Punkt gescheitert: am fehlenden Nachhaltigkeitskonzept, das sicherstellen würde, dass eine Plattform auch nach dem Auslaufen eines finanzierten Projekts weiter existieren kann und dafür Ressourcen bereitstehen (vgl. Barbot et al. 2020). Diese Diskrepanz zwischen mehr oder weniger erwiesener Nützlichkeit und langfristiger Unwartbarkeit hat Quinn Dombrowski das "Directory Paradox" genannt (Dombrowski 2021).

Im Rahmen des Horizon-2020-geförderten Projekts "Social Sciences & Humanities Open Cloud" (SSHOC), das eine Laufzeit von Januar 2019 bis April 2022 hat, wird der SSH Open Marketplace entwickelt, der einen Überblick nicht nur über digitale Tools und Services bietet, sondern auch über Trainingsmaterialien, Publikationen, Datensets und Workflows. Dabei werden diese untereinander kontextualisiert: Der Eintrag zu einem Tool verlinkt etwa Forschungspaper, die mithilfe dieses Tools entstanden sind, desweiteren passende Trainingseinheiten, zum Beispiel aus dem "Programming Historian", und, falls vorhanden, Forschungsdaten in einem für das Tool geeigneten Format. Die "Werkbänke der Digital Humanities" (Fischer et al. 2021) erscheinen auf diese Weise breit kontextualisiert. Das flexible Datenmodell und das Exponieren einer offenen API ermöglichen es, das gesammelte Praxiswissen für die Digital Humanities nutz- und erforschbar zu machen und eine neue, sich aktiv weiterentwickelnde Datenbasis dafür zu schaffen.

Der SSH Open Marketplace ist unter <https://marketplace.sshopencloud.eu/> seit Januar 2022 als stabile Version verfügbar. Um den Inhalt der Plattform aktuell zu halten, gibt es ein Kurationskonzept, das die Community mit einschließt, sowie ein eigenes Arbeitspaket, welches eine nachhaltige Governance-Struktur für dieses Projekt erarbeitet, in deren Mittelpunkt die europäischen Forschungsinfrastrukturen DARIAH, CLARIN und CESSDA stehen. Der Marketplace soll nicht nur als praktische Hilfe im Forschungsalltag dienen, sondern auch helfen die Frage zu beantworten, welche Rolle Tools in der DH-Community eigentlich spielen, im Sinne einer "Tool Science" (vgl. Wolff 2015). Auch Lücken in der Softwareversorgung sollen so sichtbar werden. Alle Daten sind frei nachnutzbar, der Code steht unter einer Open-Source-Lizenz.

2. Das Datenmodell

Dem SSH Open Marketplace liegt ein umfassendes, aber pragmatisches Datenmodell zugrunde, das auf generische Konzepte baut (vgl. Barbot et al. 2019b). Zu den grundlegenden Eigenschaften des Modells gehören (siehe auch Abb. 1):

- die fünf genannten Hauptentitäten: Tools & Services, Trainings Materials, Publications, Datasets, Workflows (Abfolgen von Arbeitsschritten)
- flexibel typisierte Relationen zwischen den Einträgen (zur Kontextualisierung)

- die detaillierte Versionierung aller Änderungen (Einträge werden automatisiert auf Konsistenz geprüft, können aber auch händisch angelegt und kuratiert werden)
- Actors (etwa Autor*innen und/oder Programmierer*innen) werden als eigene Entitäten modelliert und mit eigenen Identifiern versehen (etwa ORCID)

Eine der zentralen dynamischen Eigenschaften des Datenmodells ist "activity". Diese Eigenschaft klassifiziert die Einträge danach, in welcher Aktivität im Rahmen des Forschungsdaten-Lebenszyklus sie relevant sind. Die erlaubten Werte entsprechen dabei den Konzepten der TaDiRAH-Taxonomie (<https://vocab.s.dariah.eu/tadirah/>), wie z.B. "Scanning" oder "POS-Tagging". Weitere Beispiele für dynamische Properties sind den bibliografischen Angaben entlehnte Attribute für Publications ("conference", "journal", "year").

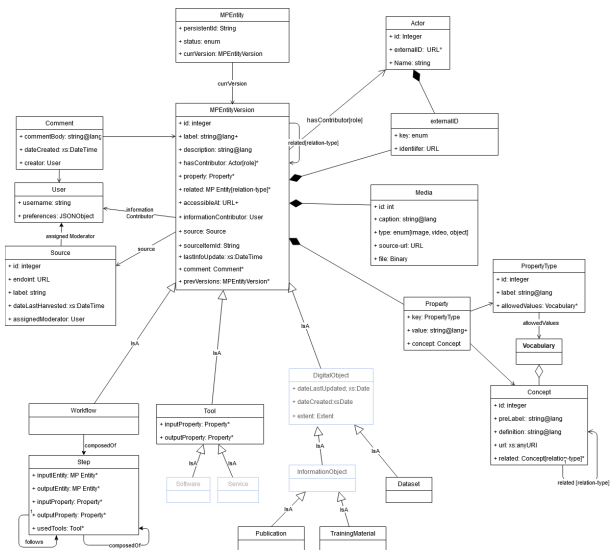


Abb. 1: Das Datenmodell des SSH Open Marketplace.

Die im Marketplace vorhandenen Daten werden über die API bereitgestellt, die online mithilfe von Swagger dokumentiert ist. Bis zum Projektende sollen die Daten auch im Sinne des LOD-Paradigmas im RDF-Format ausgeliefert werden. Als Target-Ontologie kommen die SSHOC Reference Ontology, Wikidata und Schema.org infrage, Mehrfachmappings sind denkbar und sinnvoll. Die Daten werden auch über einen SPARQL-Endpunkt abfragbar sein.

3. Überblick über die Inhalte des SSH Open Marketplace

Bisher (November 2021) haben über 5.000 Einträge ihren Weg in die Datenbank gefunden, die sich wie folgt auf die fünf Kategorien aufteilen:

- Tools & Services: 1671
- Training Materials: 321
- Publications: 2993
- Datasets: 305
- Workflows: 30

Anders als die Vorgängerprojekte, die oft von vorn angefangen haben, setzt der Marketplace darauf, bereits vorhandene Daten wiederzuverwenden, zu aktualisieren und anzureichern. Die häufigsten Quellen für die bezogenen Daten sind bisher:

- dblp computer science bibliography: 2837 Publikationen
- TAPoR: 1373 Tools
- SSK (Standardization Survival Kit): 29 Workflows und 370 Arbeitsschritte, letztere über die entsprechende Zotero-Bibliothek des SSK
- Humanities Data: 290 Datensets
- The Programming Historian: 83 Trainingseinheiten
- CLARIN Language Resource Switchboard: 56 Tools
- EOSC Marketplace: 15 Services
- SSHOC Service Catalogue: 13 Services

Dank unserer flexiblen Ingestion-Pipeline können zukünftig weitere Quellen eingespeist werden, geplant sind DARIAH Campus, die CLARIN Resource Families, das SSH Training Discovery Toolkit und SSHOC Training Material, die CESSDA Training Resources, Methodi.ca sowie die Daten nicht mehr gepflegter Projekte wie TERESAH.

4. Extraktionstask

Wie im vorangegangenen Kapitel beschrieben, speist sich der Marketplace aus verschiedenen Katalogen, deren Inhalte auf das Datenmodell gemappt und in die Kurationspipeline geschoben werden. Daneben werden aus dem Volltext wissenschaftlicher Publikationen und Übungsmaterialien erwähnte Tools extrahiert, um diese besser kontextualisieren zu können (Tool → mentionedIn → Publication). Forschungspapiere geben oft explizit Aufschluss über die Verwendung spezifischer Tools, Methoden und Datensätze, bieten daher Erfahrungswerte aus dem Forschungsalltag.

Die Systemarchitektur beinhaltet eine eigene "extraction"-Komponente, die in Volltexten von Publikationen Tools und Services identifiziert und mit entsprechenden Einträgen im Marketplace zusammenbringt und entsprechende Relationen anlegt.

Den Anfang bildete ein Experiment: die exemplarische Extraktion von Tools, die in den Beiträgen der jährlichen ADHO-Konferenzen erwähnt werden. Dafür wurde ein eigenes Kommandozeilenwerkzeug namens ToolXtractor entwickelt, das auf dem Erkennen von Zeichenketten basiert, die einer Positivliste entnommen werden (vgl. Barbot et al. 2019a und Fischer/Moranville 2020). Basierend auf der Tool-Liste des TAPoR-Projekts wurden in den Proceedings der Konferenzjahre 2015–2019 insgesamt 1.498 Erwähnungen gezählt, die auf 238 individuelle Tools zurückgingen. Die 15 am häufigsten genannten Tools im gewählten Korpus waren Gephi, Omeka, stylo, MALLET, Excel, D3.js, NLTK, WordPress, Drupal, TextGrid, CollateX, GeoNames, TXM, Solr und die Voyant Tools.

Nach diesen ersten Einblicken in die Trends der Tool-Nutzung innerhalb der DH-Forschung haben wir unseren Ansatz erweitert, um auch bisher noch nicht katalogisierte Tools zu finden. Wir haben dafür einen Datensatz geistes- und sozialwissenschaftlicher Publikationen entsprechend annotiert und mittels Transfer-Learning ein eigenes NER-Modell (Named Entity Recognition) trainiert, das Ergebnisse mit hoher Präzision und hohem Recall liefert.

Innerhalb der Extraktionspipeline (Abb. 2) werden über die API des SSH Open Marketplaces zunächst die bereits erfassten Publikationen abgerufen. Das erwähnte NER-Modell wird dann auf je-den Satz der entsprechenden Volltexte angewendet und liefert eine

Liste möglicher Tools zurück. Es wird überprüft, ob es für diese Tools bereits Einträge im Marketplace gibt; in diesem Fall wird eine Relation zwischen dem Tool und der Publikation hinzugefügt. Alle anderen extrahierten potenziellen Tools, die noch keinen Eintrag im Marketplace haben, werden in die Kurationspipeline eingespeist, wo sie entsprechend der Richtlinien bearbeitet werden.

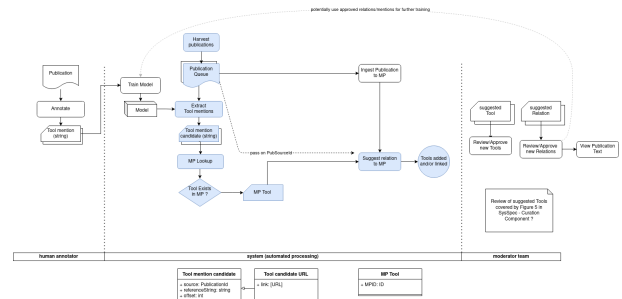


Abb. 2: Extraktionspipeline.

Eine Herausforderung bei der Datenextraktion stellen die verschiedenen Datenformate von DH-Publikationen dar. Je nach Konferenz bzw. Publikationsorgan und Jahr der Veröffentlichung findet sich ein bunter Mix aus PDF-, TEI- und HTML-Dateien. Teils sind Dateien aus bestimmten Konferenzjahren gar nicht mehr in zitierbarem Zustand verfügbar (etwa die Paper der DH2015 in Sydney), ein Missstand auch im Sinne des Konferenzmottos "Kulturen des digitalen Gedächtnisses". Ein weiteres Problem stellen Journale mit beschränktem Zugriff dar, etwa *Digital Scholarship in the Humanities* (DSH).

5. Ausblick

Der SSH Open Marketplace baut auf den Erfahrungen vieler Vorgängerprojekte auf. Durch Nutzerbefragungen und frühes Community Engagement haben wir versucht, eine inklusive Plattform zu schaffen. Die Daten sind aber auch maschinenlesbar und mit anderen Linked-Data-Projekten verknüpft, sodass der Marketplace Teil eines größeren Ökosystems ist, aus dem neue Impulse und Daten kommen. Es wird in Zukunft darauf ankommen, die Entwicklungen im Feld der Digital Humanities genau zu beobachten und zusätzliche relevante Quellen in die Ingestion-Pipeline aufzunehmen, die das Fach disziplinär weiter begreifen (vgl. dazu den Aufsatz von Luhmann/Burghardt 2021).

Der Marketplace kann und soll nicht andere Arten von Repositorien ersetzen, etwa OpenAIRE oder Dataverse, es ist keine Hosting-Plattform für Daten oder Forschungspaper. Der Fokus liegt auf der Kontextualisierung: Es werden Inhalte bevorzugt, die eine wertige Relation zwischen Tools & Services, Training Materials, Publications, Datasets und Workflows herstellen.

Im Idealfall kann der Marketplace dabei helfen, die Rolle dieser Hauptentitäten zu verdeutlichen und zu stärken. Die Bedeutung offen zugänglicher Datensets etwa, die den FAIR-Prinzipien genügen, ist innerhalb der Digital Humanities immer noch gering, was sich unter anderem darin zeigt, dass die einzige dedizierte Sammlung von DH-relevanten Datensets das Privatprojekt eines einzelnen Forschers ist (Humanities Data, <https://humanitiesdata.com/>).

Fördernachweis

Der SSH Open Marketplace wird vom Europäischen Forschungsrat (ERC) im Rahmen des Forschungs- und Innovationsprogramms Horizon 2020 (Fördervereinbarung Nr. 823782) entwickelt.

Bibliographie

Barbot, Laure / Fischer, Frank / Moranville, Yoann / Pozdniakov, Ivan (2019a): "Which DH Tools Are Actually Used in Research?" In: *weltliteratur.net*, 6. Dezember 2019. (URL: <https://weltliteratur.net/dh-tools-used-in-research/>)

Barbot, Laure / Moranville, Yoann / Fischer, Frank / Petitfils, Clara / Ďurčo, Matej / Illmayer, Klaus / Parkoła, Tomasz / Wieder, Philipp / Karampatakis, Sotiris (2019b): *SSHOC D7.1 System Specification – SSH Open Marketplace* (Version 1.0). Zenodo.

Barbot, Laure / Dombrowski, Quinn / Fischer, Frank / Rockwell, Geoffrey / Spiro, Lisa (2020): "Who Needs Tool Directories? A Forum on Sustaining Discovery Portals Large and Small." In: *DH2020: »carrefours/intersections«*, 22–24. Juli 2020. *Book of Abstracts*. University of Ottawa. (URL: https://dh2020.adho.org/wp-content/uploads/2020/07/126_WhoneedstooldirectoriesAforumonsustainingdiscoveryportalslargeandsmall.html)

Dombrowski, Quinn (2014): "What Ever Happened to Project Bamboo?" In: *Literary and Linguistic Computing*, Vol. 29, Issue 3, September 2014, S. 326–339, doi:10.1093/lc/fqu026.

Dombrowski, Quinn (2021): "The Directory Paradox." In: Anne McGrail et al. (Hg.): *Debates in the Digital Humanities: Institutions, Infrastructures at the Interstices*. University of Minnesota Press (erscheint demnächst).

Fischer, Frank / Moranville, Yoann (2020): "DH Tools Mentioned in 'The Programming Historian'." In: *weltliteratur.net*, 17. Januar 2020. (URL: <https://weltliteratur.net/dh-tools-programming-historian/>)

Fischer, Frank / Burghardt, Manuel / Luhmann, Jan / Barbot, Laure / Moranville, Yoann / Zarei, Alireza (2021): "Die Werkbänke der Digital Humanities: Zur Rolle von Tools und Software für die Forschungsarbeit." In: *vDHd2021: "Experimente"*, Zenodo, doi:10.5281/zenodo.4639228.

Luhmann, Jan / Burghardt, Manuel (2021): "Digital humanities – A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape." In: *Journal of the Association for Information Science and Technology*, 1–24, doi:10.1002/asi.24533.

Wolff, Christian (2015): "The Case for Teaching 'Tool Science'. Taking Software Engineering and Software Engineering Education beyond the Confinements of Traditional Software Development Contexts." In: *2015 IEEE Global Engineering Education Conference (EDUCON)*, Tallinn, pp. 932–938, doi:10.1109/EDUCON.2015.7096085.

Die Aktualität des Unzeitgemäßen

Krewet, Michael

m.krewet@fu-berlin.de
Freie Universität Berlin, Germany

Ernst, Felix

felix.ernst@kit.edu
Karlsruher Institut für Technologie, Germany

Götzelmann, Germaine

germaine.goetzelmann@kit.edu
Karlsruher Institut für Technologie, Germany

Hegel, Philipp

philipp.hegel@tu-darmstadt.de
Technische Universität Darmstadt, Germany

Schenk, Torsten

torsten.schenk@tu-darmstadt.de
Technische Universität Darmstadt, Germany

Söring, Sibylle

sibylle.soering@fu-berlin.de
Freie Universität Berlin, Germany

Tonne, Danah

danah.tonne@kit.edu
Karlsruher Institut für Technologie, Germany

Der Vortrag möchte die neuen Forschungspraktiken und -möglichkeiten, die sich mit der Digitalisierung jenes Bereichs des kulturellen Gedächtnisses, den das Berliner Aristotelesarchiv bewahrt, in den Fokus rücken. Nicht unerwähnt bleiben sollen dabei die mit der Digitalisierung verbundenen Herausforderungen mit Lizenzrechten und der institutionellen Kuratierung. Dabei soll sich zeigen, dass auch ein Rückgriff auf ein nicht mehr zeitgemäß scheinendes Medium wie den Mikrofilm Optionen bereithält, digitale Verfahren anzuschließen.

Das Aristotelesarchiv als Gedächtnisinstitution

Das Aristotelesarchiv an der Freien Universität Berlin ist ein weltweit einzigartiger Ort. Es besitzt Mikrofilme von allen bekannten erhaltenen Aristoteleshandschriften in aller Welt. Hinzu kommen in geringerem Umfang Farbdigitalisate von Handschriften. Zu der Sammlung des Archivs gehören Material aus großen Bibliotheken mit einem bedeutenden Bestand, wie z. B. der Biblioteca Vaticana oder der Bibliothèque nationale de France, aber auch aus nicht mehr öffentlich zugänglichen Klosterbibliotheken des Bergs Athos, oder auch aus anderen, oft nur schwer

zugänglichen Klosterbibliotheken (z. B. in Ägypten). Einige der Handschriften, von denen sich im Archiv noch Bilder befinden, gelten mittlerweile als verschollen. Hinzu kommen Fälle, in denen die Qualität der Schrift der teilweise Jahrzehnte alten Mikrofilmaufnahmen die heute in den entsprechenden Originalhandschriften vorfindbare Schrift deutlich übersteigt. Digitalisate der Handschriften, die, sofern überhaupt möglich, mit heutigen Mitteln erstellt werden, hätten also diesbezüglich Nachteile gegenüber den älteren Mikrofilmen.

Die Aristoteleshandschriften selbst überliefern nicht nur den Text. Eine Vielzahl der Handschriften wurde v. a. zwischen dem 9. und 16. Jahrhundert durch Paratexte (erklärende Glossen, Scholia, Diagramme und Randkommentare) ergänzt. Somit repräsentieren die Aristoteleshandschriften nicht nur die Aristoteles Texte, sondern auch die Auslegungstraditionen und die Kontexte, in denen der handschriftliche Text über Jahrhunderte behandelt, studiert und gelehrt wurde. Insofern sind sie bis heute Gegenstand einer Vielzahl von philologischen, philosophischen und historischen Forschungen.

Über das Sammeln aller Handschriften in Form von Mikrofilmen und Farbdigitalisaten hinaus haben Mitarbeitende des Archivs über Jahrzehnte zu jeder einzelnen dieser Handschriften umfangreiche Handschriftenbeschreibungen verfasst. Diese erfolgten in Autopsie. Die Reisen der Mitarbeitenden in die Bibliotheken wurden weitestgehend aus Drittmitteln finanziert. Nur ein erster Teil der Beschreibungen wurde in Buchform publiziert (Moraux et al. 1976). Die weiteren Beschreibungen finden sich noch unpubliziert in Form von Schreibmaschinentexten im Archiv. Diese Beschreibungen bilden gegenwärtig auch die Grundlage für die Transformation in digitale Metadaten. Schließlich besitzt das Archiv auch eine Sammlung von Sekundärliteratur zu den einzelnen Handschriften und zur Paläographie und Kodikologie.

Das Aristotelesarchiv ist damit der Ort, an dem die gesamte griechischsprachige Überlieferung eines der bedeutendsten Philosophen des Abend- und Morgenlandes zusammenkommt. Aufgrund dieses weltweiten Alleinstellungsmerkmals ist das Archiv seit Jahrzehnten der Magnet für Forschende aus aller Welt zum kulturell-materiellen und gedanklichen Erbe des Aristoteles.

Das Archiv als Ort digitaler Bewahrung

Erfreulicherweise digitalisieren große Bibliotheken (z. B. Biblioteca Vaticana, Biblioteca Medicea Laurenziana) seit einigen Jahren ihren Handschriftenbestand und stellen die Digitalisate in Farbe online. Als Fazit der Digitalisierungsbewegung der weltweit auf die Bibliotheken zerstreuten Aristoteleshandschriften bleibt aber auch festzuhalten, dass sich ein sehr großer Teil der Handschriften in kleinen Bibliotheken befindet, die noch nicht mit der Digitalisierung der Handschriften begonnen haben. Andere Bibliotheken mit einem größeren Bestand von Aristoteleshandschriften (z. B. die Biblioteca Ambrosiana in Mailand) verbinden weiterhin ein Kostenmodell mit der Bestellung von Handschrift-Digitalisaten, so dass ein freier Onlinezugang der Handschriften in näherer Zukunft als unsicher gelten muss. Es wäre vermessen, eine umfangreiche oder gar eine umfassende, frei zugängliche Sammlung von digitalisierten Aristoteleshandschriften in den nächsten Jahrzehnten zu erwarten. Damit bietet die Sammlung des Berliner Aristotelesarchivs auch perspektivisch einzigartige Möglichkeiten für um digitale Methoden erweiterte Forschungen zu dem kulturell-handschriftlichen Erbe des Autors.

Das Aristotelesarchiv mag somit mit seinem Bestand nur auf den ersten Blick antiquiert scheinen. Es besitzt den Vorteil, dass es mit dem Erwerb der Mikrofilme auch Lizenzen für die Arbeit an diesen erworben hat, die den Aufbau einer digitalen Forschungsinfrastruktur erlauben. Die einzige Einschränkung bleibt, dass für viele Handschriften die Digitalisate nicht online gestellt werden dürfen. Heute finden sich Forscher*innen vor der Situation, dass einige Bibliotheken im Ausland keine Digitalisierungen ihrer Aristoteleshandschriften mehr zur Verfügung stellen. Hinzu kommt, dass viele kleine Bibliotheken mit einem kleineren Handschriftenbestand ihre Aristoteleshandschriften mit großer Wahrscheinlichkeit angesichts einer mangelnden Infrastruktur erst einmal nicht digitalisieren oder online stellen werden. Aus diesem Grund spielt das Aristotelesarchiv für innovative und digital gestützte Forschungsansätze zu dem kulturellen Erbe der mehr als 1000 bekannten Aristoteleshandschriften eine unverzichtbare Rolle. In diesem Vortrag wird der Aufbau einer solchen Forschungsinfrastruktur für das Aristotelesarchiv vorgestellt, wie sie bislang im Rahmen des von der DFG geförderten Sonderforschungsbereichs 980 „Episteme in Bewegung. Wissenstransfer von der Alten Welt bis in die Frühe Neuzeit“ erfolgte.

Diese Forschungsinfrastruktur dient dabei nicht alleine der langfristigen Bewahrung von so genannten digitalen Objekten. Vielmehr stellt sie den Grundbaustein für die aktive Arbeit mit ihnen dar - das heißt mit Hilfe vielfältiger Werkzeuge für die Arbeit mit und an Handschriftendigitalisaten werden in den meisten Fällen erst (Forschungs-)Daten generiert, die ebenfalls bewahrt werden müssen. Durch die Ergänzung oder Veränderung vorhandener Daten und die Hinzufügung vielfältiger Verknüpfungen wird das Digitalisat Teil eines komplexen Objektes und einer modularen Infrastruktur, die für die weitere Forschung zur Verfügung gestellt werden.

Als Grundlage und Umsetzung digitaler Forschung an Archivalien werden infrastrukturelle Komponenten, die Zugriff und Durchsuchbarkeit gewährleisten, mit forschungsnahen Werkzeugen verzahnt, die wissenschaftlichen und technischen Austausch sowie Nutzbarkeit der entstehenden Forschungsdaten ermöglichen. Gemeinsam bilden Infrastruktur und Werkzeuge so die Grundlage für die aktuell vielfach diskutierten *FAIR Principles*: Findable, Accessible, Interoperable, Reusable (Wilkinson et al. 2016). Unser Beitrag soll zeigen, dass diese Prinzipien gerade dann als gelebte Praxis wichtig sind, wenn die Forschungsobjekte unzugänglich oder von Verfall bedroht sind. Und dass gerade kleine Fächer und Spezialarchive Vorreiter und Profiteure gleichermaßen sein können, wenn sie ihre digitalen Objekte und Forschungsdaten auf diese Weise zugänglich machen. Im Folgenden werden wir die verschiedenen Ausprägungen der einzelnen FAIR-Prinzipien und ihre Anwendung am Beispiel des Aristotelesarchivs kurz skizzieren.

Das Archiv als Ort digitaler Forschung

Das Herzstück der hier beschriebenen digitalen Infrastruktur ist das Forschungsdatenrepositorium, in welchem die Daten und zugehörige Metadaten als strukturierte Objekte verwaltet und nachhaltig gespeichert werden. Die reichhaltigen (Meta)daten besitzen eindeutige Bezeichner, welche bei der Nachnutzung eine unmissverständliche Referenzierung der Digitalisate erlauben und ein zielführendes Auffinden der gesuchten Informationen unterstützen. Zugänglichkeit der Daten und deren Metadaten wird durch einen Abfragedienst gesichert, der offene, freie, universelle und

standardisierte Protokolle nutzt und - wo lizenzrechtlich nötig - eine adäquate Zugriffskontrolle realisiert.

Ein solches Repositorium ermöglicht nun eine vielfältige Forschung zu den Aristoteleshandschriften, weil erstmals Digitalisierungen von allen Handschriften (schwarz-weiße Scans von Mikrofilmen und Farbdigitalisate) digital an einem Ort zusammengeführt werden können. Im Falle eines Projekts im Rahmen des genannten Sonderforschungsbereichs erfolgte zunächst eine Digitalisierung der Handschriften, welche die logischen Schriften des Aristoteles überliefern. In einer Kooperation von Forschenden der Informatik, Computerphilologie und gräzistischer Fachwissenschaft wurden Werkzeuge mit Bottom-Up-Ansatz spezifisch für Forschungsfragen entwickelt oder weiterentwickelt, die Fachwissenschaftler*innen eine Reihe von erweiterten Möglichkeiten und neue methodische Zugänge für Forschungen zur Kodikologie, Paläographie, der Überlieferungsgeschichte des Aristoteles oder auch des Wissens- oder Texttransfers ermöglichen.

Die Handschriftendigitalisate lassen sich mit Hilfe von digitalen Annotationswerkzeugen sowohl mit automatisch erzeugten als auch mit fachwissenschaftlichen Informationen gemäß W3C-Empfehlung „Web Annotation Data Model“ (Young / Ciccarese / Sanderson 2017; zur Umsetzung Tonne et al. 2019) anreichern. So kann beispielsweise ein standardisiertes kodikologisches Vokabular verlinkt werden. Dies eröffnet die Möglichkeiten, Paratexte (Interlinear- und Marginalglossen, Scholia, Randkommentare oder mit dem Text alternierende Kommentare, Diagramme) zu transkribieren, zu beschreiben, zu übersetzen oder auch mit Stichwörtern zu taggen.

Des Weiteren kann mit einer Layoutanalyse (Chandna et al. 2015) die Seitengröße und der Text-, Rand- oder Scholiabereich automatisch und semiautomatisch (d. h. über einfache korrigierende Nachzeichnungen z. B. des Textbereiches, wenn es wegen des Ineinandergreifens von Text und Scholia zu Ungenauigkeiten gekommen ist) gemessen werden. In Verbindung mit den Metadaten zur Handschrift (v. a. zur Datierung und Provenienz) stellt dieses Werkzeug somit eine innovative Hilfe für kodikologische Forschungen dar: Einzelne Handschriften können jetzt auch jenseits alleiniger paläographischer Hypothesen mit Hilfe der Ergebnisse der Layoutanalyse einer bestimmten historischen Schreibschule zugewiesen werden. Insofern durch die Analyse auf diese Weise eine Provenienz erschlossen werden konnte, werden raumzeitliche Transfers der jeweiligen Handschrift und damit auch eine historische, raumzeitliche Dissemination des Wissens von der aristotelischen Logik, das der jeweilige Codex beinhaltet, nachvollziehbar.

Über Suchmöglichkeiten in den Annotationen, auch in Zusammenhang mit den beschreibenden Metadaten sowie den Repositoriumsdaten werden eine Reihe von Text- und Wissenstransfers von Handschrift zu Handschrift nachverfolgbar. Kontakte von Handschriften untereinander konnten auch in Fällen nachgewiesen werden, in denen die Ergebnisse einer traditionellen textkritischen Analyse (Textkollation) auf keine Verwandtschaft schließen lassen würden, wenn die Texte von Handschriften beispielsweise so genannte signifikante Fehler im Text nicht teilen, gleichwohl aber über die Anwendung der digitalen Werkzeuge z. B. signifikante Fehler oder Gemeinsamkeiten der Handschriften in Paratexten gefunden werden können. Ein in dieser Weise detektierter späterer Kontakt von Handschriften miteinander kann zum einen ebenfalls Aufschluss über den raumzeitlichen Transfer mindestens einer dieser Handschriften geben (z. B. von dem Ort, an dem sie kopiert wurde, an einen Ort, an dem ihre Glossen abgeschrieben wurden oder die Glossen einer anderen Handschrift in sie eingetragen wurden). Zum anderen können durch so nachweisbare Kontakte von Handschriften untereinander Wissenstransfers

und Veränderungen innerhalb von Wissensbeständen – also beispielsweise in der handschriftlichen Überlieferung eines einzelnen Werkes des Aristoteles – nachgewiesen werden.

Das Aristotelesarchiv ermöglicht den Fachforschenden und Interessierten durch einen Zugangsaccount, der sich mit Nutzungsrechten in Übereinstimmung mit den Lizenzen des Archivs verbindet, den Zugang zu der Forschungsinfrastruktur. Auf die vom Projekt mittels dieser Forschungsinfrastruktur erzielten Ergebnisse und Forschungsdaten (z.B. Transkriptionen von Scholia, Glossen; die statistische Auswertungen, die Durchsuchbarkeit der Handschriften nach Glossen, Diagrammen; Visualisierungen und Ergebnisse von Layoutanalysen usw.) kann offen zugegriffen werden (entsprechend der CC-BY-Lizenz). Softwarekomponenten werden entweder gemäß DFG-Richtlinien als open source veröffentlicht oder sind bereits Nachnutzungen von öffentlich verfügbaren Bibliotheken und Frameworks. Trotz der lizenzrechtlichen Einschränkungen argumentieren wir, dass ein solches Archiv FAIR sein kann (Wilkinson et al. 2016, A1.2; Higman et al. 2019) und als einzelner ‘Mosaikstein’ in einer vielfältigen Forschungslandschaft bereichernd wirkt.

Analogen Archiv – digitale Gedächtnisinstitution

Die gemeinsame Forschung und Entwicklung hat maßgeblich dazu beigetragen, disziplinspezifischen Fragestellungen und dem Arbeitsalltag der Forschenden ebenso Rechnung zu tragen wie der Implementierung fachübergreifender Datenmodelle und Standards. So kann sichergestellt werden, dass einerseits der Aufbau einer Nutzergruppe zukünftig dadurch erleichtert wird, dass diese ihre Anforderungen adressiert sieht und ein Projekt mit Beispielcharakter und Vorbildfunktion bereits digital umgesetzt ist, um die Leistungsfähigkeit der Komponenten unter Beweis zu stellen. Durch die Einbindung der Infrastruktur mit ihren Werkzeugen in die tägliche Arbeit der Forschenden gehen Rückmeldungen weit über reines Nutzerfeedback hinaus. Die Bearbeitung von Forschungsfragen und die Weiterentwicklung der Werkzeuge sind eng verzahnt und befruchten einander in einem iterativen Prozess. Andererseits wurde aber von Beginn durch Nutzung verbreiteter Austauschformate (TEI) und die Umsetzung standardisierter Schnittstellen und Protokolle (Web Annotation Protocol) (Sanderson 2017) gleichermaßen zukünftige menschliche und maschinelle Nachnutzung gestärkt. Im Bereich der digitalen Geisteswissenschaften muss dabei die Nachnutzung durch auch ‘traditionell’ arbeitende Forschende naturgemäß Vorrang haben vor der Möglichkeit zu weiterer automatisierter Nutzung und Datenanalyse. Reichhaltige beschreibende Metadaten, dokumentierte Strukturierung und Frontend-Komponenten für Suche und Visualisierung sind dabei von zentraler Bedeutung. Es ist jedoch dringend angeraten, auch die andere Seite der Nutzbarkeit von vornherein mit anzulegen. Persistente Verknüpfungen, standardisierte Programmierschnittstellen und offene Softwarelizenzen leisten hier wichtige Aufgaben.

Im Falle der Erforschung hoch komplexer Textüberlieferungen – wie im Beispielprojekt der Texte des Aristotelischen *Organon* – für deren Erforschung ein Forscherleben nach Meinung der Experten kaum ausreicht (Reinsch 2001), können neben Ergebnissen auch Forschungsdaten nun über das Ende eines Projekts oder Forscherlebens hinaus für zukünftige Projekte oder Forschende gesichert und offen zur Verfügung gestellt werden. Während die meisten Digitalisate der Handschriften wegen lizenzrechtlicher Hindernisse nicht online gestellt werden können,

können die Forschungsdaten hingegen offen und in nachnutzbaren Formaten langfristig zugänglich gemacht werden. Ebenso bietet die Infrastruktur die Möglichkeit, neue Formen synergetischer Forschungen zu initiieren, indem beispielsweise Spezialistinnen und Spezialisten von den unterschiedlichsten Orten sich gemeinsam komplexen Forschungsfragen an dem digitalen Gedächtnis des Archivs zuwenden können. So entsteht neues Potential, diese individuell unlösbaren Forschungsfragen kollaborativ zu bewältigen.

Der Vortrag möchte Einblicke geben, wie man ein weltweit bedeutendes Archiv zukunftsgerichtet für digital gestützte Forschungen aufstellen und damit eine einzigartige digitale Forschungslandschaft konstituieren kann. Neben den eher technischen Anstrengungen, die Forschungsdaten auffindbar und zugreifbar zu gestalten, sind zu Beginn vor allem die inhaltlichen Arbeiten zur Sicherstellung von Interoperabilität und Nachnutzbarkeit für kleinere Projekte von zentraler Bedeutung. Auch wenn die notwendigen Anstrengungen für eine FAIRe Forschungslandschaft immens und vielleicht teilweise unüberwindbar scheinen, ist es keine Lösung, diese Aufgabe ausschließlich auf große Verbünde und Infrastrukturen zu übertragen. Ganz im Gegenteil sind Spezialarchive und allgemein die kleinen Fächer unverzichtbare Bausteine, die unabhängig von der Größe einen einzigartigen Beitrag für die zukünftige Forschung leisten. Im konkreten Fall bedeutet dies: mit Hilfe des Archivs als Gedächtnisinstitution und seiner digitalen Infrastruktur ist es über die Mikrofilme erstmals möglich, mit digitalen Methoden auf die gesamte handschriftliche Überlieferung von Aristoteles' logischen Schriften zuzugreifen und ihre historische und interdisziplinäre Bedeutung aufzudecken.

Bibliographie

Chandna, Swati / Tonne, Danah / Jejkal, Thomas / Stotzka, Rainer / Krause, Celia / Vanscheidt, Philipp / Busch, Hannah / Prabhune, Ajinkya (2015): "Software Workflow for the Automatic Tagging of Medieval Manuscript Images (SWATI)", in: *Document Recognition and Retrieval XXII. International Society for Optics and Photonics* : 940206.

Higman, Rosie / Bangert, Daniel / Jones, Sarah (2019): *Three Camps, One Destination: The Intersections of Research Data Management, FAIR and Open* <https://insights.uksg.org/articles/10.1629/uksg.468/> [letzter Zugriff 26. November 2021]

Morau, Paul / Harlfinger, Dieter / Reinsch, Diether / Wiesner, Jürgen (1976): *Aristoteles Graecus. Die griechischen Manuskripte des Aristoteles. Bd. 1: Alexandrien – London* (= Peripatoi 8). Berlin / New York: Walter de Gruyter 1976.

Reinsch, Diether Roderich (2001): "Fragmente einer Organon-Handschrift vom Beginn des zehnten Jahrhunderts aus dem Katharinenkloster auf dem Berge Sinai", in: *Philologus* 145: 57-69.

Sanderson, Robert (2017): *Web Annotation Protocol. W3C Recommendation* <https://www.w3.org/TR/2017/REC-annotation-protocol-20170223/> [letzter Zugriff 8. Juli 2021].

Tonne, Danah / Götzmann, Germaine / Hegel, Philipp / Krewet, Michael / Hübner, Julia / Söring, Sibylle / Löffler, Andreas / Hitzker, Michael / Höfler, Markus / Schmidt, Timo (2019): "Ein Web Annotation Protocol Server zur Untersuchung vormoderner Wissensbestände", in: *Konferenzabstracts DHd 2019 Digital Humanities: multimedial & multimodal* : 283-285 <http://doi.org/10.5281/zenodo.2596095> [letzter Zugriff 14. Juli 2021].

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand Jan / Appleton, Gabrielle / Axton, Myles / Baak, Arie / Blomberg, Niklas Blomberg / Boiten, Jan-Willem / Bonino da Silva Santos, Luiz / Bourne, Philip E. / Bouwman, Jildau / Brookes, Anthony J. / Clark, Tim / Crosas Mercè / Dillo, Ingrid / Dumon, Olivier / Edmunds, Scott / Evelo, Chris T. / Finkler, Richard / Gonzales-Beltran, Alejandra / Gray, Alasdair J.G. / Groth, Paul / Goble, Carole / Grethe, Jeffrey S. / Heringa, Jaap / Holt Hoen, Peter A. C. / Hooft, Rob / Kuhn, Tobias / Kok, Ruben / Kok, Joost / Lusher, Scott, J. / Martone, Maryann E. / Mons, Albert / Packer, Abel L. / Persson, Bengt / Rocca-Serra, Philippe / Roos, Marco / van Schaik, Rene / Sansone, Susanna-Assunta / Schultes, Erik / Sengstag, Thierry / Slater, Ted / Strawn, George / Swertz, Morris A. / Thompson, Mark / van der Lei, Johan / van Mulligen, Erik / Velterop, Jan / Waagmeester, Andra / Wittenburg, Peter / Wolstencroft, Katherine / Zhao, Jun / Mons, Barend (2016): "The FAIR Guiding Principles for Scientific Data Management and Stewardship", in: *Sci Data* 3, 160018 <https://doi.org/10.1038/sdata.2016.18> [letzter Zugriff 8. Juli 2021].

Young, Benjamin / Ciccarese, Paolo / Sanderson, Robert (2017): *Web Annotation Data Model. W3C Recommendation* <https://www.w3.org/TR/2017/REC-annotation-model-20170223/> [letzter Zugriff 8. Juli 2021].

Digitale Kontextualisierung und Visualisierung der Quellen-Trias Bild-Text-Realia zu historischer Kleidung, ihrer Ausformung, Zeichenhaftigkeit und Dreidimensionalität

de Günther, Sabine

sabine.de.guenther@fh-potsdam.de
Fachhochschule Potsdam, Germany

Freyberg, Linda

linda.freyberg@fh-potsdam.de
Fachhochschule Potsdam, Germany

Kleidung als Untersuchungsgegenstand

Kleidung, Tracht und Mode kommunizieren die Vorstellung des Trägers von Schönheit, Status, Alter, Geschlecht, Körper, Form, sozialen Hierarchien und religiösen Unterschieden, kurz: die Identität des Trägers. Sie kennzeichnet den Träger auch als Teil einer gesellschaftlichen Schicht (Simmel 1905: 8-9). Das Aussenden von Zeichen und Botschaften wirkt als Orientierungsmechanis-

mus in sozialen Interaktionen. Diese geben sowohl Formen als auch Normen vor, die durch Ökonomie, technischen Fortschritt und Sittengesetze reguliert werden. Die Kommunikation durch Kleidung auf der visuellen Ebene, dem "viscourse" (Cetina 1999, 245-263), ist gekoppelt mit der kulturellen Bedeutung von Zeichen und Symbolen, einer kodierten Sprache, die der Epoche und ihrem kulturellen Kontext immanent ist. Darüber hinaus wirkt der stets räumliche Aspekt der Kleidung als visueller Code. In ihrer Dreidimensionalität fungiert die Kleidung als "meaningful marker of space" (Höpflinger 2014: 177).

Kontextualisierung als Methode

Überlieferte Quellen zur Kleidung, Tracht und Mode finden sich in bildlichen Darstellungen, Schriftquellen, wie beispielsweise Luxusgesetzen, Nachlassinventaren oder Spottschriften, und überlieferten textilen Artefakten, wie beispielsweise Grabfunden. Jeder Quellentypus liefert Einzelinformationen, die interpretiert und in Verbindung mit weiteren Quellen quergelesen werden müssen, um eine Aussage über die Zeichenhaftigkeit von Kleidung treffen zu können (de Günther/Zitzlsperger 2018: 1-6).

Die Darstellung von Kleidung in Porträts, Genredarstellungen, Modegrafik oder Karikaturen liefert eine Reihe von Informationen über den oder die TrägerIn, über Identifikationsmuster, den kulturellen Kontext oder die Intention des Malers oder der Malerin. Oft genug wird jedoch hybride oder unauthentische Kleidung abgebildet und verwässert so die Sprache der Kleidung im Bild (Zitzlsperger 2015). Darüber hinaus erschwert die fehlende Mehransichtigkeit der dargestellten Kleidung die Identifizierung der dargestellten Person, ihrer Gewandung und die zeitliche oder geographische Einordnung. Eine Interpretation und Kontextualisierung von Gewand- und Schmuckelementen im Bild ist notwendig, das Hinterfragen und Querlesen des Bildes als Informationsquelle ratsam.

Während visuelle und textuelle Quellen eine Reihe von Informationen liefern, sind es die materiellen Objekte selbst, die für das Verständnis der Kleidergeschichte entscheidend sind. Textile Artefakte geben Auskunft über Schnitt, Konstruktion, Farbe, Volumen, Änderungen und textile Fertigungstechniken - kurz: über den materiellen Aspekt des Kleidungsstücks. Jene Materialität gibt Auskunft über den Wohlstand des Trägers, über Schnitt und Konstruktion, über textile Schichten, Verarbeitungsweisen und rückseitige Ausformungen. Diese Details liefern Informationen für disziplinspezifische Fragestellungen; so etwa für Forschende der Material Cultures, als auch für Forscher*innen aus den Bereichen der visuellen Studien, der Sprach- und der Kulturwissenschaften.

In dem interdisziplinären Forschungs- und Digitalisierungsprojekt „Restaging Fashion. Digitale Kontextualisierung vestimentärer Quellen“, angesiedelt am UCLAB der Fachhochschule Potsdam¹ in Kooperation mit den Staatlichen Museen zu Berlin - Preußischer Kulturbesitz und dem Germanischen Nationalmuseum Nürnberg wird die Quellen-Trias Bild-Text-Realie mit unterschiedlichen bildgebenden Verfahren digitalisiert, die Daten werden modelliert, semantisch angereichert, annotierbar gemacht und in unterschiedlichen Granularitäten visualisiert. Die digitale Verzahnung verschiedener vestimentärer Quellen ist in der Forschung ein methodisches Desiderat. Der wissenschaftliche Ansatz liegt demnach in der inhaltlichen Kontextualisierung der Quellen und darüber hinaus in dem Angebot zu disziplinenübergreifenden und kollaborativen Forschungsmöglichkeiten.

Sammlungskonvolute

Den Ausgangspunkt für das Projekt „Restaging Fashion“ bildet die im späten 19. Jahrhundert zusammengetragene Gemäldesammlung des Berliner Verlegerehepaares Franz und Frieda von Lipperheide, welche sich heute im Besitz der Kunstbibliothek Berlin – Staatliche Museen zu Berlin befindet.



Lipperheidesche Kostümbibliothek Berlin (Kunstgewerbemuseum), Fotografie, 1906

Das Konvolut umfasst über 600 Darstellungen, die Mode und Gewandung, Kostüm und Tracht aus dem 15. bis in das 19. Jahrhundert dokumentieren. Seit 1934 war dieser Sammlungsbestand nicht mehr zugänglich, eine Zusammenführung der nach dem 2. Weltkrieg in beide Landesteile verstreuten Gemälde erfolgte erst wieder 1997 am Berliner Kulturforum. Auch dort ist die Gemäldesammlung weder dokumentiert noch zugänglich. Im Kontext von „Restaging Fashion“ wurden in Zusammenarbeit mit der Restauratorin Thuja Seidel und dem Fotografen Dietmar Katz 270 Gemälde auf ihre Transportfähigkeit hin untersucht, wenn nötig gefestigt, entstaubt und auf einem Cruse-Tischscanner hochauflösend digitalisiert.



Scan des Bildnisses von Virginia Guicciardini (Guicciardina ?) im Alter von 20 Jahren, Anonym, 1560-1620, Lipperheidesche Kostümbibliothek

Die Gemäldesammlung war ursprünglich integraler Bestandteil der 1899 von Franz und Frieda von Lipperheide getätigten Schenkung an den Staat: diese enthielt neben den genannten Bildwerken einen großen Bestand an Textquellen, weitere grafische Quellen, darunter frühneuzeitliche Kostüm- und Reisebücher, Journale, Kalender und Almanache, Fotografien und Druckgrafiken, Handzeichnungen und eine große Zahl an Modezeitungen.

Für das Projekt „Restaging Fashion“ werden inhaltlich und formal relevante Handzeichnungen und Druckgrafiken aus dem Bestand ausgewählt, digitalisiert und mit den Gemälden und Miniaturen der Lipperheideschen Sammlung verknüpft. In einer webbasierten Forschungsinfrastruktur werden die digitalen Bilder eingebunden, die Werke inhaltlich tief erschlossen und mit Normdaten und Thesauri beschrieben. Die Sammlungskonvolute werden damit erstmals digital verfügbar gemacht.

Hinzu kommen Textquellen, die in Open Access verfügbar sind, sowie eine Auswahl an erhaltenen historischen Kleidungsstücken aus der Textilsammlung des Germanischen Nationalmuseums in Nürnberg. Diese exemplarisch ausgewählten Ensembles stehen in einem inhaltlichen oder formalen Zusammenhang zu den bildlichen Quellen und werden in ihrer Dreidimensionalität erfasst. Zusammengeführt werden somit Objekte, ihre Metadaten und Schriftquellen aus Archiven, Bibliotheken und Museen.

Der Fokus richtet sich dabei zunächst auf die historische und geografische Verortung von dargestellter Kleidung und darüber hinaus auf ihre Funktion als Kommunikationsmittel, als Distinktionsmittel sowie als Bildargumentation. Am Beispiel der genannten webbasierten Forschungsinfrastruktur werden die Möglichkeiten der Digitalisierung, der inhaltlichen Erschließung, der Visualisierung und deren Funktion als Erkenntnisinstrument aufgezeigt. Die Forschungsinfrastruktur beschreibt und dokumentiert nicht nur dieses Konvolut, sondern auch den Kontext zu weiteren Daten, die als LOD verfügbar sind.

3D-Digitalisierung

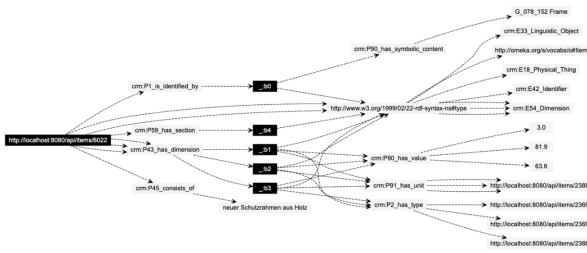
Ein besonderer Schwerpunkt in „Restaging Fashion“ liegt auf der prototypischen 3D-Digitalisierung historischer Kleidung. Die „Inszenierung“ von Kleidern in Online-Sammlungen lehnt sich bisher weitestgehend an den Darstellungen in Print-Katalogen an: die vorherrschende Dokumentation umfasst eine Schauseite und Detailaufnahmen (Rijksmuseum 2017: 13 ff.). Die digitale Bildfassung und photogrammetrische Verarbeitung dagegen wurde bisher vorrangig in Bereichen wie topographische Kartierung, Architektur und Ingenieurwesen und erst seit kurzem für 3D-Museumsobjekte wie Skulpturen oder Reliefbilder eingesetzt. Für die Erfassung der historischen Kleiderensembles steht in dem Projekt „Restaging Fashion“ die Erprobung eines objekt- und materialangemessenen 3D-Digitalisierungsprozesses² nach dem Prinzip der Photogrammetrie sowie eine webbasierte dreidimensionale Präsentation im Vordergrund. In diesem Projekt wird sowohl das einzelne Objekt in seiner Materialität und seinen Eigenschaften verfügbar gemacht als auch abstrakte Dimensionen und der größere Kontext der Sammlung aufgezeigt.

In der Photogrammetrie wird die exakte Lage und Form eines Objektes durch Einzelbilder oder Sensortechnik gemessen, wobei 3D-Koordinaten die Positionen von Objekten in einem Raum definieren. Jede der vier Variablen, d.h. äußere Orientierung, innere Orientierung, Bildkoordinaten und weitere zusätzliche Punkte definieren die Parameter des Abbildungsprozesses und des Modells. Textilien stellen ein interessantes Fallbeispiel für die digitale Dokumentation dar; sie sind zerbrechlich, detailreich und sind oft schwierig zu stabilisieren (z.B. Federn, leichte Textilien). Photogrammetrie erfordert zum einen eine Stabilisierung des Objekts während des Dokumentationsprozesses und zudem eine gleichmäßige Beleuchtung, um ein gutes Datenergebnis zu zeitigen.

Die detaillierte fotografische 3D-Reproduktion soll über die reine Präsentation hinaus mit einem Annotationswerkzeug erweitert werden. In einem weiteren Schritt kann eine Animation der 3D-Digitalisierung verwendet werden, um verblasste Farben, fehlende Kleiderelemente, realistische Bewegung, Dichte, Steifheit oder Bewegtheit des historischen Kleidungsstücks zu simulieren. Das Ziel ist es, einen Workflow zu erproben, um die Vorzüge und Grenzen der 3D-Reproduktionstechnologie für die Kleiderforschung zu untersuchen.

Inhaltliche Erschließung mit Normdaten, Vokabularien und Datenmodellierung

Eine ausführliche inhaltliche Beschreibung der Objekte bildet die Basis der semantischen Kontextualisierung. Die Objekte werden auf Metadaten-Ebene strukturell und inhaltlich erschlossen sowie semantisch angereichert. Als Wissensorganisationssystem fungiert eine technische Infrastruktur, erstellt in Omeka-S und hauptsächlich strukturiert in der erweiterbaren Ontologie CIDOC-CRM (Conceptual Reference Model)³.



Datenmodellierung in Omeka-S

Durch die Verknüpfung von Entitäten (Entities) durch Eigenschaften (properties) können in CIDOC präzise Aussagen über die Objekte getroffen werden, die unendlich erweiterbar sind. Die Modellierung von Ereignissen (Events) ermöglicht zudem komplexe Sachverhalte wie kunsthistorische Diskurse zu den Werken auszudrücken, raumzeitlich zu verorten und strukturiert zu vermitteln. So entstehen Wissensmorphologien, auf die unter anderem mittels Schlagwortsuche und kontrollierter Vokabulare zugegriffen werden kann. Diese detailliert beschriebenen Daten bilden die Basis für die Visualisierungen.

Die Fachterminologie für historische Kleidung, Mode und Tracht stellt dabei eine Herausforderung dar, weil sie regionen- und zeitspezifische, teils unpräzise Bezeichnungen beinhaltet, die nicht angemessen mit bestehenden Kleider-Thesauri oder Iconclass⁴ beschrieben werden können. Hinzu kommt die Problematik einer relativ „flachen“ Erschließungsmöglichkeit von dargestellter Kleidung. Eine detailgetreue Beschreibung jedoch stellt die Basis für die Interpretation der Zeichen- und Symbolhaftigkeit der dargestellten Kleidung dar.

Visualisierung als Forschungsinstrument

Visualisierung fungiert als epistemisches Werkzeug, welches sowohl Aussagen über die Sammlungen und als auch ihre Einzelwerke ermöglichen und fördern soll. Indem die Objekte und ihre Relationen sichtbar gemacht werden, werden Wissensmorphologien anschaulich. In ihrer Funktion als Analysewerkzeug und Erkenntnismittel erlaubt die Visualisierung dabei einen strukturierten Zugriff, auch auf große Datenmengen, wie auch einen dynamischen Zugriff auf vielfältige Dimensionen und Inhalte. Nach von Windhager et. al. (Windhager et. al. 2018: 2316-2317) existieren in der Visualisierung von Kulturerbe-Daten vier Modi der visuellen Granularität. Shneidermans „Visual Information Seeking Mantra: Overview first, zoom and filter, then details-on-demand“ (Shneiderman 1996: 337) folgend werden für das Projekt „Restaging Fashion“ die Daten in verschiedenen visuellen Granularitäten präsentiert. Sowohl Überblicke, multiple Dimensionen also auch das Heranzoomen auf Einzelobjekte sollen in dynamischen und interaktiven Visualisierungen angeboten werden. Zurückgegriffen wird dabei auf den VIKUS-Viewer, einen übertragbaren Prototypen, der von Forschenden des UCLABs für die Visualisierung von kulturellen Sammlungen entwickelt wurde.⁵

Ergebnisse

Im Ergebnis des Projekts steht eine webbasierte Verfügbarkeit von drei Sammlungskonsumenten durch Digitalisierung, inhaltli-

cher Erschließung und Interface-Design. Ein wesentlicher Teil des Projektes beinhaltet dabei die Erforschung von Bildquellen, die Modellierung der Daten und die Verknüpfung dieser Quellen. Das Projekt soll zudem übertragbare Arbeitsprozesse und Szenarien für den Umgang mit 3D-Objekten, hier historischen Kleidungsstücken, deren digitale Erfassung, Datenverarbeitung und Visualisierung im Web liefern. Den Innovationsgrad dieses Projektes stellt einerseits die Integration ausgewählter textiler Objekte wie auch die gleichwertige visuelle Präsentation der verschiedenen Quellen- und Materialarten dar. Ziel ist es die Quellentrias Bild-Text-Realie zusammenzuführen. Dabei steht die Methode der Visualisierung im Fokus, die nicht als bloße Präsentationsform der Objekte aufgefasst wird, sondern als Forschungsinstrument durch semantische Arrangements und interaktive Zugriffsmöglichkeiten die Grundlage für neue Erkenntnisse bildet.

Fußnoten

1. <https://uclab.fh-potsdam.de/projects/restaging-fashion/>.
2. Hier wird die Methode SFL (Structure from Light) verwendet. Softwarepakete, die dabei zum Einsatz für die Betrachtung, Verarbeitung und Modellierung der Daten kommen, sind PhotoScan Pro (<http://www.agisoft.com>), MeshLab (<http://www.meshlabjs.net>), SketchUp (<https://www.sketchup.com>) und CloudCompare (<https://www.cloudcompare.org>). Mit den Bildbearbeitungszeugen wie Zbrush (<http://www.pixologic.com>) können die Modelle weiterbearbeitet werden.
3. <http://www.cidoc-crm.org/>.
4. <http://www.iconclass.org/>.
5. <https://uclab.fh-potsdam.de/projects/vikus-viewer/>.

Bibliographie

- de Günther, Sabine / Zitzlsperger, Phillip (eds.) (2018): *Signs and symbols: dress at the intersection between image and realia*, Berlin.
- Höpfinger, Anna-Katharina (2014): „Clothing as a Meaningful Marker of Space: A Comparative Approach to Embodied Religion from a Cultural Studies Perspective“, in: *Religious Representation in Place: Exploring Meaningful Spaces at the Intersection of the Humanities and Sciences*: 177–192. Papers of the „Meaningful Spaces“ Conference in April 2011 at the Collegium Helveticum, London.
- Knorr-Cetina, Karin (1999): „Viskurse der Physik. Wie visuelle Darstellungen ein Wissenschaftsgebiet ordnen“, in: *Interventionen* 8: 245-63, Hochschule für Gestaltung und Kunst Zürich.
- Rijksmuseum (2015): „Manual for the Photography of 3D Objects“ <https://www.rijksmuseum.nl/en/2d3d-2017/rijksmuseum-manual-for-the-photography-of-3d-objects> [letzter Zugriff 01. Juli 2017]
- Simmel, Georg (1905): *Philosophie der Mode. Moderne Zeitfragen*, Berlin.
- Shneiderman, Ben (1996): The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: *Proceedings. IEEE Symposium on Visual Languages: September 3-6, 1996, Boulder, Colorado*. Los Alamitos, Calif: IEEE Computer Society Press, 336-343.
- Windhager, Florian / Federico, Paolo / Schreder, Gunther / Glinka, Katrin / Dork, Marian / Miksch, Silvia / Mayr, Eva (2018): „Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges“, in: *IEEE Transac-*

tions on Visualization and Computer Graphics <https://doi.org/10.1109/TVCG.2018.2830759> [letzter Zugriff 13. Juli 2021]

Zitzlsperger, Philipp (2015): „Zwischen ‚Lesbarkeit‘ und ‚Unlesbarkeit‘ der Kleider-Codes. Zur bildlichen Repräsentation unauthentischer Kleidung“, in: *Die Medialität der Mode: Kleidung als kulturelle Praxis: Perspektiven für eine Modewissenschaft*, 89-107, Bielefeld.

Digital Environmental Humanities

Zum Potential von „Computational and Literary Biodiversity Studies“ (CoLiBiS)

Langer, Lars

lars.langer@uni-leipzig.de

Spezielle Botanik und Funktionelle Biodiversität, Universität Leipzig

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de

Computational Humanities Group, Universität Leipzig

Borgards, Roland

borgards@lingua.uni-frankfurt.de

Institut für Deutsche Literatur und ihre Didaktik, Universität Frankfurt

Köhring, Esther

koehring@lingua.uni-frankfurt.de

Institut für Deutsche Literatur und ihre Didaktik, Universität Frankfurt

Wirth, Christian

cwirth@uni-leipzig.de

Spezielle Botanik und Funktionelle Biodiversität, Universität Leipzig; Deutsches Zentrum für integrative Biodiversitätsforschung (iDiv), Halle-Jena-Leipzig; Max-Planck-Institut für Biogeochemie, Jena

Einleitung: „Nature’s Contribution to People“

Spätestens seit dem Einsetzen der Industrialisierung erfährt unser Planet einen überdurchschnittlichen Rückgang der Biodiversität (Cardinale et al., 2012; IPBES, 2019; Millennium Ecosystem Assessment, 2005). Diese Biodiversität liefert einen wichtigen Beitrag für unsere Gesellschaft, etwa durch die Bereitstellung von Ressourcen, die Regulierung globaler und lokaler Ökosystemprozesse aber auch auf immaterieller Ebene als Quelle von Inspiration, Bildung und Erholung (Díaz et al., 2018; IPBES, 2019;

Schmid et al., 2009). Der Schutz der Biodiversität ist daher ein drängendes Zukunftsthema (The Global Risks Report, 2021; European Biodiversity Strategy for 2030, 2020), nicht nur in den Naturwissenschaften (Díaz et al. 2019), sondern auch in den kulturwissenschaftlichen Umweltstudien, den sogenannten *Environmental Humanities* (Vidal & Dias, 2017), in denen sich der interdisziplinäre Ansatz des *Ecocriticism* etabliert hat (Bühler, 2016). Eine Zusammenarbeit dieser beiden Forschungszugänge wird von beiden Seiten explizit gefordert (Gesing et al., 2019; Nadim, 2016; Díaz et. al., 2015; Borie & Hulme, 2015), konnte aber mangels operationalisierbarer Methoden bisher kaum realisiert werden, da nicht zuletzt im Gegensatz zu materiellen Beiträgen der Natur für die Gesellschaft (bspw. Nahrung oder Rohstoffe), immaterielle Beiträge (bspw. Erholung oder Inspiration) schwerer zu quantifizieren sind (Daily, 2000; Daniel et al., 2012; Martinez-Alier, 2002). In diesem Beitrag schlagen wir die Digital Humanities (DH) als Intermediär vor, um eine Brücke zwischen natur- und kulturwissenschaftlichen Biodiversitätsforschungen zu bauen. Durch diese interdisziplinäre Vermittlung ergibt sich das neuartige Forschungsgebiet der „Digital Environmental Humanities“, welches die Methoden der DH nutzt, um einen quantitativen Zugang zu immateriellen Beiträgen der Natur durch die computergestützte Analyse von materialisierten Kulturartefakten, wie etwa Büchern, zu untersuchen.

Der heute gebräuchliche Biodiversitätsbegriff wurde in den 1980er Jahren geprägt (Wilson & Peter, 1988). In der Konsequenz ergeben sich für historisch orientierte Untersuchungen (bspw. Literaturgeschichte der Biodiversität) Probleme mit der Anwendbarkeit des Begriffs, sodass man implizit mit einem Anachronismus, einer historischen Rückprojektion, arbeitet. Gleichzeitig hat das Konzept der Biodiversität in dieser Form eine starke wissenschaftliche und moralische Autorität (Toepfer, 2011:361). Biodiversität ist also nicht nur wissenschaftlich messbar, sondern hat auch Auswirkungen auf normatives Handeln und somit auf die Politik. Zusätzlich impliziert der Begriff der Biodiversität gemäß dem Motto „*varietas delectat*“ immer auch eine ästhetische Dimension (vgl. Abb. 1).

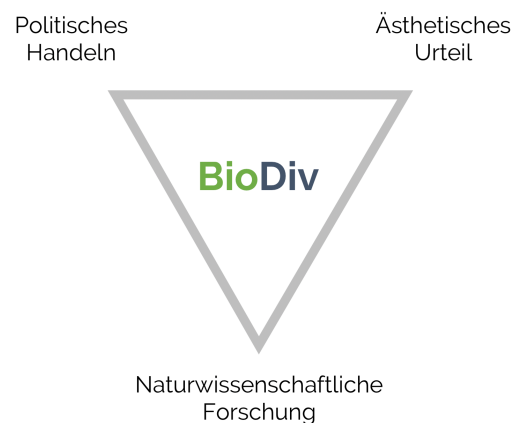


Abb. 1: Drei Dimensionen des Biodiversitätsbegriffs.

Diese Begriffstrias spiegelt sich auch in der kurzen Geschichte des ‚Weltbiodiversitätsrats‘ (IPBES = *Intergovernmental Platform for Biodiversity and Ecosystem Services*) wider: Biodiversität soll demnach wissenschaftlich quantifiziert werden und die Erkenntnisse als Grundlage für politisches (normatives) Handeln herangezogen werden. Dementsprechend wird der hier zunächst

gebrauchte, stark an ökonomischen Modellen ausgerichtete Begriff der „Ecosystem Services“ mittlerweile vom Konzept „Nature’s Contribution to People“ (Beiträge der Natur zur Gesellschaft) ersetzt, welches besser die engen Verflechtungen von Kultur und Natur inklusive ihrer immateriellen Komponenten zu berücksichtigen vermag.

Methodik der CoLiBiS

Bisher leiden Methoden der Quantifizierung immaterieller Beiträge der Natur für die Gesellschaft an subjektiver Verzerrung (vgl. etwa Ainscough, 2019) oder sind von überschaubarem Umfang (vgl. etwa Celis-Diez et al., 2016; Kesebir & Kesebir, 2017; Prévot-Julliard et al., 2015). Ein bislang kaum untersuchter immaterieller Beitrag der Natur ist deren Einfluss auf unsere Kommunikation. So wurden etwa die Nennungen von Lebewesen in unserer Sprache im Laufe der Zeit zu unterschiedlichen Zwecken genutzt, bspw. in Vergleichen („Augen wie ein Luchs“), Metaphern („Wurzel“ als Ursache/Ursprung), Redewendungen („auf den Busch klopfen“), Wortspielen („Katzten wohnen im Miezhaus“), Wortneuschöpfungen („Computermaus“) und *entity naming* (Leipzig – „die Stadt im Lindenhain“). Wir gehen davon aus (vgl. auch Langer et al., 2021), dass Biodiversität in unserer Kommunikation als Indikator für das Bewusstsein des Menschen für seine Lebenswelt betrachtet werden kann (Gagliano, Ryan, & Vieira, 2017; Mesoudi, 2011; Tüür & Tønnessen, 2014). Als Teilbereich der Digital Environmental Humanities schlagen wir für die Untersuchung des Einflusses der Natur auf die literarische Kommunikation den neuartigen Ansatz der „Computational and Literary Biodiversity Studies“ (CoLiBiS) vor. Konkret soll dabei die Analyse der Biodiversität in textuellem Kulturgut durch die Verknüpfung von Methoden aus der Ökologie und den DH erfolgen.

Ökologische Methoden

In der Ökologie wird Biodiversität über Maßzahlen quantifiziert und bezieht sich auf eine bestimmte taxonomische Stufe, wobei entsprechende Vorkommen innerhalb eines vordefinierten Raumes erfasst werden. Ein Taxon ist eine Einheit in der biologischen Systematik der Lebewesen, das aufgrund gleicher Eigenschaften bestimmten Lebewesen zugeordnet wird (bspw. die Säugetiere als Gruppe aller säugenden Wirbeltiere). Die taxonomische Stufe kennzeichnet dabei die Höhe eines Taxons in der biologischen Systematik (bspw. sind die Säugetiere eine „Klasse“ und gehören zum „Unterstamm“ der Wirbeltiere). Die am häufigsten untersuchte Stufe in ökologischen Studien ist die „Art“ (*species*). Übliche Maßzahlen für Biodiversität sind *richness* (Anzahl verschiedener Arten), *abundance* (Anzahl aller Individuen jeglicher Arten), Shannon-Diversität und Simpson-Diversität. Die Shannon-Diversität (Whittaker, 1960) quantifiziert die Unsicherheit der Vorhersage der Art eines zufällig gewählten Individuums im Untersuchungsraum. Die Simpson-Diversität quantifiziert hingegen die Wahrscheinlichkeit, dass zwei zufällig gewählte Individuen aus dem Untersuchungsraum der gleichen Art angehören. Üblicherweise wird hier der Kehrwert genommen, damit höhere Werte mit höherer Diversität korrelieren. Beide Maßzahlen sind abhängig von der Verteilung der Zahl der Individuen zwischen den Arten, so dass eine hohe Gleichverteilung (*evenness*) höhere Werte erzeugt. Der Untersuchungsraum kann stark variieren und in Abstufungen betrachtet werden. So können bspw. verschiedene Plots (vordefinierte Teilabschnitte) einer Wiese als kleinste

Stufe dienen, in der man die Shannon-Diversität von Grasarten bestimmt. Auf der niedrigsten räumlichen Stufe wird dann von α -Diversität (Magurran & McGill, 2011) geredet. Die nächste Stufe könnte nun die gesamte Region (also bspw. die gesamte Wiese) sein, in der man die γ -Diversität bestimmt. Unterschiedliche Plots auf einer homogenen Wiese können möglicherweise die gleichen Arten beinhalten, wodurch γ ähnlich hoch ist wie α , während sie auf einer heterogenen Wiese verschiedene Arten beinhalten, wodurch γ deutlich höher ist als α . Um die Biodiversität zwischen Plots zu quantifizieren nutzt man bspw. den Quotienten von γ geteilt durch α und bezeichnet diesen als β -Diversität. Wir möchten nun im Rahmen der CoLiBiS diese Indizes etablieren, um Biodiversität auch in der literarischen Kommunikation zu quantifizieren. In unserem Fall werden sodann nicht Wiesen und Plots, sondern Bücher und Kapitel untersucht. Die Vorgehensweise hierzu wird in den nachfolgenden Fallstudien näher erläutert.

DH-Methoden

Aus dem Bereich der DH kommen in den CoLiBiS vor allem Text Mining und NLP-Methoden zum Einsatz, um Lebewesen in großen Literaturkorpora automatisiert zu detektieren und damit einer Quantifizierung mithilfe der genannten Biodiversitätsindizes zugänglich zu machen. Zu den relevanten Methoden gehören grundlegende NLP-Verfahren zum *text preprocessing*, etwa Lemmatisierung, um Flexionsformen miterfassen zu können, aber auch *POS-Tagging* und *Named Entity Recognition* zur Disambiguierung von Naturbegriffen und anderen Begriffen (bspw. „to bear“ vs. „the bear“; „Rose“ als Name vs. „rose“ als Blume). Weiterhin kommen Methoden der Segmentierung zum Einsatz, um das Korpus in Plots und Regionen unterteilen zu können und so die Biodiversitätsindizes zuordnen zu können.

Fallstudien

In ersten Fallstudien wurde bereits das Potential des CoLiBiS-Ansatzes erfolgreich erprobt. Das konkrete Szenario war hier die Untersuchung der Vielfalt, die der Erwähnung von Lebewesen in belletristischer Literatur im Zeitraum von 1705 bis 1969 zu-grunde liegt (Langer et al., 2021). Zu diesem Zweck analysierten wir ein englisches Sub-Korpus des *Standardized Project Gutenberg Corpus* (SPGC) (Gerlach & Font-Clos, 2020), welches knapp 16.000 literarische Werke von etwa 4.000 Autoren enthält. Zu den bestehenden Metadaten, die im Wesentlichen Titel und Kategorisierung der Werke sowie Namen und Lebensdaten der Autoren beinhalten, wurden zahlreiche weitere Parameter, wie etwa Geschlecht und Herkunftsort der Autoren, manuell ergänzt und aufbereitet. Die Identifizierung der Lebewesen im Korpus realisieren wir durch den Abgleich mit einem großen Lexikon, welches über 240.000 englische Begriffe für Lebewesen enthält. Dieses Lexikon wurde aus Wikidata und Wikispecies (<https://dumps.wikimedia.org>) extrahiert.

Biodiversität in unserer Kommunikation als Spiegel von Bewusstsein gegenüber der Natur (Fallstudie 1)¹

Alle Werke des eingangs erwähnten SPGC-Subkorpus wurden für die diachrone Analyse in Fünfjahresabschnitte eingeteilt und dann unter Verwendung des beschriebenen Lexikons englischer

Bezeichnungen für Lebewesen und mit Hilfe von NLP-Methoden auf deren Biodiversität durchsucht. Insgesamt wurden so mehr als 4,4 Millionen Okkurrenzen von ca. 6.000 unterschiedlichen Bezeichnungen für Lebewesen gefunden (Beispielsatz: „We sat down on a rude bench, under a group of magnificent *lime trees*). Zur Berechnung der Biodiversitätsindizes eines Werkes unterteilen wir jedes Werk in Abschnitte zu je 1.000 Wörtern (Plots), in denen wir die α -Diversität bestimmten. Nach einer Größennormalisierung wurden bspw. *richness* sowie α -, β - und γ -Diversität bestimmt. Um die Entwicklung über die Zeit nachvollziehen zu können, wurden sämtliche Werke gemäß ihres Fünfjahresabschnitts mithilfe eines Bootstrapping-Ansatzes gemittelt. Alle größen-normalisierten Biodiversitätsindizes, wie bspw. die Shannon-Diversität pro Werk (γ -Diversität), zeigen auf, dass nach einem anfänglichen Anstieg von Biodiversität in Literatur (BiL) in den 1830er Jahren ein Maximum erreicht wurde, und danach ein stetiger Ab-fall von BiL – auch relativ zum jeweiligen Gesamtvokabular – folgte (siehe Abb. 2). Die Ergebnisse der β -Diversität weisen zudem darauf hin, dass nach diesem Höchstwert zusätzlich eine weniger spezifische Nennung von Lebewesen innerhalb von Werken stattfindet.

Diese Ergebnisse interpretieren wir dahingehend, dass BiL bis zur Industrialisierung aufgrund mehrerer simultaner Einflüsse, wie etwa der Öffnung der fiktionalen Literatur, der Weiterentwicklung des Bildungssystems und einer möglichen Bewusstwerdung von Biodiversitätsverlust während der Romantik, zunimmt. Da diese Einflüsse andauerten und wir keine Hinweise auf weitere BiL-reduzierende Vorgänge fanden, wie bspw. Verschlankung biologischen Vokabulars durch Synonymbildung, vermuten wir, dass die darauffolgende Umkehrung der Entwicklung von BiL ab 1837 das Resultat menschlicher Entfremdung von der Natur infolge umfassender sozialer Veränderungen durch die Industrialisierung ist. Inwiefern an dieser Umkehr auch Prozesse beteiligt sind, die sich ausschließlich im literarischen Feld abspielen, wie z.B. epochenspezifische Funktionen von und Ansprüche an Literatur, also sich literaturhistorisch wandelnde poetologische Konzepte, entzieht sich bisher einer Quantifizierung.

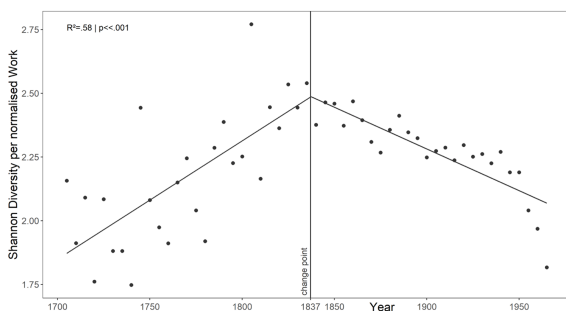


Abb. 2: Durchschnittliche Shannon-Diversität pro größen-normalisiertem Werk über die Fünfjahresabschnitte hinweg. Eine Change-point-Regression erklärte mit $R^2=0,58$ die Messwerte gut, war mit $p < 0,001$ hochsignifikant und zeigte einen Change-point im 1835er Fünfjahresabschnitt.

Parameter für Empfindsamkeit gegenüber Natur (Fallstudie 2)

In einer weiteren, aktuell laufenden Fallstudie, soll untersucht werden, ob es autoren- oder genrespezifische Parameter gibt, die in unmittelbarem Zusammenhang mit einem erhöhten BiL-Wert und damit einer höheren Empfindsamkeit gegenüber der Natur

stehen. Hierzu werden die eingangs erwähnten zusätzlichen Metadaten, die für jedes Werk und jeden Autor durch Daten aus Wikidata ergänzt wurden, herangezogen. Um einen Überblick zu relevanten Parametern zu bekommen, die mit der Wahrnehmung und Verarbeitung von Biodiversität korrelieren, bestimmten wir mit einer *Random - Forest*-Regression die Wichtigkeit und die gegenseitige Abhängigkeit (*partial dependency*) einiger ausgewählter Parameter. *Partial dependencies* einer *Random - Forest*-Regression zeigen den isolierten Einfluss eines Parameters, d.h. unter der Annahme, dass alle anderen konstant gehalten werden. Die *partial dependencies* für den Parameter *Hauptregion* des Autors (siehe Abb. 3) und die *Literaturform* der Werke (siehe Abb. 4) weisen, gerade auch unter Betrachtung der Fehlerbalken, auf einen deutlichen Einfluss dieser Parameter auf die resultierende BiL hin.

Diese ersten Ergebnisse deuten darauf hin, dass die Wahrnehmung eines Autors und die damit verbundene schriftliche Verarbeitung der umgebenden Biodiversität in starkem Zusammenhang mit seiner Hauptregion steht. Es liegt die Vermutung nahe, dass sich dies u.a. mit kulturellen Unterschieden, biogeographischen Eigentümlichkeiten der einzelnen Gegenden und regional verschiedenen Zuständen der Entfremdung von Natur zu tun hat. So können wir vermutlich davon ausgehen, dass im deutlich weniger erschlossenen Nordamerika des 19. Jahrhunderts naturnahe Themen eine höhere Alltagsrelevanz besaßen und Lebewesen eine stärkere Immersion der Wahrnehmung erwirkten als im dichter besiedelten Europa oder gar auf den Britischen Inseln, wo die Industrialisierung am frühesten begann und am stärksten vorangeschritten war (siehe Abb. 3).

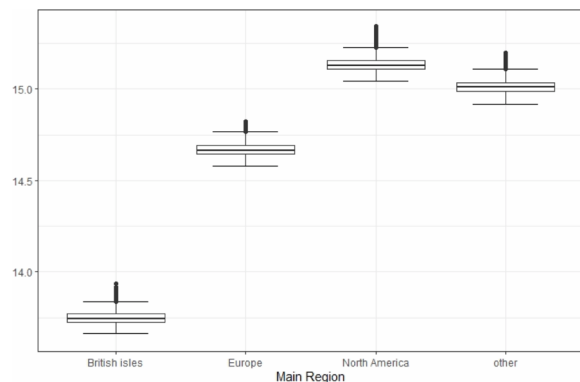


Abb. 3: *Partial dependencies* bezüglich des durchschnittlichen Artenreichtums pro normalisiertem Werk in isolierter Abhängigkeit von der Hauptregion des Autors.

Die Prädisposition eines Werks bezüglich BiL scheint zudem stark mit seiner Literaturform zu korrelieren, wobei vor allem für Lyrik ein stark erhöhter Wert zu beobachten ist (siehe Abb. 4).

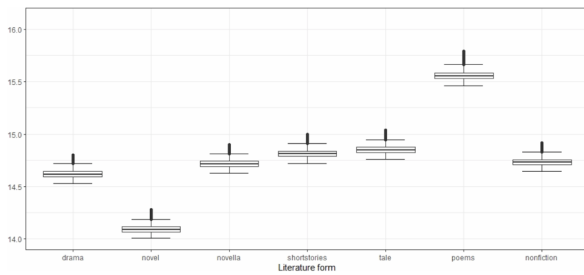


Abb. 4: Partial dependencies bezüglich des durchschnittlichen Artenreichtums pro normalisiertem Werk in isolierter Abhängigkeit von der Literaturform des Werks.

Fazit und Ausblick

In diesem Beitrag schlagen wir mit den „Computational and Literary Biodiversity Studies“ einen interdisziplinären „Digital Environmental Humanities“-Ansatz zur Untersuchung von Biodiversität in der Literatur vor. Zur konkreten Umsetzung von CoLiBiS stellen wir einen Methodenmix aus ökologischen Maßzahlen und textanalytischen Verfahren der DH vor und illustrieren deren Potential anhand zweier Fallstudien. Wir hoffen damit einen Anreiz für weitere, vergleichbare Studien zu schaffen, die weitere Korpora (Weltliteratur vs. Nationalliteraturen) untersuchen und zur Kontextualisierung der Biodiversitätsmaße ggf. weitere Metadaten heranziehen. Weiterhin ist geplant, die bislang erarbeiteten Ressourcen für Folgeforschung frei zugänglich zu machen. Dies betrifft im Wesentlichen eine durchsuchbare Datenbank, welche sämtliche Sätze enthält, in denen mindestens ein Begriff für ein Lebewesen vorkommt. Die Datenbank soll eine Suche nach Keywords sowie auch eine Filterung nach den unterschiedlichen Metadaten erlauben. Weiterhin sollen die erweiterten und manuell redigierten Metadaten zu einem Subkorpus des SPCG als separater Datensatz veröffentlicht werden.

Weiteres Potential für weiterführende Ansätze ergibt sich zudem auf Ebene der DH-Methodik. So könnten künftig etwa *word embeddings* für eine verbesserte Erkennung von Begriffen für Lebewesen jenseits des aktuellen, statischen Lexikonansatzes eingesetzt werden. Dies scheint vor allem im Kontext literarischer Produktion vielversprechend, wo gehäufte, kreative Umschreibungen von Lebewesen („Isgrim“ als Wolf, „Mäusefresser“ als Katze, etc.) zu erwarten sind, die nicht Teil der Wikispecies-Begriffsliste sind. Weiterhin könnte mit LDA Topic Modeling der Kontext der Naturbegriffe erfasst (beispielhafte Kontexte: Stadt, Land, Traumszene, etc.) und mithilfe von Sentiment Analyse hinsichtlich der emotionalen Polarität erfasst werden (positive oder negative Naturdarstellung im Werk von Autor X).

Fußnoten

1. Sämtliche Details der nachfolgenden Fallstudie werden ausführlich in (Langer et al., 2021) beschrieben.

Bibliographie

Ainscough, J., de Vries Lentsch, A., Metzger, M., Rounsevell, M., Schröter, M., Delbaere, B., ... Staes, J. (2019): "Na-

vigating pluralism: Understanding perceptions of the ecosystem services concept", in: *Ecosystem Services*, 36. doi: 10.1016/j.ecoser.2019.01.004

Borie, M., & Hulme, M. (2015): "Framing global biodiversity: IPBES between mother earth and ecosystem services", in: *Environmental Science & Policy*, 54, 487–496. doi: 10.1016/j.envsci.2015.05.009

Bühler, B. (2016): *Ecocriticism*. doi: 10.1007/978-3-476-05489-0

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., ... Naeem, S. (2012): "Biodiversity loss and its impact on humanity", in: *Nature*, 486, (7401), 59–67. doi: 10.1038/nature11148

Celis-Diez, J. L., Díaz-Forestier, J., Márquez-García, M., Lazzarino, S., Rozzi, R., & Armesto, J. J. (2016): "Biodiversity knowledge loss in children's books and textbooks", in: *Frontiers in Ecology and the Environment*, 14 (8), 408–410. doi: 10.1002/fee.1324

Daily, G. C. (2000): "Ecology: The Value of Nature and the Nature of Value", in: *Science*, 289 (5478), 395–396. doi: 10.1126/science.289.5478.395

Daniel, T. C., Muhar, A., Arnberger, A., Aznar, O., Boyd, J. W., Chan, K. M. A., ... von der Dunk, A. (2012): "Contributions of cultural services to the ecosystem services agenda", in: *Proceedings of the National Academy of Sciences of the United States of America*, 109 (23), 8812–8819. doi: 10.1073/pnas.1114773109

Díaz, Sandra, Demissew, S., Joly, C., Lonsdale, W. M., & Larigauderie, A. (2015): "A Rosetta Stone for Nature's Benefits to People", in: *PLOS Biology*, 13 (1). doi: 10.1371/journal.pbio.1002040

Díaz, Sandra, Pascual, U., Stenseke, M., Martín-López, B., Watson, R. T., Molnár, Z., ... Shirayama, Y. (2018): "Assessing nature's contributions to people", in: *Science*, 359 (6373), 270–272. doi: 10.1126/science.aap8826

Díaz, Sandra, Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneeth, A., ... Zayas, C. N. (2019): "Pervasive human-driven decline of life on Earth points to the need for transformative change", in: *Science*, 366 (6471). doi: 10.1126/science.aax3100

EU Biodiversity Strategy for 2030. (2020): *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. Retrieved from https://ec.europa.eu/environment/strategy/biodiversity-strategy-2030_en

Gagliano, M., Ryan, J. C., & Vieira, P. (2017): "The language of plants: Science, philosophy, literature", in: *The Language of Plants: Science, Philosophy, Literature*. University of Minnesota Press.

Gerlach, M., & Font-Clos, F. (2020): "A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics", in: *Entropy*, 22 (1), 126. doi: 10.3390/e22010126

Gesing, F., Knecht, M., Flitner, M., & Amelang, K. (Eds.). (2019): *NaturenKulturen. Denkräume und Werkzeuge für neue politische Ökologien*. Bielefeld: transcript.

IPBES. (2019): *Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services* (S. Díaz, J. Settele, E. S. Brondízio, H. T. Ngo, M. Guèze, J. Agard, ... C. N. Zayas, eds.). doi: 10.5281/zenodo.3553579

Kesebir, S., & Kesebir, P. (2017): "A growing disconnection from nature is evident in cultural products", in: *Perspectives on Psychological Science*, 12 (2), 258–269. doi: 10.1177/1745691616662473

Langer, L., Burghardt, M., Borgards, R., Böhning-Gaese, K., Seppelt, R., & Wirth, C. (zur Publikation angenommen;

2021): "The rise and fall of biodiversity in literature: A comprehensive quantification of historical changes in the use of vernacular labels for biological taxa in Western creative literature", in: *People and Nature*.

Magurran, A. E., & McGill, B. J. (2011): *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press.

Martinez-Alier, J. (2002): *The Environmentalism of the Poor*. doi: 10.4337/9781843765486

Mesoudi, A. (2011): *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. University of Chicago Press.

Millennium Ecosystem Assessment. (2005): *Ecosystems and Human Well-Being: Synthesis* (W. V. Reid, H. A. Mooney, A. Cropper, D. Capistrano, S. R. Carpenter, K. Chopra, ... M. B. Zurek, eds.). Washington DC: Island Press.

Nadim, T. (2016): "Biodiversität erfassen: von Suppen und Satelliten", in: A. Blum, N. Zschocke, V. Barras, & H.-J. Rheinberger (Eds.), *Diversität. Geschichte und Aktualität eines Konzepts* (pp. 68–83). Königshausen & Neumann.

Prérot-Julliard, A.-C., Julliard, R., & Clayton, S. (2015): "Historical evidence for nature disconnection in a 70-year time series of Disney animated films", in: *Public Understanding of Science*, 24 (6), 672–680. doi: 10.1177/0963662513519042

Schmid, B., Balvanera, P., Cardinale, B., Godbold, J., Pfisterer, A., Raffaelli, D., ... Srivastava, D. (2009): "Consequences of species loss for ecosystem functioning: Meta-analyses of data from biodiversity experiments", in: *Biodiversity, Ecosystem Functioning, and Human Wellbeing*, 14–29. doi: 10.5167/uzh-25528

The Global Risks Report. (2021): *The Global Risks Report*. World Economic Forum, Geneva.

Toepfer, G. (2011): "Diversität", in: *Historisches Wörterbuch der Biologie. Geschichte und Theorie der biologischen Grundbegriffe. Bd. 1*. Stuttgart, 351–365.

Tüür, K., & Tønnessen, M. (2014): "The semiotics of animal representations", in: *The Semiotics of Animal Representations* (pp. 7–30). doi: 10.1163/9789401210720_002

Vidal, F., & Dias, N. (Eds.). (2017): *Endangerment, Biodiversity, and Culture*.

Whittaker, R. H. (1960): "Vegetation of the Siskiyou Mountains, Oregon and California", in: *Ecological Monographs*, 30 (4), 407–407. doi: 10.2307/1948435

Wilson, E. O., & Peter, F. M. (Eds.). (1988): *Biodiversity*. doi: 10.17226/989

Dokument, Transkription, Forschungsdatum Technische und kulturelle Überlegungen für interdisziplinäre Transkriptionspraxis

Baierer, Konstantin

konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Boenig, Matthias

boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Engl, Elisabeth

engl@hab.de
Herzog August Bibliothek Wolfenbüttel

Geestmann, Mareen

geestmann@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Hinrichsen, Lena

hinrichsen@hab.de
Herzog August Bibliothek Wolfenbüttel

Neudecker, Clemens

clemens.neudecker@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Pestov, Paul

pestov@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Weidling, Michelle

weidling@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Transkription im weitesten Sinne, d.h. die schriftliche Kodierung von Informationen, bildet die Grundlage für eine Vielzahl von Methoden in den Digital Humanities (DH). Zu transkribieren heißt, ein Quelldokument mittels Regeln und Konventionen in ein Zieldokument zu übertragen.¹ Das Dokument besteht grob betrachtet aus dem Träger, der eine Stele, Tontafel, Papyrusrolle, Papierblätter oder auch eine Datei sein kann, und dem Text.

Der Text bildet sowohl im Zusammenspiel mit dem Trägermaterial als auch in sich selbst mehrere Dimensionen aus. In den meisten Fällen soll durch die Transkription der Text erfasst werden. Die Gesamtheit der Dimensionen kann nur sehr begrenzt kodiert werden.² In Kulturerbeeinrichtungen finden sich verschiedenste Formen von Transkriptionen wie handschriftliche Kopien, Übertragungen von gesprochener Sprache in Parlamentsprotokollen oder auch Transkriptionen, die mit dem Ziel einer Edition erstellt werden. Darüber hinaus kann die Transkription aber auch als Eingabe für Software-Verfahren dienen. Ein Beispiel dafür wäre das Trainieren eines Neuronales Netzes (NN) für die Optical Character Recognition (OCR). Wird dieses trainierte OCR-Modell wiederum auf Dokumente angewendet, entsteht so eine automatische Transkription. Da Transkription auf verschiedenen Ebenen mit unterschiedlichen Rahmenbedingungen innerhalb der DH Anwendung findet, wollen wir uns in diesem Vortrag einigen technischen und kulturellen Aspekten dieser Praxis nähern, diese im Kontext der praktischen Transkriptionsarbeit verorten und schließlich ihren hohen Wert herausstellen.

Das Konzept der Transkription ist so alt wie die Schrift selbst. Die Kulturtechnik des Schreibens wird seit jeher gelernt durch das Abschreiben bestehender Texte (Kopie) oder das schriftliche Festhalten mündlicher Sprache (Diktat). Nur durch die kon-

tinuierliche Transkriptionsarbeit von Kopist:innen und Übersetzer:innen im Laufe der Jahrhunderte können wir heute Texte und die ihnen zugrunde liegenden vormodernen Kulturen rezipieren. Die schriftliche Überlieferung wurde durch die Übertragungs- und Abschreibprozesse auch Veränderungen ausgesetzt, die häufig nicht dokumentiert wurden (vgl. z.B. Kammer 2017: 33–37; Schulz 2014: 290). Mit dem Aufkommen des Buchdrucks in Europa in der Renaissance kam das Setzen als neue Transkriptionsform hinzu, bei der der Aufwand für die Anfertigung von Kopien im Gegensatz zum manuellen Abschreiben mit einer steigenden Anzahl an zu produzierenden Exemplaren immer weiter sinkt. Die Folgen waren eine deutliche Ausweitung des kulturellen Austauschs, die Wiederentdeckung von bisher nur in vereinzelten Manuskripten manifesten Texten und die Ausgestaltung vieler Wissenschaftsdisziplinen wie wir sie heute kennen. Seit dem Aufkommen der Mikrochiptechnologie und des Internets befinden wir uns in der nächsten kulturtechnischen Evolution, in der IT-gestützte Verfahren auf allen Ebenen von Transkriptionsarbeit zum Einsatz kommen.

Da Informationen in IT-Systemen immer binär gespeichert werden, d.h. als Abfolge von Nullen und Einsen, stellt sich die Frage, wie und vor allem wie einheitlich Texte als Binärcode kodiert werden sollen. Bis in die 2000er Jahre hinein war ASCII, eine auf das Englische zentrierte Kodierung aus den 1960er Jahren, mit sprachspezifischen Erweiterungen (Windows-1252, KOI-8 etc.) üblich. Heute hat sich hingegen der Unicode-Standard UTF-8 weitgehend durchgesetzt, dessen Ziel es ist, für jede Letter aus jeder Schrift, historisch wie gegenwärtig, einen äquivalenten Codepoint in Unicode zu definieren. Damit ist es heute möglich, beinahe jedes Zeichen und jede Variante eindeutig zu kodieren. Es ist dementsprechend gängige Praxis, konsequent alle Texte in UTF-8 (und nicht in einer anderen, obsoleten Unicode-Codierung wie UTF-16) zu kodieren. Da Unicode ein offener und aktiv entwickelter Standard ist, werden laufend neue Zeichen darin aufgenommen. Für sehr spezielle Fälle, die (noch) nicht im Unicode-Standard vorhanden sind, gibt es Erweiterungen der Private Use Area (PUA) von Unicode, bspw. die Medieval Unicode Font Initiative (MUFI³). Die Kodierungen entsprechend MUFI sind zwar nicht Teil des offiziellen Unicode-Standards, sie sind aber dennoch gegenüber Eigenlösungen klar zu bevorzugen und werden auch in unregelmäßigen Abständen in Unicode integriert.

Neben der Schrift können in der digitalen Transkription noch weitere Bedeutung tragende (semantische) Dimensionen der Vorlage kodiert werden, bspw. Elemente des typografischen Erscheinungsbilds oder die jeweilige Hand in einem Manuskript (vgl. Neuber 2020). Der gängige Standard für die Erstellung kritischer Editionen mit reichem, semantischem Markup ist das XML-Format der Text Encoding Initiative (TEI-XML). Seit 1987 in Entwicklung, ist TEI-XML ein äußerst detaillierter XML-Standard, der eine Vielzahl von Vorlagen abdeckt (Korrespondenz, Drama, Prosa, Lyrik, Handschriften uvm.). Gegenüber eigenentwickelten Kodierungen ist TEI-XML in den meisten Fällen zu bevorzugen. TEI-XML möchte die Vielzahl der textuellen Dimensionen abbilden, aus diesem Grund ist es ein sehr umfangreicher und flexibler Standard. Da die konkrete Transkription nicht alle Dimensionen erfassen soll, ist es empfehlenswert, vorhandene und in der Praxis erprobte TEI-XML-basierte Transkriptionsrichtlinien wie bspw. das DTABf (DTA-Basisformat, vgl. Haaf et al. 2014) zu nutzen. Das DTABf wird bereits seit vielen Jahren und im Austausch mit der Community gepflegt. Es bietet die Sicherheit, dass für gängige Konstrukte und textuelle Phänomene bewährte Lösungen bereits enthalten sind. Aspekte, die diese Schemas und Richtlinien nicht enthalten, können entsprechend der umfassenden TEI-Kompatibilität ergänzt werden. Dadurch kann weitgehende Interoperabi-

lität zwischen verschiedenen Texten erreicht werden. Bei dieser Vorgehensweise sind die Transkriptionsentscheidungen bereits in bestehenden Richtlinien dokumentiert und der Transformationsaufwand der zusätzlichen textuellen Aspekte in andere Formate oder Sammlungskontexte kann minimiert werden (vgl. Fisseni et al. 2021).

Nicht immer ist das Ziel der digitalen Transkription eine kritische Edition. Moderne Verfahren der OCR und der Handwritten Text Recognition (HTR) beruhen auf Methoden des Deep Learning (DL) bzw. auf NN. Ziel dieser Verfahren ist es, die Transkription von bilddigitalisierten Werken zu automatisieren, indem der Computer lernt, Zeile für Zeile den Text im Bild zu erkennen. Diese Verfahren müssen zunächst mit manuell transkribierten Zeilenbild-Zeilentext-Paaren (Ground Truth, GT) trainiert werden, die repräsentativ für die zu erkennenden Vorlagen sind (vgl. auch Boenig et al. 2018). Während die Zielgruppe für kritische Editionen vornehmlich menschliche Rezipienten sind, werden Transkriptionen für das Training von OCR/HTR von einem Computerverfahren interpretiert und haben dadurch andere Anforderungen. Zum einen kann die semantische Auszeichnung der Transkriptionen von den derzeit implementierten Verfahren nicht verwendet werden, sie sind ausschließlich auf zeilenweisen Text fokussiert. Zum anderen ist eine durchgängig einheitliche Kodierung der GT unerlässlich, da jede Inkonsistenz das Erlernen einer inkorrekten Transkription zur Folge hat. Deshalb sind auch hier gut dokumentierte und auf den Anwendungsfall ausgerichtete Transkriptionsrichtlinien essentiell. Wir empfehlen nachdrücklich die OCR-D-Ground-Truth-Richtlinien (vgl. Boenig et al. 2019), die von OCR-D seit mehreren Jahren aktiv gepflegt werden, eine Transkription auf mehreren Ebenen der Textgenauigkeit ermöglichen und Beispiele für Sonderzeichen wie historische Ligaturen samt möglicher Normalisierungen enthalten.⁴ Diese Richtlinien nehmen Erfahrungen und Methoden auf, die in der Editionswissenschaft, in Editionsprojekten, im DFG-Projekt Deutsches Textarchiv⁵ entwickelt und gesammelt wurden⁶. Einen Schwerpunkt bildet dabei die Beschreibung und Normierung des Verhältnisses zwischen Vorlage und transkribiertem Text. Dazu wurde eine Level-Einteilung⁷ entwickelt, die den Grad der Interpretation der Übertragung in drei Levels unterteilt. Aber nicht nur zur Transkription an sich kann dieses Einteilungssystem genutzt werden, sondern auch zur Messung vorhandener Transkriptionen. So kann einfacher beurteilt werden, ob sich eine Transkription zur Verwendung als Ground Truth für das OCR-Training oder zur Evaluation eignet.

Auch in nachgelagerten Schritten der Textarbeit kommen Verfahren zum Einsatz, die als Transkriptionen angesehen werden können. Die aus Textverarbeitungsprogrammen bekannte Rechtschreibkorrektur transkribiert einen Text mit potentiellen Fehlern zu einem korrigierten Text. Ähnlich verhält es sich mit der (semi-)automatischen Nachkorrektur von OCR mithilfe historischer Sprachmodelle. Auch die Anreicherung mit Informationen über benannte Entitäten im Text (Named Entity Recognition, NER) und deren Verknüpfung mit Normdaten (Named Entity Linking, NEL) sind Texttransformationen, deren Eingabe und Ausgabe Text mit semantischem Markup ist. Daneben gibt es eine Vielzahl von Aufgaben in der Weiterverarbeitung von Text in den diversen DH-nahen Disziplinen, die den Text anreichern, strukturieren, auswerten usw. Auch diese Aufgaben sind häufig Transkriptionen in dem Sinne, dass sich hier dieselben Fragen im Hinblick auf die Kodierung des Textes und dessen Auszeichnung mit Markup stellen. Entsprechend ist es auch hier empfehlenswert, Transkriptionsrichtlinien zu nutzen und von Anfang an explizit festzulegen, wie sich die Ausgabe eines Verfahrens im Text niederschlägt.

Schließlich sei noch auf die sich zunehmend verbreitenden Ansätze hingewiesen, die die Dichotomie von manueller und automatischer Transkription auflösen und Aspekte von beiden Vorgehensweisen kombinieren. Bspw. ist es möglich, mit einer kleinen Menge an manuell transkribierter GT zu beginnen und ein erstes, noch stark fehlerbehaftetes OCR-Modell zu trainieren. Dieses wird anschließend auf weitere Texte angewendet, deren Fehler wiederum manuell korrigiert werden. In einem iterativen Zyklus aus Training, OCR und manueller Korrektur wird das Modell schließlich so lange verbessert, bis es zufriedenstellende Ergebnisse liefert. Neben diesem Bootstrapping-Verfahren, das bspw. in OCR4all (vgl. Reul et al. 2019) Anwendung findet, gibt es auch die Möglichkeit zum Nachtrainieren. Hierbei wird ein bereits vorhandenes Modell mit zusätzlicher GT angereichert und damit bspw. für eine bestimmte Schrifttype, ein bestimmtes Werk oder einen bestimmten Zeitabschnitt angepasst. Dadurch kann mit relativ geringem Aufwand ein für die jeweilige Vorlage maßgeschneidertes OCR-Modell erstellt werden. Auch bei diesen Ansätzen ist es wichtig, nach dafür geeigneten einheitlichen Transkriptionsrichtlinien wie den OCR-D-GT-Guidelines vorzugehen, damit ein fehlerhaftes Training durch Inkonsistenzen und eine damit einhergehende schlechtere Erkennung vermieden werden.

Während die technischen Aspekte von Transkription mit fortschreitender Digitalisierung an Bedeutung gewinnen, dürfen die soziokulturellen Umstände von Transkriptionsarbeit nicht außer Acht gelassen werden. Zunächst sei darauf hingewiesen, dass Transkribieren keine neutrale Aktivität ist, sondern immer im Kontext der Transkribierenden verortet werden muss (vgl. z.B. Alpert-Abrams 2016). Bei manueller Transkription entscheidet ein Mensch, welche Phänomene – und welche nicht – in welcher Form kodiert werden sollen. Diese Selektion und Kodierungspraxis ist abhängig von der Forschungsfrage bzw. der intendierten weiteren Verarbeitung der transkribierten Daten. Während für eine stilometrische Analyse der reine Text benötigt wird, kann für andere Perspektiven auf das Werk das Layout wichtig sein. Für weitere Ansätze sind wieder andere Faktoren interessant, wie Abbildungen, handschriftlich Anmerkungen, die Materialität des Werkes betreffende Informationen wie Wasserzeichen und Papiertextur oder die Lagenformel einer Handschrift. Die Landkarte ist nicht das Gebiet – es ist unmöglich, eine Vorlage unter allen denkbaren Gesichtspunkten exakt zu beschreiben, ohne sie zu reproduzieren. Es ist aber sehr wohl möglich, diese impliziten Selektionskriterien in Form von Transkriptionsrichtlinien explizit zu machen, um potentiellen Nachnutzenden die Einschätzung der Bedeutung einer Transkription für ihre Forschungsfrage zu erleichtern.

Für die automatischen Transkriptionsverfahren gilt diese Beobachtung umso mehr, da jedes trainierte Modell nur so präzise, effektiv und umfassend sein kann wie die Trainingsdaten bzw. das kodierte Kontextwissen, das in das Verfahren einfließt. Bei heuristischen Verfahren, in denen in Form von Software abgefasste Regeln die Verarbeitung bestimmen, sollten diese Regeln klar dokumentiert werden. Nur so kann sichergestellt werden, dass Verfahren und Daten zueinander passen. Wird bspw. ein OCR-Verfahren so trainiert, dass das "lange l" zu einem modernen "runden s" normalisiert wird, wird es niemals das "lange l" produzieren können. Für eine automatische Sprachverarbeitung könnte diese Normalisierung sogar hilfreich sein, für Forschungsfragen zu historischer Orthografie hingegen wären die Erkennungsergebnisse dann ungeeignet. Deshalb ist es auch bei automatischer Transkription wichtig, diese vorgelagerten Entscheidungen und impliziten Annahmen zu dokumentieren.

Transkription ist eine Aktivität von vielen im wissenschaftlichen Diskurs und wird daher auch von dessen Konventionen ge-

prägt. Das oft wiederholte Mantra von "publish or perish" gilt in den DH ebenso wie in anderen Disziplinen. Mit Blick auf das Ziel, eine oder mehrere Publikationen auf Basis der Transkription zu erarbeiten, wird der eigentlichen Transkriptionsarbeit repräsentativ häufig kaum Bedeutung beigemessen. Insbesondere im Bereich der Informatik und der STEM-Disziplinen etabliert sich jedoch zunehmend die Ansicht, dass ein wohlkuriertes Datenset, das gemäß der FAIR-Prinzipien⁸ in einem Forschungsdatenrepository frei nachnutzbar publiziert ist, sehr wohl eine anerkennenswerte akademische Leistung ist. Auch Drittmittelgeber wie die DFG oder die Europäische Kommission haben die Wichtigkeit von Forschungsdaten erkannt und fordern zunehmend eine Forschungsdatenstrategie als Voraussetzung für erfolgreiche Förderung. Qualitativ hochwertige Transkriptionen setzen Domänenwissen sowie größte Sorgfalt voraus und sind zeitintensiv, werden aber bislang akademisch kaum wertgeschätzt und nur selten überhaupt und wenn dann erst sehr spät veröffentlicht. Gerade im Hinblick auf die Synergiemöglichkeiten, die der rasante Fortschritt im Bereich der NN in den letzten Jahren mit sich bringt, ist es bedauerlich, wie wenig rohe Transkriptionsdaten zur Verfügung stehen, um diese Systeme zu trainieren. Ein Kulturwandel hin zu mehr Offenheit und Anerkennung für die Transkriptionsarbeit wäre somit ein doppelter Gewinn für interdisziplinäre Forschung.

Mit mehr Respekt für die Transkriptionsarbeit sollte zudem ein stärkerer, auch interdisziplinärer Austausch zur Transkriptionspraxis einhergehen, bspw. über Fragen, ob TEI-XML das beste Datenmodell für eine rohe Transkription ist oder ob nicht – wie in OCR-D (vgl. Engl et al. 2020), OCR4all, eScriptorium,⁹ Kraken¹⁰ oder Transkribus (vgl. Kahle et al. 2017) – PAGE-XML (vgl. Plutschacher und Antonacopoulos 2010) die bessere Wahl ist; welche Vor- und Nachteile alternative Kodierungsformen für verschiedene Anwendungsfälle haben können; wer in derselben Domäne bereits nach welchen Richtlinien transkribiert hat und Erfahrungen teilen kann; ob es für eine etwaige semi-automatische Erfassung mithilfe von OCR bereits passend trainierte Modelle gibt; uvm. Es ist für die DH im Speziellen, aber auch für den wissenschaftlichen Diskurs insgesamt, ein Gewinn, wenn wir eine interdisziplinäre Diskussion zur Transkriptionspraxis weiterführen. Dadurch kann es uns gelingen, reflektierter mit Ground-Truth-Daten umzugehen, bei deren Erstellung und Qualitätsbestimmung es durch fehlende Richtlinien an Transparenz und universeller Anwendbarkeit fehlt (vgl. Boenig et al. 2018).

Fußnoten

1. Vgl. Theorie der Transkription bei Sahle 2013, S. 279-283.
2. Begriff des Textes bei: Sahle 2013 und Hermes 2011.
3. <https://mufi.info>
4. <https://ocr-d.de/de/gt-guidelines/>
5. Richtlinien zur Transkription: <https://www.deutschestextarchiv.de/doku/basisformat/transkription.html>
6. Anzuführen seien auch die Dokumentationen und Richtlinien des Referenzkorpus Mittelniederdeutsch/Niederrheinisch (https://corpora.uni-hamburg.de/hzsk/de/islandora/object/file:ren-0.1_transkriptionshandbuch/datastream/PDF/transkriptionshandbuch.pdf) sowie Grundsätze für die Textbearbeitung im Fachbereich Historische Hilfswissenschaften hrsg. Archivschule Marburg Version vom 26.04.2009 (https://www.archivschule.de/uploads/Ausbildung/Grundsätze_fuer_die_Textbearbeitung_2009.pdf) und die Blogbeiträge zum Thema Ground Truth aus dem DFG-Projekt "Rechtsprechung

- im Ostseeraum" (<https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/>).
7. Übersicht der Ground-Truth-Level: <https://ocr-d.de/de/gt-guidelines/trans/trLevels.html>
8. <https://www.go-fair.org/fair-principles/>
9. <https://gitlab.com/scripta/escrptorium/>
10. <https://github.com/mittagessen/kraken>

Bibliographie

Alpert-Abrams, Hannah (2016): "Machine Reading the *Primeros Libros*", in: *Digital Humanities Quarterly* 10 (4) <http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html> [letzter Zugriff 15. Juli 2021].

Boenig, Matthias / Baierer, Konstantin / Hartmann, Volker / Federbusch, Maria / Neudecker, Clemens (2019): "Labelling OCR Ground Truth for Usage in Repositories", in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATECH19), Brüssel 09.05.2019* : 3-8 <https://dl.acm.org/doi/abs/10.1145/3322905.3322916> [letzter Zugriff 15. Juli 2021].

Boenig, Matthias / Federbusch, Maria / Herrmann, Elisa / Neudecker, Clemens / Würzner, Kay-Michael (2018): "Ground Truth: Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities?", in: *DHd 2018. Kritik der digitalen Vernunft. 5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum. Konferenzabstracts*. Universität zu Köln 26. Februar bis 2. März 2018: 219-223 10.5281/zenodo.4622316.

Engl, Elisabeth / Boenig, Matthias / Baierer, Konstantin / Neudecker, Clemens / Hartmann, Volker (2020): "Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke", in: *Zeitschrift für Historische Forschung* 47 (2): 223-250 <https://elibrary.duncker-humboldt.com/journals/id/28/vol/47/iss/5737/art/58179/> [letzter Zugriff 15. Juli 2021].

Fisseni, Bernhard / Sandler, Simon / Schulz, Daniela / Boenig, Matthias / Meiners, Hanna-Lena / Sikora, Uwe (2021): "Das DTABf in der Edition – zusammenfassender Evaluationsbericht" (= CLARIAH-DE-Arbeitsberichte 1) 10.14618/ids-pub-10496 [im Druck].

Haaf, Susanne / Geyken, Alexander / Wiegand, Frank (2014): "The DTA 'base format': A TEI subset for the compilation of a large reference corpus of printed text from multiple sources", in: *Journal of the Text Encoding Initiative* 8 10.4000/jtei.1114.

Hermes, Jürgen (2011): *Textprozessierung - Design und Applikation*. Dissertation, Universität zu Köln <https://kups.uni-koeln.de/4561/> [letzter Zugriff 15. Juli 2021].

Kahle, Philipp / Colutto, Sebastian / Hackl, Günter / Mühlberger, Günter (2017): "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents", in: *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* : 19 - 24 <https://ieeexplore.ieee.org/document/8270253> [letzter Zugriff 15. Juli 2021].

Kammer, Stephan (2017): *Überlieferung: Das philologisch-antiquarische Wissen im frühen 18. Jahrhundert* (= Halle'sche Beiträge zur Europäischen Aufklärung 58). Berlin / Boston: De Gruyter 10.1515/9783110520286.

Neuber, Frederike (2020): *Das Konzept einer typografiezentrierten digitalen Edition der Werke Stefan Georges samt einem Modell zur Beschreibung von Mikrotypografie*. Dissertation, Uni-

versität zu Köln <https://kups.uni-koeln.de/12202/> [letzter Zugriff 15. Juli 2021].

Pletschacher, Stefan / Antonacopoulos, Apostolos (2010): "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", in: *Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010*. IEEE-CS Press: 257 – 260 https://www.primaresearch.org/www/assets/papers/ICPR2010_Pletschacher_PAGE.pdf [letzter Zugriff 15. Juli 2021].

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): "OCR4all — An open-source tool providing a (semi-) automatic OCR workflow for historical printings", in: *Applied Sciences* 9 (22) 10.3390/app9224853.

Sahle, Patrick (2013): *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung* (= Schriften des Instituts für Dokumentologie und Editorik 9). Dissertation, Universität zu Köln. Norderstedt: BoD <https://kups.uni-koeln.de/5013/> [letzter Zugriff 15. Juli 2021].

Schulz, Matthias (2014): "Deutsch in Handschrift und gedrucktem Buch im 15. und 16. Jahrhundert", in: Korn, Lorenz / Hoffmann, Birgitt / Stricker, Stefanie (eds.): *Aus Buchwerkstatt und Bibliothek. Manuskriptkulturen des Mittelalters in Orient und Okzident* (= Bamberger interdisziplinäre Mittelalterstudien Vorträge und Vorlesungen 3). Bamberg: University of Bamberg Press: 271-304. <https://fis.uni-bamberg.de/handle/uniba/21106> [letzter Zugriff 15. Juli 2021].

Dramatische Metadaten Die Datenbank deutschsprachiger Einakter 1740–1850

Çakir, Dîlan Canan

dilan.cakir@aol.com
Universität Stuttgart

Fischer, Frank

frank.fischer@dariah.eu
Higher School of Economics, Moskau

Das Hauptinteresse der digitalen Literaturwissenschaft gilt noch immer Volltexten und entsprechenden Korpora. Diese bilden denn auch die Grundlage für die entwickelten Methoden und Praktiken. Allerdings sind in jüngster Zeit Zugänge zu Metadaten wichtiger geworden, nicht nur durch die Initiativen von GLAM-Institutionen wie der Deutschen Nationalbibliothek oder des DLA Marbach (Barnet et al. 2021), die immer mehr Informationen über Schnittstellen verfügbar machen. Der Rückgriff auf Standards und offene Daten führt langfristig zu einem digitalen Ökosystem, in dem jedes Projekt von seiner Umgebung profitiert und auch selbst Daten in die Umgebung einspeisen kann. In diesem Kontext ist die *Datenbank deutschsprachiger Einakter 1740–1850* angesiedelt, die seit Juni 2020 unter der Adresse <https://einakter.dracor.org/> zugänglich ist.

Die Einakter-Datenbank stellt Metadaten für tendenziell alle zwischen 1740 und 1850 recherchierbaren deutschsprachigen Ein-

akter des Sprechtheaters zur Verfügung. Das Datenmodell orientiert sich dabei an Reinhart Meyers *Bibliographia Dramatica et Dramaticorum* (Meyer 1986 – 2011), wurde aber für die digitale Nutzung angepasst. Meyers opulentes Verzeichnis, das seit 1986 alle im 18. Jahrhundert im ehemaligen deutschen Reichsgebiet gedruckten, als Manuskript erhaltenen und gespielten Theaterstücke bibliografiert – schätzungsweise etwa 50.000 Texte bzw. Inszenierungen – hat von Anfang an »das erschreckende Desinteresse einer Wissenschaft an ihren ›Gegenständen‹« belegt, wie es in einer Rezension hieß (Krämer 1998). Denn die literatur- und theaterwissenschaftliche Forschung gründet ihre Aussagen auf einen verschwindend geringen Bruchteil überlieferter Texte: »Die Folge der literaturwissenschaftlichen Selektion ist, daß etwa 90% der dramatischen Produktion des letzten Drittels des 18. Jahrhunderts aus dem Gesichtsfeld der Germanisten entschwand.« (Meyer 2012: 341; ähnliche Zahlen vgl. Brandt-Schwarze/Oellers 2000: 7) Auch die hier behandelten Einakter, die im Untersuchungszeitraum durchaus populär waren, sind durch keinen Kanon gesichert – früh als »Niederungen der Poesie« (Meyer 1920: 4) betitelt, wurden sie größtenteils aus dem kulturellen Gedächtnis verdrängt.

Zum Begriff ›Einakter‹

Nachdem vorangegangene Untersuchungen zum Einakter bereits die »spezifische Eigenständigkeit« dieser Form herausgearbeitet haben (Pazarkaya 1973: 1), möchte unsere Datenbank die Erforschung des deutschsprachigen Einakters auf eine neue, breitere, inklusivere und vor allem digitale Basis stellen. Dabei musste zunächst der Terminus ›Einakter‹ (der erst im ausgehenden 19. Jahrhundert gebräuchlich wird) operationalisiert werden, da er selbst in der Forschungsliteratur meist eher unbestimmt gebraucht wird. So kann er sich auf nicht näher bezeichnete Kurzdramen ebenso beziehen wie auf abendfüllende Dramen ohne Akte bzw. Akteinteilung. Ohne eine Arbeitsdefinition droht die Zusammenstellung eines Einakterkorpus willkürlich zu werden. Für die Einakter-Datenbank haben wir uns hauptsächlich auf die explizite Einaktermarkierung (im Peri- oder Epitext) konzentriert. Denn der aktive Gebrauch von Etiketts wie »in einem Akt«, »in einem Aufzug« usw. – oft von Autor*innen oder Herausgeber*innen selbst hinzugefügt – bildet eine sinnvolle, einfach zu beschreibende Gemeinsamkeit dieser Dramen und grenzt das Korpus nach außen ab. Solche Untertitel sind für das deutschsprachige Drama erst ab der Mitte des 18. Jahrhunderts üblich. Die meisten Stücke in unserem Korpus bringen diese Markierung im Untertitel mit. Für Dramen wie Lessings *Philotas* (Erstdruck 1759), das nur mit dem Untertitel »Ein Trauerspiel« publiziert wurde, ist dann die explizite Bezeichnung als Einakter auf zeitgenössischen Theaterzetteln sowie in der Forschungsliteratur ausschlaggebend. In unserer Datenbank wird jeweils auch auf derlei wechselnde Untertitel, sofern ermittelt, verwiesen.

Für den fokalen Zeitraum von 1740–1850 haben wir momentan 2.305 Einakter verzeichnet und entsprechend unseres Datenmodells mit Titel-, Veröffentlichungs- und Aufführungsdaten, Personenverzeichnissen und Links zu benachbarten digitalen Projekten versammelt. Als Beispieleintrag diene Christian Leberecht Heynes Einakter *Die beiden Billets* (Abb. 1).

Einakter: Anton-Wall: Die beiden Billets	
<p>Heyne, Christian Leberecht (Anton-Wall)</p> <p>Die beiden Billets</p> <p>Nachspiel in einem Akt</p>	
<p>Erscheinungsdaten 9 July 1783 1783</p>	
<p>Anzahl Szenen 11</p>	
<p>In anderen Projekten</p> <ul style="list-style-type: none"> DraCor: per000482 Wikidata: Q79349104 Weber-Gesamtausgabe: A021103 	
<p>Ausgaben</p> <ul style="list-style-type: none"> Königliches Theater der Franzosen. Für die Deutschen. Herausgegeben von J. G. Dyk. Achtel Theil. Leipzig: Dyk 1783, S. 217–268 Zweite Ausgabe. Leipzig: Dyk 1790 Wien: Wallishauser 1802 Deutsche Schaubühne seit Lessing und Schröder bis auf die neueste Zeit. Neuntes Bändchen. Wien: Schade 1825, S. 5–35 	
<p>Personenverzeichnis</p> <ul style="list-style-type: none"> Görge, (m) Rösigen, (f) Schnapps, ein Dorfbarbier, (m) 	
<p>Handlungsort</p> <ul style="list-style-type: none"> Die Szene ist auf einem freyen Platze vor Rösigen's Hause. 	
<p>Vorlage</p> <ul style="list-style-type: none"> Jean-Pierre Claris de Florian Q051740: <i>Les deux billets. Comédie en un acte et en prose</i> [v0000647] [06529417] (1779) «French» 	
<p>In Nachschlagewerken</p> <ul style="list-style-type: none"> <i>Bibliographia dramatica et dramaticorum</i>, 2. Abteilung, Band 28, S. 490–491 	
<p>Formale Aspekte</p> <ul style="list-style-type: none"> Prosa 	

Abb. 1: Ansicht eines einzelnen Einakters.

In einer der wenigen größeren Arbeiten zum deutschen Einakter zählte der schon zitierte Yüksel Pazarkaya für das 18. Jahrhundert um die 500 Stücke (vgl. ebd.: 67), von denen er etwa 200 bis 300 ausgewertet hat (vgl. ebd.: 163 u. passim). Unsere Datenbank geht eine ganze Größenordnung nach oben (wobei unser Untersuchungszeitraum um ein halbes Jahrhundert verschoben wurde, insgesamt aber auch ca. ein Jahrhundert abdeckt). Die Analyse der Gattung wird also auf eine viel breitere Basis gestellt, indem hunderte Stücke in den Blick gelangen, die bisher für die Forschung schlicht nicht sichtbar waren.

Beobachtungen an einer Grundgesamtheit: Eigenschaften des deutschsprachigen Einakters

Durch die systematische Durchforstung der verfügbaren Quellen und dank des enormen Digitalisierungsfortschritts in den letzten zwanzig Jahren kommen wir erstmals der Grundgesamtheit der gedruckten und gespielten Einakter (die oft nur als Erstdruck oder Manuskript überliefert sind) im genannten Zeitraum nahe. Auch wenn die Datenbank laufend ergänzt wird, wird sich an der Größenordnung nicht mehr viel ändern. Daher beschreiben die quantitativen Aussagen, die wir nun treffen können, den deutschsprachigen Einakter im ausgewählten Untersuchungszeitraum auf Basis nahezu des gesamten überlieferten Materials.

Untertitel

Gemäß der vergebenen Untertiteln sind etwa zwei Drittel aller Einakter Lustspiele, Komödien oder Posen. Weniger Anteile verfallen auf neutralere Bezeichnungen wie Schauspiel, Nachspiel oder Drama.

Auffällig ist, dass nur ca. 2,5% der Einakter zu den Tragödien bzw. Trauerspielen gehören (d.h. im Titel als solche bezeichnet werden oder tragisch ausgehen), was vor allem formale Gründe hat. Pazarkaya etwa spricht von der »Unmöglichkeit der Tragödie in einem Akt« (ebd.: 129), da sich eine tragische Handlung nur mühsam und defizitär in die Kürze eines Aktes zwingen lasse. Prozentual mag der gemessene Anteil gering ausfallen, allerdings zeigt unsere Datenbank doch, dass es wesentlich mehr einaktige Tragödien gibt als in der Forschung bisher angenommen wurde.

Diese können nun als Phänomen erstmals zusammen präsentiert werden. Unter diesen tragischen Stücken befinden sich neben mehreren Schicksalsdramen von unter anderem Zacharias Werner, Adolph Müllner und Ernst von Houwald auch Trauerspiele ohne tragischen Ausgang, etwa Gustav Freytags Einakter *Der Gelehrte* (1848).

Autor*innenschaft

Zu den produktivsten Autor*innen im Korpus zählen August von Kotzebue (ca. 90 Einakter), Ignaz Franz Castelli (über 70) und Franz August von Kurländer (über 60). Vielschreiber*innen sind in jener Zeit insgesamt keine Seltenheit (vgl. Schonlau 2014). Kotzebue, der etwa 260 Dramen geschrieben hat, »rühmte sich [...], ein Stück in drei Tagen fertig schreiben zu können« (Wiese 1972: 8). Daneben haben sich vor allem auch – damals wie heute fast gänzlich unbekannte – Laienautor*innen in der dramatischen Kurzform versucht. Zirka 10% der Einakter wurden anonym veröffentlicht, allerdings konnten einige Anonymate durch die Arbeit von Reinhart Meyer und anderen inzwischen aufgelöst werden.

Der Einakter galt zu jener Zeit auch als Einübungsform und hat viele Erstlingswerke hervorgebracht, beispielhaft seien die später mit umfangreicheren Werken äußerst erfolgreich gewordenen Dramatiker Gotthold Ephraim Lessing oder Adolph Müllner genannt.

Die Datenbank macht auch den Anteil von Autorinnen an der Dramenproduktion sichtbarer (dank der Verknüpfung der Autor*innen mit ihren Wikidata-Einträgen können die Informationen zum Geschlecht automatisch bezogen werden; eine Statistik dazu befindet sich auf unserer Website). Zwar sind nur knapp 5% der erfassten Einakterautor*innen weiblich, allerdings treten so neben bekannteren Vertreterinnen wie Luise Adelgunde Victorie Gottsched und produktiven Theaterautorinnen wie Johanna Franul von Weißenthurn oder Charlotte Birch-Pfeiffer auch unbekannte Verfasserinnen von Einaktern zutage.

Übersetzungen

Als erster deutschsprachiger Einakter wird mitunter Johann Ulrich von Königs allegorisches Stück *Die verkehrte Welt* (1725) angeführt. Allerdings prangt der explizite Hinweis auf dessen Einakterigkeit erstmals nachträglich in einer Ausgabe von 1749 (»Ein Lust-Spiel von einem Aufzuge«). In Frankreich gab es den Untertitelzusatz »en un acte« mindestens 20 Jahre bevor er auch im Deutschen eingeführt wurde. Das erste gedruckte deutschsprachige Drama mit einem Einakterhinweis ist wahrscheinlich *Die Widersprecherin*, das 1741/42 in Gottscheds Dramenanthologie *Die Deutsche Schaubühne* abgedruckt wurde.

Der Hinweis auf Frankreich ist auch deshalb wichtig, weil ein knappes Drittel der Einakter nachweislich aus Übersetzungen oder Adaptionen besteht, und zwar in großer Mehrzahl (über 90%) aus dem Französischen; andere Sprachen spielen kaum eine Rolle. Die Vorlagen für diese Übersetzungen oder Bearbeitungen haben wir ebenfalls in der Datenbank erfasst, inklusive Metadaten zu Werken und Autor*innen.

Der hohe Anteil an Übersetzungen hatte zwei Reaktionen zur Folge: Entweder man beschwerte sich über ihre Massenhaftigkeit oder man fand diese berechtigt, da es so wenige gute deutsche Einakter gebe. Castelli, einer der produktivsten Autoren im Korpus, ist gleichzeitig ein Vielübersetzer – Vielschreiberei galt im Untersuchungszeitraum allerdings oft als verwerflich, über Castelli liest man das Folgende: »Auch bei ihm also hat die Quantität

der Qualität geschadet, Castelli hat zu viel geschrieben, um Bedeutendes geschrieben zu haben.« (Anonym 1854: 14) In Frankreich erfolgreiche Stücke wurden teils auch mehrfach übersetzt, Molières Einakter *Les précieuses ridicules* ist dabei ein Extrembeispiel im Korpus, mit mindestens sieben individuellen einaktigen Übersetzungen zwischen 1752 und 1824.

Schauplätze

Unter dem Personenverzeichnis eines Dramas werden typischerweise Ort und Zeit der nachfolgenden Handlung vermerkt. In mehraktigen Stücken kann sich der Ort im Verlauf des Stücks mehrfach verlagern, bei Einaktern ist das normalerweise nicht der Fall. In unserer Datenbank sind geeignete Ortsangaben via Wikidata kodiert (zu den nicht kodierbaren Informationen gehören fiktive Orte wie »Krähwinkel«, unbestimmte Angaben wie »Eine große Stadt in Deutschland« oder allzu allgemeine Lokalisierungen à la »Die Handlung geht vor in Amerika«).

Über Wikidata werden dann auch die Geokoordinaten bezogen und mit der JavaScript-Bibliothek Leaflet auf eine Weltkarte gemappt, die ebenfalls auf der Website zu finden ist. Dadurch wird auch ein geodatenbasierter Zugang zum Korpus möglich. Auf besagter Karte wird deutlich, dass sich die Handlung deutschsprachiger Einakter für den von uns beobachteten Zeitraum vor allem in den Grenzen des Alten Reichs entfaltet. Neben Berlin und Wien ist allerdings Paris die mit Abstand häufigste Ortsangabe, was freilich an der Vielzahl von Übersetzungen liegt. Einakter, die in der Neuen Welt, im Nahen, Mittleren oder Fernen Osten spielen, bilden die absolute Ausnahme. Durch unseren exhaustiven Ansatz bei der Korpuszusammenstellung lassen sich diese Ausnahmen aber zum ersten Mal systematisch erfassen.

Anzahl der Szenen

Ein Einakter entspricht formal und inhaltlich nicht einfach nur einem einzelnen Akt eines mehraktigen Dramas. Im Schnitt hat ein genuiner Einakter mehr Auftritte als die Einzelakte umfangreicherer Dramen.

Abb. 2 zeigt die Anzahl von Szenen pro Einakter in chronologischer Verteilung. Der Durchschnitt liegt bei 14 Szenen. Für die Mehrzahl der Stücke ist die Szenenanzahl zwischen 7 und 20 angesiedelt, wie man im dunklen Innern der Datenwolke erkennen kann. Als Vergleich seien die 131 zwischen 1740 und 1850 publizierten fünftaktigen deutschsprachigen Dramen des GerDraCor-Korpus (vgl. Fischer et al. 2019) herangezogen, deren durchschnittliche Aktlänge bei knapp unter 7 Szenen liegt.

Die Spieldauer für die Einakter beträgt typischerweise zwischen 15 Minuten und einer Stunde (teils ist diese explizit mit angegeben).

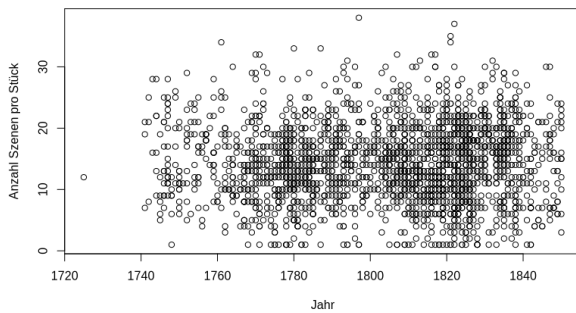


Abb. 2: Anzahl von Szenen pro Einakter (chronologisch).

Anzahl der Figuren

Zu den Eigenschaften des Einakters gehört neben der insgesamt geringen Anzahl an Szenen auch eine reduzierte Anzahl an Figuren. Folge davon ist auch eine potenziell weniger komplexe Handlung. Goethe spricht im Bezug auf Michael Beers Einakter *Der Paria. Trauerspiel in einem Aufzuge* (Erstaufführung 1823) lobend von einer »bewiesenen großen Oekonomie«: »Ohne Zwang sind alle jene tragischen Motive in einen einzigen Akt zusammengebracht, die Handlung entwickelt sich an einem einzigen Ort und der handelnden Person sind nur drey.« (Goethe 1824: 107)

Die Figurenanzahl der Einakter liegt durchschnittlich bei 7, wobei die Mehrzahl der Stücke zwischen 5 und 7 Personen bzw. Sprechinstanzen im Personenverzeichnis auflistet (Abb. 3). Um wieder mit GerDraCor zu vergleichen: Dort liegt die durchschnittliche Anzahl der Figuren für die 131 Fünfakter des Zeitraum 1740–1850 bei 29. Das Verhältnis zwischen weiblichen und männlichen Figuren liegt für dieselben Fünfakter bei 23:100; hingegen in der Einakter-Datenbank (momentan 2305 Stücke) bei 46:100. Der höhere Anteil weiblicher Figuren in einaktigen Stücken lässt sich unter anderem mit bestimmten Handlungsschwerpunkten erklären (viele einaktige Eheanbahnungskomödien bei nur wenigen Tragödien und historischen Dramen).

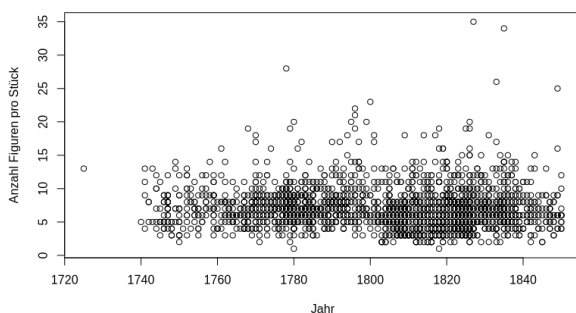


Abb. 3: Anzahl von Figuren pro Einakter (chronologisch).

Die Einakter-Datenbank als Forschungsservice

Unser Datenmodell ist zugeschnitten auf den konkreten Anwendungsfall und im Projekt-Wiki dokumentiert. Die Daten selbst werden im YAML-Format erfasst, die Versionierung erfolgt per Git. Die entstandene Software steht unter der MIT-Lizenz, die Daten unter einer CC-BY-Lizenz.

Übliche bibliografische Metadaten zu den gesammelten Einaktern werden kombiniert mit Links in die Umgebung – darunter DraCor, Wikidata und Projekte wie die Carl-Maria-von-Weber-Gesamtausgabe, aber auch analoge literaturwissenschaftliche Nachschlagewerke wie Meyers *Bibliographia* (Meyer 1986 – 2011), das *Dramenlexikon des 18. Jahrhunderts* (Hollmer/Meier 2000) oder die im Wehrhahn Verlag erschienenen Kotzebue- und Iffland-Lexika (Birgfeld et al. 2011 und Dehrmann/Košeninina 2009).

Die hier präsentierten Ergebnisse können jeweils auf dem aktuellen Stand der Datenbank überprüft werden. Vorgehaltene und errechnete Daten werden über leicht zugängliche Endpunkte exponiert, die unsere Daten im CSV- und JSON-Format anbieten. Die Daten können entweder für die Nutzung in Tabellenkalkulationen wie Microsoft Excel oder LibreOffice Calc heruntergeladen oder direkt über eine Programmiersprache bezogen werden. Beispiele für die Verwendung in R gibt es auf der Website des Projekts.

Bibliographie

Anonym (1854): *Ignaz Franz Castelli. Mit Portrait*. Cassel: Balde.

Barnet, Arno / Dietrich, Elisabeth / Kolbe, Ines / Schmidgall, Karin (2021): »Vom Nutzen vernetzter Werke. Das Kooperationsprojekt »Werktitel als Wissensraum« des Deutschen Literaturarchivs Marbach und der Herzogin Anna Amalia Bibliothek Weimar«, in: *Zeitschrift für Bibliothekswesen und Bibliographie* 68/3: 138–151. (doi:10.3196/186429502068327)

Birgfeld, Johannes / Bohnengel, Julia / Košenina, Alexander (Hg.) (2011): *Kotzebues Dramen. Ein Lexikon*. Hannover: Wehrhahn.

Brandt-Schwarze, Ulrike / Oellers, Norbert (2000): *Die Dramen der Fürstlichen Bibliothek Corvey 1805–1832*. München: Fink. (urn:nbn:de:bvb:12-bsb00040797-1)

Dehrmann, Mark-Georg / Košenina, Alexander (Hg.) (2009): *Ifflands Dramen. Ein Lexikon*. Hannover: Wehrhahn.

Fischer, Frank et al. (2019): »Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama«, in: *Proceedings of DH2019: Complexities*, Utrecht University. (doi:10.5281/zenodo.4284002)

Goethe, Johann Wolfgang (1824): »Die drey Paria«, in: *Ueber Kunst und Alterthum. Fünften Bandes erstes Heft*. Stuttgart: Cotta, 101–108.

Hollmer, Heide / Meier, Albert (Hg.) (2001): *Dramenlexikon des 18. Jahrhunderts*. München: C.H. Beck.

Krämer, Jörg (1998): »Reinhart Meyer, Bibliographia Dramatica et Dramaticorum. Kommentierte Bibliographie der im ehemaligen deutschen Reichsgebiet gedruckten und gespielten Dramen des 18. Jahrhunderts nebst deren Bearbeitungen und Übersetzungen und ihrer Rezeption bis in die Gegenwart. 2. Abteilung: Einzeltitel. 6 Bde. 1993–1996« [Rezension], in: *Arbitrium* 16/2: 131–134. (doi:10.1515/arbi.1998.16.2.131)

Meyer, Elise Marie (1920): *Der Einakter in der deutschen Dichtung des achtzehnten Jahrhunderts*. Leipzig.

Meyer, Reinhart (1986–2009): *Bibliographia dramatica et dramaticorum. Kommentierte Bibliographie der im ehemaligen deutschen Reichsgebiet gedruckten und gespielten Dramen des 18. Jahrhunderts nebst deren Bearbeitungen und Übersetzungen und ihrer Rezeption bis in die Gegenwart*. 1.–2. Abteilung, 1.–30. Band. Tübingen: Niemeyer.

Meyer, Reinhart (2010–2011): *Bibliographia dramatica et dramaticorum. Kommentierte Bibliographie der im ehemaligen deutschen Reichsgebiet gedruckten und gespielten Dramen des 18. Jahrhunderts nebst deren Bearbeitungen und Übersetzungen und ihrer Rezeption bis in die Gegenwart*. 2. Abteilung, 31.–34. Band. Berlin / New York: De Gruyter.

Meyer, Reinhart (2012): »Der Anteil des Singspiels und der Oper am Repertoire der deutschen Bühnen in der zweiten Hälfte des 18. Jahrhunderts« [1981], in: Matthias J. Pernerstorfer (Hg.): *Ders.: Schriften zur Theater- und Kulturgeschichte des 18. Jahrhunderts*. Wien: Hollitzer, 341–400.

Pazarkaya, Yüksel (1973): *Die Dramaturgie des Einakters. Der Einakter als eine besondere Erscheinungsform im deutschen Drama des 18. Jahrhunderts*. Göttingen: Kümmerle.

Schonlau, Anja: »Es war nicht immer Kotzebue. Eine Revision der Kanonisierung des populären Theaterstücks um 1800«, in: Ina Karg / Barbara Jessen (Hg.): *Kanon und Literaturgeschichte. Facetten einer Diskussion*. Frankfurt/M.: Peter Lang 2014, 259–282.

Wiese, Benno (1972): »Einführung«, in: Jürg Mathes (Hg.): *August von Kotzebue: Schauspiele*. Frankfurt/M.: Athenäum, 7–42.

Emotionen im kulturellen Gedächtnis bewahren

Dennerlein, Katrin

katrin.dennerlein@uni-wuerzburg.de
Institut für Deutsche Philologie, JMU Würzburg

Schmidt, Thomas

thomas.schmidt@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Wolff, Christian

Christian.Wolff@sprachlit.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Einleitung

Zwischen dem Ende des Dreißigjährigen Krieges und dem Anfang der Restaurationsepoche entwickelt sich das Drama rasant und wird im deutschsprachigen Gebiet zur publikumswirksamen Gattung dieses Zeitraums (Baumann 1985; Jahn 1996; Krämer 1998; Urchueguía 2015; Meid 2009: 327–501). Es wird zu einer ‚Schule der Affekte‘ (*palaestra affectum*, Rotermund 1972: 25), in der man lernen soll, gewünschte Emotionen zu empfinden und mit unerwünschten wie Angst, Neid oder Leid angemessen umzugehen (Schings 1971; Schings 1980; Wiegmann 1987; Meier 1993; Schulz 1988; Zeller 2005; Lukas 2005; Schonlau

2017). Heutige Konzepte von Emotionen bilden sich dabei erst nach und nach heraus, so dass sich einzelne Figurenaussagen und die Motivation der Handlung für Rezipient*innen der Gegenwart nicht immer unmittelbar erschließen. Die derzeitigen Bemühungen um eine digitale Konservierung von Dramentextbeständen und ihre Analyse müssen folglich um eine Erschließung der Emotionsstrukturen ergänzt werden, wenn dieser Teil des kulturellen Erbes so im kulturellen Gedächtnis konserviert werden soll, dass er auch verständlich bleibt (vgl. zum Begriff des kulturellen Erbes Tauschek 2013). Gemeint ist dabei nicht ein Alltagsverständnis von Gedächtnis im Sinne dessen, „was das Bewußtsein bewußt erinnert“ (Luhmann 1999: 44). Es geht vielmehr um das Gedächtnis sozialer Systeme, das durch Kommunikation in der aktuellen Gegenwart benutzt und reproduziert wird. Plädiert wird vorwiegend dafür, die Emotionen in Dramen des Untersuchungszeitraumes im Speichergedächtnis zu bewahren (vgl. Assmann 2003: 133 ff.). Für das Funktionsgedächtnis, das nur das aktuell Anschlussfähige und Zukunftsorientierte verfügbar hält, sind jedoch zumindest diejenigen Emotionsverteilungen und -verläufe von Interesse, die auch heute noch die Dramaturgie von Dialogformaten bestimmen.

Literaturwissenschaftler haben bisher schwerpunktmäßig wenige Texte einzelner Gattungen hinsichtlich der Frage untersucht, welche emotionale und rationale Wirkung durch sie erreicht werden soll (Pikulik 1965; Sauder 1974–1980; Schings 1971; Nolting 1986).¹ Nur wenige Wissenschaftler*innen erforschen Emotionen der Figuren selbst (Anz 2011; Schulz 1988; Mönch 1993, 344–350; Schonlau 2017). Das führt dazu, dass sehr wenig darüber bekannt ist, welche Figurenemotionen in Dramen dargestellt werden und wann sie im Handlungsverlauf zum Einsatz kommen.

Obwohl die Untersuchung von Gefühlen und Emotionen in den letzten Jahren in der computergestützten Literaturwissenschaft für zahlreiche Textgenres an Bedeutung gewonnen hat (siehe Mohammad 2011; Nalisnick/Baird 2013; Reagan et al. 2016; Zehe et al. 2016; Schmidt/Burghardt 2018; Mc Hardy/Adel/Klinger 2019; Kim/Klinger 2019; Jacobs 2019; Pianzola/Rebora/Lauer 2020), wurden speziell emotionale Aspekte in Dramentexten bisher nur vereinzelt untersucht. Der Fokus lag dabei auf der Analyse von Valenz oder Polarität, d.h. der positiven respektive negativen Konnotation von Sätzen und Textteilen, und zumeist auch auf einzelnen Autoren bzw. Werken (Mohammad 2011; Nalisnick/Baird 2013; Schmidt/Burghardt 2018; Schmidt/Burghardt/Dennerlein 2018a; 2018b; Schmidt/Burghardt/Wolff 2019; Schmidt/Burghardt/Dennerlein/Wolff 2019a; 2019b). Die Untersuchung von Emotionsverteilungen und -verläufen in einzelnen Texten, in verschiedenen Genres und Epochen ist das Ziel des Projekts „Emotions in Drama“, das seit April 2020 Emotionen in Dramen dieses Zeitraums in einem kombinierten Verfahren aus händischer Annotation und ihrer Vorhersage mittels Deep Learning-basierter Sprachmodelle erforscht.² Im Folgenden sollen konzeptionelle Überlegungen und erste Ergebnisse dieses Projekts vorgestellt werden.

Emotionsset

Im Anschluss an Schwarz-Friesel verstehen wir Emotionen als [...] *mehrdimensionale, intern repräsentierte und subjektiv erfahrbare Syndromkategorien, die sich vom Individuum ichbezogen und introspektiv-geistig sowie körperlich registrieren lassen, deren Erfahrungswerte an eine positive oder negative Bewertung gekoppelt sind und die für andere in wahrnehmbaren Ausdrucksvarianten realisiert werden (können)*. (Schwarz-Friesel 2007: 55).

Der Ausdruck von Emotionen kann sich physiologisch, sprachlich oder im Verhalten zeigen. Wir annotieren die von der Figur gemeinte eigene oder zugeschriebene Emotion und nicht diejenigen Emotionen, die Produzenten haben oder die bei Rezipienten ausgelöst werden sollen.³ Diese gemeinten Emotionen können von den explizit in der Sprache repräsentierten abweichen, so dass bei der Annotation oftmals mehrstufige Schlussprozesse und kontextabhängige Interpretationen nötig sind.

Im Untersuchungszeitraum gibt es mehrere historische Kategoriensysteme, die oftmals eine Mischung aus Tugenden und Affekten aufweisen (Zeller 2005; Grimm 2010). Um die Veränderung in der Gewichtung von Emotionen im Untersuchungszeitraum fassen zu können, ist es nötig, solchermaßen von diesen Konzepten zu abstrahieren, dass sie sowohl für den gesamten Zeitraum anwendbar sind als auch Umgewichtungen und Entwicklungen abbilden können.⁴ Deshalb haben wir uns dafür entschieden, die Oberkategorien *Angst*, *Leid*, *Freude* und *Zuneigung* zu verwenden und diese ausdifferenzieren. Die Notwendigkeit der Ausdifferenzierung sei an einigen Beispielen kurz erläutert: Die Kategorie des Leids, die zweifellos zentral für das gesamte europäische Drama ist, ist um die neu aufkommende Kategorie des Mitleids zu ergänzen, die mit Lessing ab der Mitte des 18. Jahrhunderts zur Zentralkategorie wird (Schings 1980). Ähnlich verhält es sich mit Liebe, die um zwei Kategorien zu ergänzen ist. Einerseits die der Lust, die bspw. im Barockdrama oder später im Drama des Sturm und Drang eine zentrale Rolle bei der Abwertung von Figuren spielt. Andererseits ist hier eine Oberkategorie nötig, die mit ‚Zuneigung‘ betitelt ist, um auch Freundschaft einbeziehen zu können. Diese Emotion ist zentral für die Anthropologie der Aufklärung und zusammen mit Mitleid und der Abwesenheit von Lust vermutlich charakteristisch für den Epochenumbruch vom barocken zum aufklärerischen Drama (Sauder 1974-1980; Lukas 2005). Hinzugekommen sind zudem eine Kategorie der Ablehnung, die Ärger und Wut/Abscheu beinhaltet, weil sie zentral für ein grundlegendes Verständnis der Handlung ist. Die Kategorie der ‚emotionalen Bewegtheit‘ schließlich erscheint zusätzlich nötig, um starke gefühlsmäßige Bewegtheit, die aber inhaltlich nicht genauer spezifiziert wird, auszeichnen zu können. Ab und an wissen die Figuren selbst nicht, was sie fühlen (sollen) und/oder schwanken zwischen mehreren Gefühlen.

Diese Kategorien entsprechen, mit leicht abweichenden Benennungen, auch anderen Emotionstaxonomien, die in der Linguistik und in der Literaturwissenschaft verwendet werden (Schwarz-Friesel 2007; Flüh 2020). Die meisten Emotionsrepräsentationen im Natural Language Processing (NLP) folgen Kategoriensystemen der Psychologie, zumeist den Basisemotionen von Plutchik (Plutchik 1980; Wood et al. 2018a; 2018b). Basal sind hier bspw. Angst, Wut, Freude und Trauer, die wir auch verwenden, aber auch einige andere, die wir nicht als Hauptemotionsbezeichnungen übernommen haben, wie ‚Vertrauen‘, ‚Vorfriede‘, ‚Überraschung‘ und ‚Ekel‘. Diese Emotionen können in unserem Schema ebenfalls annotiert werden, sind jedoch in andere Konzepte integriert. Die ‚Vorfriede‘ ist der ‚Freude‘ eingemeindet, ‚Überraschung‘ der ‚emotionalen Bewegtheit‘, ‚Ekel‘ ist in der Form von Abscheu im Sinne starker Ablehnung berücksichtigt.

Das Annotationsset setzt sich nach jetzigem Stand wie folgt zusammen (+/-: Polarität):⁵

- Emotionen der Zuneigung
 - Lust (+)
 - Liebe (+)
 - Freundschaft (+)
 - Verehrung, Bewunderung (+)
- Emotionen der Freude

- Freude (+)
- Schadenfreude (+)
- Emotionen der Angst
 - Angst (-)
 - Verzweiflung (-)
- Emotionen der Ablehnung
 - Ärger (-)
 - Abscheu, Wut, Hass (-)
- Emotionen des Leids
 - Leid (-)
 - Mitleid (-)
- Emotionale Bewegtheit (keine Polarität)

Annotiertes Korpus

Bisher wurden in den folgenden 11 Dramen aus dem GerDraCor-Korpus (Fischer et al. 2019) und aus einem Korpus von im deutschsprachigen Gebiet sehr erfolgreichen Werken des *Leopoldstädter Theaters* in Wien (Kasperl-Stücke) Emotionen ausgezeichnet.⁶ Diese Dramen sind repräsentativ für das noch weitgehend kanonische Gesamtkorpus, enthalten jedoch mit dem *Mandollettikrämer* und dem *Postzug* auch bereits zwei Werke außerhalb des Kanons:

- Luise Adelgunde Victorie Gottsched: Das Testament (1745), Komödie
- Johann Elias Schlegel: Canut (1746), Tragödie
- Christian Fürchtegott Gellert: Die zärtlichen Schwestern (1747), Komödie
- Johann Gottlieb Benjamin Pfeil: Lucie Woodvil (1757), Tragödie
- Joachim Wilhelm von Brawe: Der Freigeist (1758), Tragödie
- Gotthold Ephraim Lessing: Minna von Barnhelm, oder das Soldatenglück (1767), Komödie
- Cornelius von Ayrenhoff: Der Postzug (1769), Komödie
- Ferdinand Eberl: Kasperl' der Mandollettikrämer (1789), Komödie [Libretto]
- Friedrich Schiller: Kabale und Liebe (1784), Tragödie
- August von Kotzebue: Menschenhass und Reue (1790), Komödie
- Johann Wolfgang Goethe: Faust. Eine Tragödie (1807), Tragödie

Insgesamt haben wir bislang 12.364 Annotationen mit einer durchschnittlichen Länge von ca. 25 Tokens gesammelt. Jeweils zwei Annotator*innen haben diese Dramen mit Emotionsinformationen angereichert. Die Übereinstimmung der Annotator*innen (*Inter-Annotator Agreements*) entsprechen Kappa-Werten auf der Ebene der Einzelemotionen zwischen 0,3 und 0,4 je nach Drama, was einer schwachen bis moderaten Übereinstimmung entspricht (Landis/Koch 1977). Für die Polaritätsklasse erhöhen sich die Werte auf 0,4-0,6. Diese vergleichsweise geringen Übereinstimmungswerte sind üblich für die Annotation historischer und literarischer Texte (Alm/Sproat 2005; Sprugnoli et al. 2016; Schmidt/Burghardt/Dennerlein 2018b; Schmidt/Burghardt/Dennerlein/Wolff 2019b; Schmidt et al. 2019). Es ist geplant, die Werte durch kontinuierliche Verbesserung der Annotationsanleitung und Schulung der Annotator*innen noch weiter zu verbessern. Ausgezeichnet werden Textabschnitte von variabler Größe, die von einem Wort bis zu mehreren Sätzen reichen, weil sich der Ausdruck von Emotionen auf unterschiedlich lange Textabschnitte beziehen kann. Neben Emotionen werden auch Quell-

und Zielinformationen annotiert. Weiterführende Informationen zur Annotation finden sich bei Schmidt/Dennerlein/Wolff (2021b; 2021c).

Ergebnisse

Hauptklassen und Sub-Emotionen	absolut	%
Emotionen der Zuneigung	2 928	22
Lust	52	0
Liebe	1 755	13
Freundschaft	345	3
Verehrung, Bewunderung	776	6
Emotionen der Freude	1 943	15
Freude	1 619	13
Schadenfreude	324	2
Emotionen der Angst	1 257	9
Angst	721	5
Verzweiflung	536	4
Emotionen der Ablehnung	3 028	23
Ärger	1 625	12
Abscheu, Wut, Hass	1 403	11
Emotionen des Leids	2 700	20
Leid	2 069	16
Mitleid	631	4
Emotionale Bewegtheit	1 408	11
Gesamt	13 264	100

Abb. 1: Absolute und prozentuale Verteilung der annotierten Emotionen.

Die am häufigsten annotierte Einzelemotion ist Leid (vgl. Abb. 1). Dies erscheint zunächst verwunderlich, weil das Verhältnis von Tragödien zu Komödien im annotierten Korpus mit fünf zu sechs eine positivere Emotion als Hauptemotion erwarten lässt. Zu bedenken ist jedoch, dass Komödien vor dem obligatorischen guten Ende (Kraft 2011) oftmals beträchtliche Hindernisse enthalten, so dass negative Emotionen eben doch eine zentrale Rolle spielen.

Für Verlaufsanalysen der Emotionsannotationen wird jedes Drama nach der Zeichenzahl in fünf gleiche Teile (Quintile) geteilt und die Zahl an Emotionsannotationen pro Kategorie und pro Quintil berechnet. Das Verfahren der Aufteilung in Quintile wird in Anlehnung an eine fünfstufige Struktur gewählt, die zahlreiche Dramen, aber eben nicht alle aufweisen. Es ermöglicht normalisierte Vergleiche von Trends in Dramen, die unterschiedliche viele Szenen und Akte haben. Zum Vergleich der Genres (Tragödien vs. Komödien) wird pro Quintil der Durchschnitt aus den jeweiligen Annotationen aller Dramen eines Genres gebildet.

Wirft man nun einen Blick auf die Verteilung der Emotion ‚Leid‘ im Handlungsverlauf der Dramen, so sieht man, dass diese Emotion in Tragödien durchschnittlich genau doppelt so häufig annotiert wurde wie in Komödien, wie an der y-Achse der Abb. 2 abzulesen ist (durchschnittlich 27-32 Textstellen mit Leid in der Komödie, 45-60 in der Tragödie):

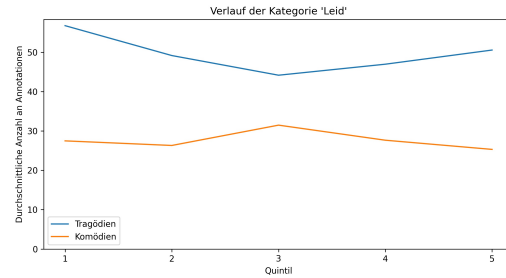


Abb. 2: Verlauf der Annotationen der Emotion ‚Leid‘ in Tragödien und Komödien.

Zudem kann man in Abb. 2 sehen, dass Leid zu Beginn und am Ende von Tragödien gehäuft auftritt. In der Mitte der Dramen besteht jedoch offensichtlich Hoffnung auf eine Verbesserung der Situation und die Figuren empfinden weniger Leid. In Komödien erkennen wir hingegen nach einer kurzen Abnahme von Leid einen leidvollen Höhepunkt, der jedoch zugleich der Wendepunkt zum Besseren, weniger leidvollen Geschehen ist.

Die Tatsache, dass Leid auch in Komödien die wichtigste Emotion ist, hat sicher auch damit zu tun, dass es sich bei den von uns annotierten Komödien vorwiegend um rührende und satirische Komödien handelt, in denen Zustände und Verhaltensweisen kritisiert und auch einige Tränen über leidvolle Ereignisse vergossen werden. Erst zum Schluss hin fügt sich alles zu einem unverhofft freudigen Ausgang. Diese Thesen über den Handlungsverlauf in der Komödie lassen sich auch durch die Verlaufskurve der Emotion ‚Freude‘ belegen. In Komödien ist in der Mitte der Handlung, wenn sich die Verwirrungen und Probleme häufen, Freude am wenigsten annotiert, zum Schluss hin steigen die Werte wieder fast auf das Niveau des Anfangs (vgl. Abb. 3). In Tragödien finden sich hingegen kurz vor der Mitte der Handlung die meisten freudigen Aussagen von Figuren (vgl. Abb. 3). Dieser Befund des plötzlichen Abfalls von Freude korreliert mit dem dramaturgischen Konzept der Peripetie, dem Glückswechsel. Nach der idealtypischen aristotelischen Definition führt der Handlungsumschwung unvermeidlich zum schlechten Ausgang. Die Ergebnisse unserer Annotationsanalyse zeigen eine dazu passende stetige Abnahme von Freude in der Tragödie:

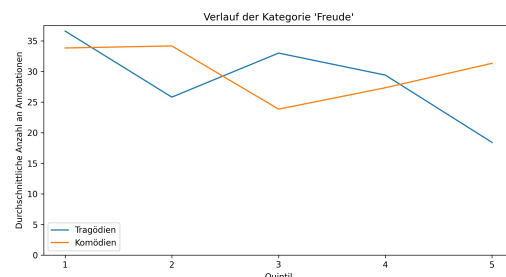


Abb. 3: Verlauf der der Annotationen der Emotion ‚Freude‘ in Komödien und Tragödien.

Diskussion und Ausblick

Die bisherigen Ergebnisse sind vielversprechend, was die Erkennung emotionaler Schlüsselstellen betrifft. Poetologisch lassen sie sich gut mit dem erklären, was wir über die kanonischen Dramen wissen, allerdings wurden bisher nur elf ausgewählte Dramen mit Emotionen ausgezeichnet. Ziel ist es einerseits, alle 300 bisher digitalisierten Dramen zu analysieren und neue Erkenntnisse über die Verteilung von Emotionen zu gewinnen. Andererseits soll auch geprüft werden, wie sich die Emotionsverteilung verändert, wenn im größeren Umfang nicht-kanonische Dramen einbezogen werden. Zu diesem Zweck erweitern wir das Korpus um ausgewählte Dramen der Wanderbühne, um Dramen des Schultheaterautors Christian Weise und um Libretti der Hamburger Oper.⁷

Methodisch gehen wir so vor, dass mithilfe der annotierten Textstellen transformerbasierte Sprachmodelle mit Architekturen wie BERT oder ELECTRA auf die Erkennung von Emotionsklassen trainiert werden (Schmidt/Dennerlein/Wolff 2021b; 2021c). Hier gibt es durch die Entwicklungen der letzten Jahre vielversprechende kontextsensitive Verfahren. Während bisherige Verfahren aus dem traditionellen maschinellen Lernen oder für statische Sprachmodelle Wörter meist ohne Beachtung des Kontextes (also der Wortumgebung) repräsentieren, haben Wörter und Ausdrücke in dynamischen Sprachmodellen je nach Kontext eine andere Repräsentation. Auch können die Modelle besser als bisherige Verfahren mit Wörtern und Phrasen umgehen, die nicht im Vokabular enthalten sind. Diese Modelle sind jedoch zumeist mit zeitgenössischem Sprachmaterial trainiert (z. B. mit Web-Texte wie der Wikipedia oder News-Texten, vgl. Chan/Schweter/Möller 2020). Deshalb planen wir, die bestehenden Sprachmodelle auf historische und fiktionale Sprache, später dann auf dramenspezifische Sprache zu trainieren. Erste vielversprechende Projekte konnten auf diese Weise im Kontext deutschsprachiger und historischer Texte erfolgreiche Prädiktionsleistungen erreichen (Labusch et al. 2019; Schweter/Baiter 2019; Brunner et al. 2020a; Schweter/März 2020). Bisher haben wir die Leistung verschiedener transformerbasierter Modelle evaluiert, die auf zeitgenössischer Sprache vortrainiert sind (unter anderem von Chan/Schweter/Möller 2020) und solche, die mit historischen und/oder literarischen Texten nachtrainiert wurden (unter anderem Schweter 2020, Brunner et al. 2020), sowie das Fine-Tuning von BERT-Modellen mithilfe unserer eigenen Korpora und Theaterstücke untersucht. Ein erstes Zwischenergebnis ist, dass die Vorhersagen mit *gbert-large* und *gelectra-large* von *deepset* (Chan/Schweter/Möller 2020) mit 83% Prädiktionsgenauigkeit die besten sind. Diese Zahl bezieht sich auf die Einstufung einer Textstelle als positiv oder negativ. Soll vorausgesagt werden, ob eine Textstelle zu einer unserer Emotionsoberklasse gehört, beträgt die Genauigkeit 57%, für die Vorhersage einer der 13 Einzelemotionen beträgt sie 47%.⁸ Weitere Details und Informationen zu ersten Ergebnissen der Emotionsklassifikation auf dem annotierten Korpus findet man bei Schmidt/Dennerlein/Wolff (2021a; 2021c). Hier ist noch ein erheblicher Aufwand nötig, um Sprachmodelle mit Texten unseres Untersuchungszeitraumes und mit Dramen nachzutrainieren.

Fußnoten

1. Während ‚Gattung‘ im Anschluss an Fricke als literaturwissenschaftlicher Ordnungsbegriff verwendet wird, soll im Folgenden der Begriff ‚Genre‘ für alle gruppenförmigen Erscheinungen

von Dramen gebraucht werden, die durch historisch identifizierbare „ge- und bewußte[] Normen [bestimmt sind, die] die Produktion und Rezeption von Texten bestimmen“ (Fricke 1981: 132).

2. Das Projekt wird von der Deutschen Forschungsgemeinschaft (DFG) für drei Jahre im Rahmen des Schwerpunktprogramms *Computational Literary Studies* (SPP 2207/1) gefördert https://dfg-spp-cls.github.io/projects_en/2020/01/24/TP-Emotions_in_Drama/ (Projektnummer 424207618, Sachbeihilfen DE 2188/3-1 und WO 835/4-1).

3. In der literaturwissenschaftlichen Emotionsforschung lassen produktions-, rezeptions-, text- und kontextbezogene Ansätze unterscheiden (Winko 2003).

4. Das Annotationsschema wurde zu Projektbeginn iterativ bei der Annotation erprobt und mehrfach angepasst, wie es für die Geisteswissenschaften üblich und empfohlen ist (Reiter 2020).

5. Die Bewertung erfolgt ausgehend von der empfindenden Figur. Schadenfreude ist hier folglich deshalb als positive Emotion ausgewiesen, weil sie für die Figur, die sie empfindet, eine freudig-positive Empfindung ist.

6. http://lithes.uni-graz.at/maeze/maeze_startseite.html.

7. Die Ergänzungskorpora müssen teilweise aufwändig volltextdigitalisiert und alle noch mit TEI ausgezeichnet werden. Vgl. z.B. <https://www.germanistik.uni-wuerzburg.de/lehrstuehle/computerphilologie/mitarbeiter/dennerlein/digitalisierung-von-libretti-der-hamburger-gaensemärkte-opern-von-1678-1730/>.

8. Die Zuordnung einzelner Emotionen zu Oberklassen funktioniert hier noch leicht abweichend.

Bibliographie

Alm, Cecilia Ovesdotter / Sproat, Richard (2005): „Emotional sequencing and development in fairy tales.“, in: *International Conference on Affective Computing and Intelligent Interaction*. Springer.

Anz, Thomas (2011): „Todesszenarien: literarische Techniken zur Evokation von Angst, Trauer und anderen Gefühlen.“, in: Lisanne Sauerwald (Hrsg.): *Emotionale Grenzgänge. Konzeptualisierungen von Liebe, Trauer und Angst in Sprache und Literatur*. Königshausen & Neumann: 54–59.

Arntzen, Helmut (1968): *Die ernste Komödie. Das deutsche Lustspiel von Lessing bis Kleist*. Nymphenburger Verl. 1968.

Assmann, Aleida: *Erinnerungsräume. Formen und Wandlungen des kulturellen Gedächtnisses*. 2. Auflage. Beck.

Bauman, Thomas: *North German opera in the age of Goethe*. Cambridge University Press, 1985.

Brunner, Annalen / Duyen Tanja Tu, Ngoc / Weimer, Lukas / Jannidis, Fotis (2020): „To BERT or not to BERT-Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation.“, in: *SwissText/KONVENS 2020*.

Chan, Branden / Schweter, Stefan / Möller, Timo (2020): „German's Next Language Model.“, in: *arXiv preprint arXiv:2010.10906*

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hecht, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): „Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama.“ Zenodo. <https://doi.org/10.5281/zenodo.4284002>

Flüh, Marie (2020): „Emotionsanalyse“. In *forTEXT. Literatur digital erforschen*. <https://fortext.net/ressourcen/tagsets/emotionsanalyse>

- Fricke, Harald** (1981): *Norm und Abweichung. Eine Philosophie der Literatur*. C.H. Beck.
- Grimm, Gunther E.** (2010): "Affekt.", in: Karlheinz Barck, Martin Fontius, Dieter Schlenstedt, Burkhard Steinwachs und Friedrich Wolfzettel (Hrsg.), *Ästhetische Grundbegriffe*. Metzler, Bd. 1., 16-49.
- Jacobs, Arthur M.** (2019): "Sentiment analysis for words and fiction characters from the perspective of computational (Neuro-) poetics.", in: *Frontiers in Robotics and AI* 6 (2019): 53.
- Jahn, Bernhard** (1996): "Das Libretto als literarische Leitgattung am Ende des 17. Jahrhunderts? Zu Zi(e)glers Roman Die Asiatische Banise und seinen Opernfassungen", in: Eleonore Sent (Hrsg.): *Die Oper am Weißenfeller Hof*. Hain: 143–170.
- Kim, Evgeny / Klinger, Roman** (2019): "A survey on sentiment and emotion analysis for computational literary studies.", in: *Zeitschrift für digitale Geisteswissenschaften*.
- Kraft, Stefan** (2011): *Zum Ende der Komödie. Eine Theoriegeschichte des Happyends*. Wallstein.
- Krämer, Jörg**: *Deutschsprachiges Musiktheater im späten 18. Jahrhundert: Typologie, Dramaturgie und Anthropologie einer populären Gattung*. Niemeyer.
- Labusch, Kai / Neudecker, Clemens / Zellhofer, David** (2019): "BERT for Named Entity Recognition in Contemporary and Historical German.", in: *Proceedings of the 15th Conference on Natural Language Processing*, Erlangen, Germany.
- Landis, J. Richard / Koch, Gary G.** (1977): "The measurement of observer agreement for categorical data.", in: *biometrics* (1977): 159-174.
- Luhmann, Niklas** (1999): "Kultur als historischer Begriff". In: Niklas Luhmann: *Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft*. Bd. 4. Suhrkamp, 31-54.
- Lukas, Wolfgang** (2005): *Theodizee und Anthropologie. Studien zum Moraldiskurs im deutschsprachigen Drama der Aufklärung (ca. 1730–1770)*. Vandenhoeck & Rupprecht.
- Mc Hardy, Robert / Adel, Heike / Klinger, Roman** (2019): "Adversarial Training for Satire Detection: Controlling for Confounding Variables.", in: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics.
- Meid, Volker** (2009): *Die deutsche Literatur im Zeitalter des Barock. Vom Späthumanismus zur Frühaufklärung 1570–1740*. C.H. Beck, 327-501.
- Meier, Albert**: *Dramaturgie der Bewunderung: Untersuchungen zur politisch-dramatistischen Tragödie des 18. Jahrhunderts*. Klostermann, Vittorio.
- Mohammad, Saif** (2011): "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales.", in: *arXiv preprint arXiv:1309.5909*
- Mönch, Cornelia** (1993): *Abschrecken Oder Mitleiden: Das Deutsche Bürgerliche Trauerspiel Im 18. Jahrhundert: Versuch Einer Typologie*. Niemeyer.
- Nalisnick, Eric T. / Baird, Henry S.** (2013): "Character-to-character sentiment analysis in Shakespeare's plays." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.
- Neuhuber, Christian** (2003): *Das Lustspiel macht Ernst. Das Ernste in der deutschen Komödie auf dem Weg in die Moderne: von Gottsched bis Lenz*. Berlin.
- Nolting, Winfried** (1986): *Die Dialektik der Empfindung: Lessings Trauerspiele „Miss Sara Sampson“ und „Emilia Galotti“: Mit Einer Einleitung, Gemischte Gefühle: Zur Problematik eines explikativen Verstehens der Empfindung*. F. Steiner Verlag.
- Pianzola, Federico / Rebora, Simone / Lauer, Gerhard** (2020): "Wattpad as a Resource for Literary Studies. Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins.", in: *PLOS ONE*, vol. 15, no. 1, Public Library of Science, p. e0226708. *PLoS Journals*.
- Pikulik, Lothar** (1966): *Bürgerliches Trauerspiel und Empfindsamkeit*. Böhlau.
- Plutchik, Robert** (1980): *Emotion. A psychoevolutionary synthesis*. Harper & Row.
- Reagan, Andrew J. / Mitchell, Lewis / Kiley, Dylan / Danforth, Christopher M. / Dodds, Peter Sheridan** (2016): "The emotional arcs of stories are dominated by six basic shapes." *EPJ Data Science* 5.1: 1-12.
- Reiter, Nils** (2020): "Anleitung Zur Erstellung von Annotationsrichtlinien". *ResearchGate*, <doi: 10.1515/9783110693973-009>
- Rotermund, Erwin** (1972): *Affekt und Artistik: Studien zur Leidenschaftsdarstellung und zum Argumentationsverfahren bei Hofmann von Hofmannswaldau*. W. Fink.
- Sauder, Gerhard** (1974-1980): *Empfindsamkeit*. Metzler. 3 Bde.
- Schings, Hans-Jürgen** (1971): "Consolatio Tragoediae. Zur Theorie des barocken Trauerspiels". in: *Deutsche Dramentheorien. Beiträge zu einer historischen Poetik des Dramas in Deutschland*, herausgegeben und eingeleitet von Reinhold Grimm. Athenäum, Bd. 2, 1-44.
- Schings, Hans-Jürgen** (1980): *Der mitleidigste Mensch ist der beste Mensch: Poetik Des Mitleids von Lessing Bis Büchner*. C.H. Beck.
- Schmidt, Thomas** (2019): "Distant Reading Sentiments and Emotions in Historic German Plays", in: *Abstract Booklet, DH_Budapest_2019*. Budapest, Hungary, 57-60.
- Schmidt, Thomas / Burghardt, Manuel** (2018): "An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing", in: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, 139-149.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin** (2018a): "'Kann man denn auch nicht lachend sehr ernsthaft sein?' – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen", in: *Book of Abstracts, DHD 2018*.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin** (2018b): "Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior.", in: Sandra Kübler, Heike Zinsmeister (eds.), *Proceedings of the Workshop on Annotation in Digital Humanities (annDH 2018)*. Sofia, Bulgaria, 47-52.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin / Wolff, Christian** (2019a): "Katharsis - A Tool for Computational Drametrics", in: *Book of Abstracts, Digital Humanities Conference 2019 (DH 2019)*. Utrecht, Netherlands.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin / Wolff, Christian** (2019b): "Sentiment Annotation in Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts", in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Leipzig, Germany.
- Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian** (2021a): "Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language", in: *Proceedings of the 5th Joint SIGHUM*

Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 67–79.

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021b): "Towards a Corpus of Historical German Plays with Emotion Annotations", in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021c): "Using Deep Neural Networks for Emotion Analysis of 18th and 19th century German Plays", in: *Fabrikation von Erkenntnis: Experimente in den Digital Humanities*. Melusina Press. DOI:10.26298/melusina.8f8w-y749-udlf

Schmidt, Thomas / Winterl, Brigitte / Maul, Milena / Scharf, Alina / Vlad, Andrea / Wolff, Christian (2019): "Inter-Rater Agreement and Usability: A Comparative Evaluation of Annotation Tools for Sentiment Annotation", in: Draude, C., Lange, M. & Sick, B. (Hrsg.), *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*. Bonn: Gesellschaft für Informatik e.V., 121–133. DOI: 10.18420/inf2019_ws12

Schonlau, Anja (2017): *Emotionen im Dramentext: eine methodische Grundlegung mit exemplarischer Analyse zu Neid und Intrige 1750–1800*. De Gruyter.

Schulz, Georg-Michael (1988): *Tugend, Gewalt und Tod: das Trauerspiel der Aufklärung und die Dramaturgie des Pathetischen und des Erhabenen*. Niemeyer.

Schwarz-Friesel, Monika (2007): *Sprache und Emotion*. Francke

Schweter, Stefan (2020): *Europeana BERT and ELECTRA models*. <https://doi.org/10.5281/zenodo.4275044>.

Schweter, Stefan / Baiter, Johannes (2019): "Towards robust named entity recognition for historic german.", in: *arXiv preprint arXiv:1906.07592* (2019).

Schweter, Stefan / März, Luisa (2020): "Triple E-Effective Ensembling of Embeddings and Language Models for NER of Historical German.", in: *CLEF (Working Notes)*.

Sprugnoli, Rachele / Tonelli, Sara / Marchetti, Alessandro / Moretti, Giovanni (2016): "Towards sentiment analysis for historical texts.", in: *Digital Scholarship in the Humanities* 31.4: 762–772.

Tauschek, Markus: *Kulturerbe. Eine Einführung*. Reimer 2013.

Urchueguía, Cristina: *Allerliebste Ungeheuer: Deutsches komisches Singspiel 1760–1790*. Stroemfeld 2015.

Weiss-Schletterer (2005): *Das Laster des Lachens. Ein Beitrag zur Genese der Ernsthaftigkeit im deutschen Bürgertum des 18. Jahrhunderts*. Böhlau.

Wiegmann, Hermann (1987): *Die ästhetische Leidenschaft: Texte zur Affektenlehre im 17. und 18. Jahrhundert*, herausgegeben von Hermann Wiegmann. Olms.

Winko, Simone (2003): *Kodierte Gefühle. Zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900*. Erich Schmidt Verlag.

Wood, Ian / McCrae, John / Andryushechkin, Vladimir / Buitelaar, Paul (2018a): "A comparison of emotion annotation schemes and a new annotated data set.", in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wood, Ian / McCrae, John / Andryushechkin, Vladimir / Buitelaar, Paul (2018b): "A comparison of emotion annotation approaches for text.", in: *Information* 9.5: 117.

Zehe, Albin / Becker, Martin / Hettinger, Lena / Hotho, Andreas / Reger, Isabella / Jannidis, Fotis (2016): "Prediction of happy endings in German novels based on sentiment informa-

tion.", in: *3rd Workshop on Interactions between Data Mining and Natural Language Processing*, Riva del Garda, Italy.

Zeller, Rosemarie (2005): "Tragödien-theorie, Tragödien-praxis und Leidenschaften". In *Passion, Affekt und Leidenschaft in der Frühen Neuzeit*, herausgegeben von Johann Anselm Steiger, Harrassowitz, Bd. II: 691–704.

Empirische Aufmerksamkeitseffekte multimodaler Kohäsion im Film

Laubrock, Jochen

laubrock@uni-potsdam.de

Universität Potsdam; Medizinische Hochschule Brandenburg, Germany

Tseng, Chiao-I

tseng@uni-bremen.de

Universität Bremen

Wir kombinieren die Diskursmethode multimodaler Kohäsion mit empirischen Daten zu Aufmerksamkeit und narrativem Verstehens. Multimodale Kohäsion bezieht systematisch die in auditiver, visueller und verbaler Modalität auftretenden Ereignisse auf modalitätsübergreifende Diskursstrukturen. Wir nutzen diese Diskursstrukturen, um daraus theoriegeleitet empirisch prüfbare Vorhersagen abzuleiten. Wir überprüfen mit Blickbewegungsexperimenten und Fragebogenstudien, wie kohäsive Hinweise Aufmerksamkeit und das Verständnis des Narrativs beeinflussen. Konkret haben wir mit Videobearbeitungssoftware kritische kohäsive Hinweisreize z.B. aus der Eingangsszene von Hitchcock's "The Birds" entfernt und mittels Eyetracking die Aufmerksamkeitsverteilung von insgesamt 114 Betrachtern gemessen. Unterschiedliche Gruppen von Probanden sahen Originale und manipulierte Versionen. Die kritischen kohäsiven Hinweisreize wurden im Original deutlich häufiger betrachtet als äquivalente Orte in der manipulierten Version. Also werden kohäsive Hinweise im Normalfall tatsächlich beachtet. Die Effekte kohäsiver Hinweise wirken nach: In einer anschließenden, für beide Versionen identischen Szene zeigten Betrachter ohne narrative Hinweise eine deutlich diffusere Orientierungsverteilung. Narrative Elemente im Film lenken die Aufmerksamkeit des Betrachters.

Von verbaler zu multimedialer Kohäsion

Die Analyse von Kohäsionsmitteln ist seit langem etabliert als Methode der linguistischen Textanalyse (Halliday & Hasan 1976). Sie basiert auf der Beobachtung, dass z.B. Wiederholungen und Rekurrenzen linguistischer Muster funktional dafür sind, wie Sätze in einem Text als zusammenhängend betrachtet werden. Obwohl alle verbalen Texte verschiedene Formen von Kohäsionsmitteln nutzen, hat es sich bisher trotz einiger diesbezüglicher Corpusstudien (z.B. Flowerdew & Mahlberg 2009; Tanskanen 2006;

Hoffmann 2012) als schwierig erwiesen, direkte Verbindungen zwischen verschiedenen Arten von Kohäsion und Diskursinterpretationen zu ziehen, beispielsweise bezüglich Anaphern oder Koreferenzketten.

Martin (1992: Chapter 3) entwickelte eine funktional organisierte Diskurssemantik für verbal Texten. Kohäsion wird betrachtet als Menge kommunikativer Ressourcen zur Präsentation und Verfolgung von Diskursreferenzen über Text hinweg mit Fokus vor allem auf Menschen, Orten und Objekten und zur Klassifikation der Verbindungen korrespondierender Elemente durch spezifizierte Präsentations- und Verfolgungsstrategien. Dadurch dass Martin für Kohäsion explizit eine höhere Abstraktionsebene vorsieht, die sich von den spezifischen linguistischen Elementen und Formen unterscheidet, wird es leichter, semiotisch unterschiedliche (z.B. verbale und visuelle) Elemente in gemeinsame Diskursstrukturen zu integrieren.

Hier erweitern wir diese Linie auf das audiovisuelle Medium des Films. Film eignet sich besonders gut zur explorativen kohäsiven Analyse. Erstens kombinieren Filme nicht nur Text und Bild, sondern auch gesprochene und geschriebene Sprache, Klänge, Bewegungen und andere visuell getragene Information wie Betrachtungspunkte, Gestik, Mimik, Nähe etc. (Bordwell 2007) in einer bewusst integrativen Art und Weise. Zweitens sind Filme trotz ihrer semiotischen Komplexität immer noch primäre lineare expressive Formen, die sich normalerweise linear in der Zeit entfalten. Drittens gibt es zunehmend Arbeiten zu Gemeinsamkeiten der kognitiven und neuronalen Korrelate des Diskursverständnisses in Film und Text (Zacks & Magliano 2011; Zacks et al. 2007; Kurby & Zacks 2008; Radvansky & Zacks 2017). Aus experimentellen Studien wissen wir, dass Leser wie Filmbetrachter mit dem Ziel des Diskursverständnisses Orte, Zeiten, Handelnde und Kausalbeziehungen eng verfolgen. Wenn sich zentrale Merkmale der Situation ändern, wird dies als Ereignisgrenze wahrgenommen und das aktuelle Ereignismodell muss aufgefrischt werden (Zacks et al. 2009). Solche Merkmale gleichen denen, die in der Kohäsionsanalyse im Mittelpunkt stehen.

Multimodale Kohäsion als Bestandteil filmischen Diskurses

Filmwissenschaftler sind sich einig darüber, dass im audiovisuellen Medium des Films Wiederholungen, Rekurrenzen und Ähnlichkeiten in der Form systematisch eingesetzt werden, um es dem Betrachter zu erleichtern, zu kohärenten Interpretationen des Materials zu gelangen und ihn emotional und ästhetisch zu involvieren. Obwohl die allgemeine Konzeption von Kohäsion als Hinweis darauf, wie Diskursentitäten eingeführt und verfolgt werden können, beibehalten wird, existieren beim Film im Vergleich zum Text eine weitaus größere Zahl möglicher Kommunikationsmittel. Diese erweiterte Form der Kohäsion wird in Tseng (2013) im Detail diskutiert; wir nutzen sie hier zur Analyse spezifischer Filmsequenzen.

In diesem Rahmenmodell werden kohäsive Mechanismen als Strategien angesehen, den Betrachter zu bestimmten Interpretationen zu führen beim Versuch, Ereignisse in audiovisuellen Medien zu verstehen. Hier zeigen wir, wie Kohäsionsanalyse es uns ermöglicht zu verstehen, wie technische, beobachtbare Merkmale eines Filmes die Interpretation anleiten, während sich dynamische Ereignisse entwickeln. Die kohäsive Analyse beschreibt, wie Bild, Ton, verbale Sprache, geschriebene Sprache, Kamerabewegung, *Framing*, Farbe und viele weitere Aspekte es bewirken, dass Per-

sonen, Orte und Objekte in einer gegebenen Ereignissequenz eingeführt und verfolgt werden können.

Gibt es systematische empirische Korrelate der abstrakten Kohäsionsanalyse? Mit anderen Worten, stehen die tatsächlichen Prozesse des Diskursverständnisses in einem systematischen Zusammenhang mit den Mustern der kohäsiven Analyse? Hier zeigen wir, dass wir aus der abstrakten kohäsiven Analyse spezifischer Filmsequenzen Elemente ableiten können, die für das weitere Verständnis des Narrativs von entscheidender Bedeutung sind. Wir untersuchen dies mit einem der Experimentalpsychologie entlehnten Ansatz der isolierten Bedingungsvariation. Wir manipulieren gezielt und subtil das Filmmaterial an Stellen, die gemäß der theoretischen Kohäsionsanalyse als besonders wichtig erscheinen, und beobachten die Effekte auf das Verständnis der Betrachter. Wir testen die Hypothesen mittels Fragebogendaten und Blickmessung (*Eye-Tracking*). Wir illustrieren dies am Beispiel einer kohäsiven Analyse der Anfangssequenz von Alfred Hitchcocks *The Birds* (Die Vögel, 1963), aus der wir theoriegeleitet Vorhersagen ableiten, die wir anschließend in Verhaltensexperimenten empirisch überprüfen.

Zu Beginn des Films wird die Hauptperson narrativ vom Hintergrund in den Vordergrund geholt. In der gesamten Außenszene hört man immer wieder kreischende Möwen. Der Film beginnt mit einer vorbeifahrenden Straßenbahn, die den Blick auf eine Gruppe von Menschen an einer belebten Kreuzung freigibt, welche darauf warten die Straße zu überqueren. Eine weibliche Person trennt sich von der Menge und die Kamera fokussiert auf sie, was sie visuell salient macht. Sie verschwindet hinter einem Poster der Golden Gate Bridge mit dem Schriftzug San Francisco und taucht dann wieder auf als Person im Vordergrund vor einer Tierhandlung. Man sieht, wie sie sich nach dem Geräusch der kreischenden Möwen am Himmel umdreht, woraufhin die Kamera auf den Möwenschwarm schwenkt. Nun wird wieder die Frau fokussiert, und man sieht, wie sie die Tierhandlung betritt (die zeitgleich Hitchcock verlässt). Schließlich wird sie in der Tierhandlung gezeigt, wie sie eine Treppe heraufsteigt.

Das Ergebnis der kohäsiven Analyse ist in Abbildung 1 dargestellt. Für die empirische Untersuchung ist besonders wichtig zu erkennen, dass schon frühzeitig, etwa in 6, der Hinweis auf die Tierhandlung als verbaler Hinweisreiz zu erkennen ist („Davidson's Pet Shop“). Dieser Hinweisreiz etabliert das Setting, nachdem die Handelnde durch die Eingangstür der Handlung gegangen ist und erleichtert es dem Betrachter, Kohäsion wahrzunehmen.

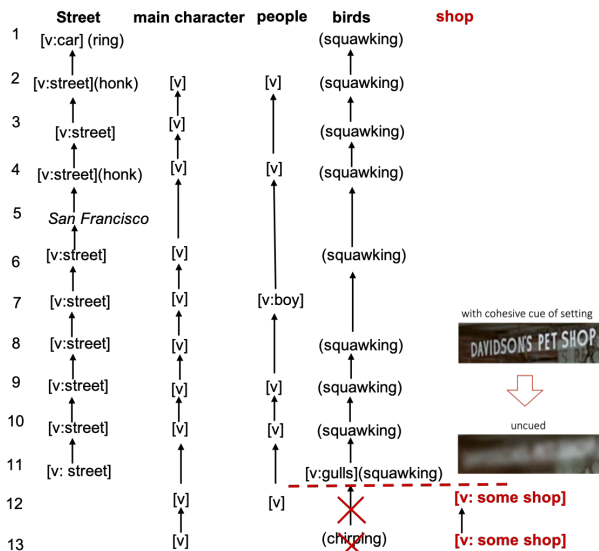


Abb. 1: Kohäsionsanalyse (schwarz) und Veränderungen in der manipulierten Version (rot) der Eingangssequenz von Hitchcock's *The Birds*.

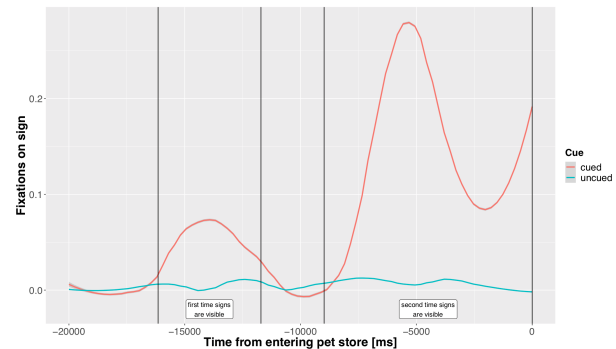


Abb. 2: Anzahl der Fixationen auf den Schildern in der Version mit und ohne Hinweisreiz (cued vs. uncued), relativ zum Eintreten in die Tierhandlung (Zeit = 0).

Empirische Überprüfung

Wenn wir nun Hinweisreize auf die Tierhandlung entfernen, sollte dem Betrachter die Interpretation des *Settings* schwerer fallen. Wir haben den Film subtil verändert, indem wir die Schilder durch Tiefpassfilterung unleserlich gemacht haben. Zusätzlich haben wir die Vogelgeräusche im Inneren der Tierhandlung durch generische sanfte Hintergrundmusik ersetzt. Die Sequenz innerhalb der Tierhandlung war somit visuell identisch für das Original und die manipulierte Version. Obwohl die Manipulation flüchtigen Betrachtern nicht auffiel, erwarteten wir durch die Zerstörung kohäsiver Ketten einen Effekt auf das narrative Verständnis der Szene innerhalb der Tierhandlung. Die roten Elemente in Abbildung 1 zeigen, dass das Entfernen der visuell-verbalen Hinweise auf die Tierhandlung das Setting von einer Tierhandlung zu einem generischen Geschäft änderte. In ähnlicher Weise sollte sich das Ersetzen des Vogelzwitscherns durch Aufzugsmusik auswirken. Dadurch dass der Betrachter weniger vorbereitet ist, erwarteten wir mehr Orientierungsverhalten.

In einer Fragebogenstudie erfragten wir unmittelbar im Anschluss an das Filmbetrachten das offene Wissen um das Setting: „Wohinein ging die handelnde Person?“. In zwei unterschiedlichen Stichproben in Bremen (n=45) und in Potsdam (n=74) gab es deutliche Unterschiede zwischen den Gruppen, die das Original und die manipulierte Version gesehen hatten ($p = 0.009$ bzw. $p = 0.033$). Jeweils war die Anzahl richtiger Antworten höher in der Originalversion.

In einer Blickbewegungsstudie haben wir zunächst geschaut, ob sich das Entfernen der kohäsiven Hinweisreize auf die Aufmerksamkeitsverteilung ausgewirkt hat. Abbildung 2 zeigt, dass dies eindeutig der Fall war. Wenn Schrift zu lesen war, war ein nennenswerter Anteil (18%) aller Fixationen auf die Schilder fokussiert, die also stark aufmerksamkeitslenkend wirkten. Wenn sie dagegen verschwommen dargestellt wurden, wurden sie kaum beachtet (1% der Fixationen). Dieser Gruppenunterschied war hoch signifikant, $p < 0.001$. (3% der Probanden mit Hinweisreiz, aber nur 15% der Probanden ohne Hinweisreiz fixierten die Schilder mindestens einmal.

In einem weiteren Schritt haben wir untersucht, ob das Fehlen kohäsiver Hinweise zur Etablierung des Tierhandlungs- *Settings* zu mehr Orientierungsverhalten bei der visuell identischen Szene *innerhalb* der Tierhandlung führten. Dazu haben wir ein Maß benutzt, das als Erweiterung der Standardabweichung auf zwei Dimensionen angesehen werden kann (die Quadratwurzel der Determinante der Kovarianzmatrix, Paindaveine 2008). Abbildung 3 zeigt wie sich die Streuung der Blickverteilung über die Zeit in der Tierhandlung entwickelt. In beiden Bedingungen steigt die Streuung anfangs und sinkt gegen Ende der Szene. Jedoch beginnt das Absinken in der Bedingung mit Hinweisreiz früher und es ist deutlich. Der Beginn des Absinkens scheint mit dem Moment übereinzustimmen, an dem die Handelnde ihren Kopf dreht. Wir spekulieren, dass das Kopfdrehen ein soziales Signal für den Betrachter ist, ihrer Aufmerksamkeit zu folgen, und dass dieser Cue in der Originalversion häufiger beachtet wurde, während Betrachter der manipulierten Version weiter das Geschäft explorierten. Über die gesamte Szene aufsummiert ist die Streuung der Verteilung eindeutig größer in der Bedingung ohne Hinweisreiz, obwohl sich die Verteilungen zu Beginn mehr oder weniger parallel entwickeln.

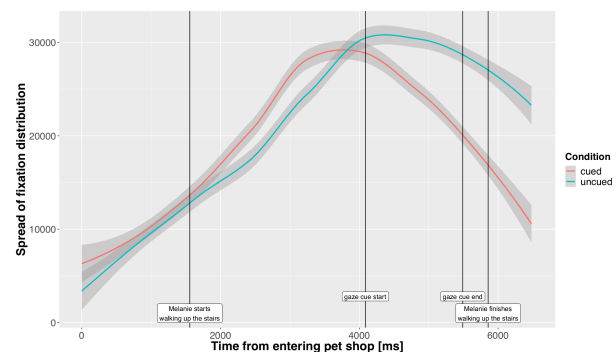


Abb. 3: Entwicklung der Streuung der Blickverteilung innerhalb des Geschäfts über die Zeit.

Während die Fragebogendaten zeigen, dass die Teilnehmer weniger wahrscheinlich den Ort identifizieren, wenn kohäsive Hinweise fehlen, legen die Blickbewegungsdaten nahe, dass Betrachter des Originals tatsächlich die "Pet Shop"-Schilder lesen und dazu nutzen, die Identität des Geschäfts zu etablieren, während Betrachter der manipulierten Version später innerhalb des Geschäftes aktiv nach Information suchten, um ihre Unsicherheit zu reduzieren. Zusammenfassend betrachtet haben hier verbale und auditive Hinweise die Aufmerksamkeit des Betrachters gelenkt und dadurch ihren narrativen Verstehensprozess beeinflusst.

Diskussion

Die Kohäsionseffekte waren relativ klein gemessen an dem Effekt der visuellen Manipulation. Dies hatten wir erwartet vor dem Hintergrund dass Aufmerksamkeit sowohl durch den Reiz als auch durch höhere Ziele gelenkt werden kann, und die im Film dominanten Bewegungssignale sehr stark reizgetriebene die Aufmerksamkeit aus sich ziehen. Tatsächlich ist diese reizgetriebene, bottom-up Steuerung der Aufmerksamkeit und des Blickes bei gleichzeitig nur schwer detektierbaren top-down-Effekten so ausgeprägt, dass sie als "tyranny of film" bezeichnet wurde (Loschky et al. 2015).

Auch in unserer Studie drückten sich solche reizgetriebenen Effekte aus in der Homogenität der Blickverteilung und vor allem auch im starken Effekt, den unsere Manipulation hinsichtlich der Betrachtung der manipulierten Reize selbst hervorrief (vgl. Abb. 2). Nichtsdestotrotz haben wir auch kleinere, zielgesteuerte top-down-Effekte auf die Aufmerksamkeitsverteilung nachweisen können, die durch Unterschiede in der Kohäsion verursachte Unterschiede im Verstehen der Szenesemantik widerspiegeln. Letztlich führten diese zu Unterschieden der Aufmerksamkeitsverteilung auf visuell identischen nachfolgenden Szenen.

Inwieweit liefert dieser Beitrag einen Mehrwert für die DH? Wir haben digitale Methoden verwendet bei der Bearbeitung der Filmausschnitte, der Messung der Blickdaten und bei der Erstellung von R-Skripten für Auswertungen und Visualisierungen. Jedoch haben wir hier keinen *Distant Viewing*-Ansatz mit rein informatischer Inhaltsanalyse, sondern eher einen *Close Viewing*-Ansatz verfolgt (im Sinne des *Close Reading*). Rein computerbasierte Inhaltsanalysen mit tiefen neuronalen Netzen, wie sie von uns etwa bei der stilometrischen Analyse graphischer Romane implementiert wurde, liefern sicher perspektivisch wichtige zusätzliche Erkenntnisse; sie sind aber derzeit noch nicht auf dem Niveau entwickelt, mit dem sich hypothesengeleitete semantisch-narrative Fragestellungen wie die unsrige mit vertretbarem Aufwand sinnvoll verfolgen ließen. Unsere Studie werten wir eher als eine erste explorative empirische Vorarbeit, die Fragestellungen für spätere, möglicherweise „digitalere“ Analysen aufzeigen kann.

In diesem Beitrag haben wir einen theoretischen Rahmen multimodaler Diskursanalyse verknüpft mit einer empirischen Überprüfung von daraus abgeleiteten Vorhersagen. Wir haben gezeigt, dass kohäsive Hinweise zum Verständnis des Films beitragen und dem Betrachter bei der Interpretation helfen. Wir wissen nun etwas besser, wie kohäsive Hinweise im Film funktionieren. Spannend für zukünftige Forschung wäre die Frage, ob sich daraus auch Konsequenzen für Filmschaffende ableiten lassen.

Bibliographie

- Bordwell, D.**, 2007, *Poetics of Cinema*, Routledge, London and New York.
- Flowerdew, J. and Mahlberg, M.**, eds (2009), *Lexical Cohesion and Corpus Linguistic*, John Benjamins, Amsterdam.
- Halliday, M. A. K. and Hasan, R.**, 1976, *Cohesion in English*, Longman, London.
- Hoffmann, C. R.**, 2012, *Cohesive Profiling: Meaning and Interaction in Personal Weblogs*, John Benjamins, Amsterdam.
- Kurby, C. A. and Zacks, J. M.**, 2008, 'Segmentation in the perception and memory of events', in: *Trends in Cognitive Sciences* 12, 72–79.
- Loschky, L. C., Larson, A. M., Magliano, J. P. and Smith, T. J.**, 2015, 'What would jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension', in: *PLOS ONE* 10(11), 1–23.
- Martin, J. R.**, 1992, *English text: Systems and structure*, Benjamins, Amsterdam.
- Paindaveine, D.**, 2008, 'A canonical definition of shape', in: *Statistics & Probability Letters* 78(14), 2240–2247.
- Radvansky, G. A. and Zacks, J. M.**, 2017, 'Event boundaries in memory and cognition', in: *Current Opinion in Behavioral Sciences* 17, 133–140.
- Tanskanen, S.-K.**, 2006, *Collaborating towards Coherence: Lexical Cohesion in English Discourse*, John Benjamins, Amsterdam.
- Tseng, C.**, 2013, *Cohesion in Film: Tracking Film Elements*, Palgrave Macmillan, Basingstoke.
- Zacks, J. M., Speer, N., Swallow, K., Braver, T. and Reynolds, J.**, 2007, 'Event perception: a mind/brain perspective', in: *Psychological Bulletin* 133, 273–293.
- Zacks, J. M. and Magliano, J. P.**, 2011, 'Film, Narrative and Cognitive Neuroscience', in: F. Bacci and D. P. Melcher, eds, *Art and the Senses*, Oxford University Press, Oxford and New York, pp. 435–454.

Erweiterungen der Digital Humanities durch kulturwissenschaftliche Perspektiven

Franken, Lina

lina.franken@soziologie.uni-muenchen.de
LMU München, Germany

Die Digital Humanities haben sich zu einer interdisziplinären „Transformationswissenschaft“ (Jannidis/Kohle/Rehbein 2017: XI) entwickelt, die aus zahlreichen Richtungen diskutiert und deren Perspektive immer wieder auch durch neue Herkunftsdisziplinen bereichert wird, die sich hin zur DH öffnen. Bisher fehlt allerdings die kulturwissenschaftliche Perspektive weitestgehend. Der vorliegende Beitrag möchte diese Leerstelle beleuchten und fragt danach, warum die Kulturwissenschaften in den Digital Humanities aktuell noch kaum vertreten sind. Er geht außerdem der Frage nach, welche Spezifika die Kulturwissenschaften in die DH einbringen und wie die entstehende Transformationswissenschaft

im Sinne einer „Big Digital Humanites“ (Svensson 2016) durch eben jene Perspektiven bereichert werden kann.

Kulturwissenschaften im Allgemeinen, und die Empirische oder Vergleichende Kulturwissenschaft im Speziellen, fragen nach Bedeutungen von kulturellen Äußerungen, wobei Kultur mit einem weiten Begriff als Alltagskultur oder als „whole way of life“ (Williams 1960) verstanden wird. Dabei gibt es vor allem aus methodischer Perspektive ein Alleinstellungsmerkmal im Vergleich zu anderen Disziplinen, die sich den DH bereits geöffnet haben: In den Kulturwissenschaften werden Daten bearbeitet, die zum Teil im Forschungsprozess entstehen und ihre Relevanz erst während der Analyse zeigen. Damit ist die Datenauswahl und -erhebung grundsätzlich anders gestaltet als in Bereichen der DH, die bestehende Korpora auswählen, aufbereiten und analysieren. Literaturwissenschaftliche, geschichtswissenschaftliche oder kunsthistorische Korpora etwa liegen bei Entwicklung der jeweiligen Fragestellung bereits vor und müssen zwar ausgewählt, nicht aber erst erzeugt werden. Kulturwissenschaftliche Korpora sind flexibler, sie entstehen insbesondere durch empirische Erhebungen und werden in der Analyse beständig erweitert und gefiltert (Koch/Franken 2019). Über den konkreten Forschungsgegenstand ist im Vorfeld in der Regel verhältnismäßig wenig bekannt, die Expertise steigt im Laufe der Erhebung und durch die Beschäftigung mit den oft emergenten Phänomenen in ihren Kontexten. Damit steigt auch das Wissen darum, welche Teilkorpora für die weitere Analyse relevant sind oder doch sein könnten, iterativ. Selbstverständlich ist dies auch für andere an den DH beteiligte Disziplinen zu konstatieren, finden sich doch auch hier relevante Ausschnitte aus den Korpora erst in der Auseinandersetzung mit eben diesen. In empirischen Zugängen zu Welt, etwa auf Grundlage von Methodologien wie der Grounded Theory (Glaser/Strauss 1967) und im feldforscherischen oder ethnografischen Zugriff (Hess/Schwertl 2013; Faubion 2009) mit einem weiten Kulturbegriff (Reckwitz 2000; Geertz 1983) arbeitend, wird das jeweilige kulturwissenschaftliche Forschungsfeld erst in der Analyse näher erschlossen. Hier wird immer wieder auch neu entschieden, einzelne Daten mit einzubeziehen, da sie für die präzisierte Fragestellung interessant werden. Datenmaterial wird dabei sowohl forschungsinduziert erzeugt als auch in Form von prozessproduzierten Daten nachgeutzt (zur Unterscheidung Baur/Graeff 2020).

Damit einher geht eine komplexe Datengrundlage, die vielfältig geschachtelt ist: stehen etwa am Anfang einzelne Texte oder erste Interviews in ihren Transkriptionen im Mittelpunkt, so kann je nach Zuschnitt der Fragestellung später der Fokus auf Text-Bild-Kompositionen, eigenen Beobachtungen und deren Verschriftlichung, auf Dokumentationen mit Fotos oder Videos liegen oder sich im Bereich der digitalen Ethnographie hin zu Äußerungen in den sozialen Medien oder in Blogs erweitern (zur letzteren jüngsten Methodenentwicklung Hine 2015; Pink et al. 2016). Eben weil Erkenntnisse aufeinander aufbauen, werden Datensätze erst im Laufe der Erhebung und Analyse als relevant identifiziert und iterativ ergänzt. Viele andere DH-Forschungen hingegen betrachten das eigene Korpus relativ zu Beginn des Forschungsprozesses als abgeschlossen und nehmen nur noch in Ausnahmefällen weiteres Quellenmaterial hinzu. Die Kulturwissenschaften bringen damit eine spezifisch multimodale Datengrundlage in die Digital Humanities ein, die nicht nur verschiedene Standards je nach Datentyp notwendig macht, sondern auch die Kombination von unterschiedlichen analytischen Ansätzen. Denn das jeweils individuell wachsende Datenmaterial richtet sich nicht nach Datentypen oder möglichen Verfahren, sondern nach den mit diesen ermöglichten Erkenntnissen. Damit multiplizieren sich auch die Ansprüche an Forschende, die sich aus den Kulturwissenschaften kommend in die DH einbringen wollen: eine ganze Vielfalt an Möglichkeiten

eröffnet sich, gleichzeitig sind nicht alle Ansätze für die spezifischen Korpora dabei gewinnbringend und sie müssen stets kombiniert werden, um dem Datenmaterial gerecht zu werden.

Erst in einigen wenigen Ansätzen werden die dargestellten kulturwissenschaftlichen Forschungsperspektiven bisher mit den Digital Humanities in Verbindung gebracht. So entwickelten Hoffmeister, Marguin und Schendzielorz (2018) einen Ansatz der Aufbereitung von Feldnotizen für digitale Analysen, der bisher jedoch nicht in ein nutzbares Tool übersetzt wurde. Auch konzeptionelle Überlegungen zum Umgang mit der spezifischen multimodalen Datengrundlage bestehen (etwa Wiedemann 2016), wurden jedoch bisher erst selten in Forschungspraxis umgesetzt. Vorarbeiten der Autorin bleiben bisher ebenfalls auf der konzeptionellen Ebene (Franken 2020a) sowie ersten Auslotungen von Möglichkeiten (etwa Adelman et al. 2019; Adelman/Franken 2020) und arbeiteten in einer empirischen Untersuchung heraus, dass kulturwissenschaftliche Forschende aktuell vor allem generische Tools verwenden und wenig mit den Verfahren der DH in Berührung kommen (Franken 2020b). International sind die Ansätze bereits deutlicher konturiert, insbesondere im Bereich der Cultural Analytics (Manovich 2020) und auch in den Digital Folkloristics (Tolbert/Johnson 2019), folgen jedoch nicht immer dem ethnografischen Paradigma innerhalb der Kulturwissenschaften, sondern sind wiederum stärker an historischen und auch medienwissenschaftlichen Perspektiven orientiert. Gleichzeitig bilden diese in den kulturwissenschaftlichen Forschungen aktuell eine absolute Ausnahme, wie es in so vielen Disziplinen zu Beginn der Etablierung von DH-Ansätzen war und ist.

Durch ihre spezifische Perspektive und das empirische Vorgehen in eigener Forschung stehen die Kulturwissenschaften an der Schnittstelle zu den Computational Social Sciences (CSS), die ihren Fokus eindeutig auf die Gegenwart legen, in vielen Bereichen vor allem Daten aus den sozialen Medien analysieren sowie Simulationen und Modelle erstellen sowie deutlich aus der quantitativen Sozialforschung heraus argumentieren (Salganik 2018; Stützer/Welker/Egger 2018). Doch kulturwissenschaftliche Forschung arbeitet einerseits qualitativ, andererseits versteht sie Gegenwart als geworden und fragt in ihrer Perspektive immer auch nach historischen Dimensionen (Lipp 2013; Hirschfelder 2012). Gerade die Verbindung von gegenwartsbezogenen und historischen Perspektiven prägt ihre Forschungen: Es besteht hier ein Dazwischen, welches die (empirische) Kulturwissenschaft zwischen gegenwartsbezogen und historisch arbeitenden Disziplinen vertritt. Je nach Forschungsfrage spielt für die Analyse der Gegenwart die historische Genese eine große Rolle: warum etwa heutiges immaterielles Kulturerbe vom Brot bis zum Oktoberfest in seinen Ausdrucksformen gestaltet und gelebt wird (Kirschenblatt-Gimblett 2014; Bendix 2007; Tauschek 2013), ist ohne Geschichte nicht zu erklären – aber auch noch ohne Gegenwart.

Kulturwissenschaften sind in ihren Fragestellungen und in ihren Datengrundlagen in der Summe deutlicher den DH als den CSS zuzuordnen, steht jedoch an der Grenze. Denn forschungsinduzierte Quellen legen zwar den CSS-Bezug nahe, dort wird die DH-spezifische historische Dimensionierung jedoch ausgeblendet. Mehr noch: Kulturwissenschaften können zentral dazu beitragen, die zahlreichen vorhandenen Bezüge zu stärken, die zwischen CSS und DH bestehen, aktuell jedoch kaum als Synergien genutzt werden. Die DH könnten damit einmal mehr als „Brückenfach“ (Sahle 2016) verstanden werden, um die bestehenden Entwicklungen zusammen zu bringen. Als Schnittstelle zwischen beiden Ansätzen bieten sich die Kulturwissenschaften an, um Gemeinsamkeiten zu identifizieren und seitens der DH in einen längst fälligen engeren Dialog mit den empirisch arbeitenden Disziplinen zu treten, die sich unter dem Dach der CSS sammeln:

Beispielsweise die Politik- und Medienwissenschaften, aber auch die Geographie und Soziologie steigen jüngst in die Nutzung und Weiterentwicklung von digitalen Verfahren ein, etablieren entsprechende Studiengänge und Professuren. Der Austausch gerade innerhalb der deutschsprachigen Community der Geistes- und Sozialwissenschaften hin zur Nutzung von Informationstechnologien ist durch die Scharnierfunktionen der Kulturwissenschaften deutlich zu stärken.

Schließlich haben die Kulturwissenschaften die Entwicklungen selbst zum Forschungsgegenstand, welche DH und CSS erst ermöglichen: die digitale Durchdringung unserer Alltage und damit auch Forschungsalltage. Diese nicht nur zu beschreiben, sondern mit kritischen Perspektiven und Reflexionen zu begleiten, ist zentrale Aufgabe der Kulturwissenschaften (Fortun et al. 2014; Beck 2019). Insbesondere die Critical Code Studies (Marino 2020; Introna 2016) ebenso wie die Critical Data Studies (Kitchen/Lauriault 2018; boyd/Crawford 2012) werden in den DH bisher kaum wahrgenommen. Daten werden hier als kontextabhängig und prozesshaft verstanden, als nicht neutral, sondern „broken“ oder „messy“ (Pink et al. 2018). Sie sind nie roh, es gibt keine unbearbeiteten Daten (Bowker 2014). Computercode und Algorithmen werden als eingebunden in Handlungsweisen verstanden und sind immer in Standards und Konventionen eingebunden, damit als kulturelles Artefakt zu interpretieren (Mackenzie 2005). Diese Perspektiven sind gewinnbringend für die Reflexion dessen, wie sich Forschungsprozesse verändern, wenn sie digital erweitert werden. Kulturwissenschaftliche Forschungen stellen entsprechende Fragen in den Mittelpunkt und können damit das konkrete Nutzen von DH-Zugängen mit deren Infragestellung verbinden. Eine epistemologische Weiterentwicklung der DH kann somit noch einmal ganz anders dimensioniert werden: Welche Rolle spielen die verwendeten Skripte und Tools für unsere Forschungsprozesse? Wie sind Codestrukturen und Datensätze in ihren materiellen wie immateriellen Dimensionen mit Forschenden verstrickt und wo bestehen wechselseitige Abhängigkeiten? Technik, und damit eben auch Daten und Code, muss eine eigene Handlungsmacht zugeschrieben werden: Sie ist nicht passiv durch menschliche Akteure verwendet, sondern prägt Handlungen, ermöglicht und begrenzt sie. Es reicht also nicht aus, nur die Technik selbst, in unserem Fall die Tools und Verfahren, anzuschauen, sondern diese muss in soziale Relationen eingeordnet verstanden werden.

Durch eine stärkere Verankerung der Kulturwissenschaften in den DH können sich DH und Kulturwissenschaften wechselseitig gewinnbringend erweitern. Eine entsprechende Etablierung rückt näher: In den deutschsprachigen ethnologischen Fachverbänden hat die Diskussion von und Bezugnahme auf die Digital Humanities in den letzten Jahren ebenso zugenommen wie die Referenz auf Prinzipien der Open Science, die Digitalisierung und Erschließung von Archivmaterialien oder die Hinwendung zu digitalen Plattformen zur Analyse und Veröffentlichung von Forschungsergebnissen. Es ist anzunehmen, dass sich in naher Zukunft weitere Fachwissenschaftler:innen gerade auch der Nachwuchsgeneration hin zu den DH öffnen und ihre Expertise, aber auch ihre spezifischen Bedürfnisse und Problematisierungen einbringen. Konkrete methodische Perspektiven der DH sind nicht immer auf diese ausgerichtet, die Lücke von Struktur und Bedeutung ist oft noch sehr groß. Deshalb bieten digitale Methoden aktuell insbesondere als Vorarbeit für die eigentliche Interpretationsleistung Potential. Das computationelle Vorgehen führt zu granularen Perspektiven, die zusammengefügt werden müssen. Das Kontextwissen der Forschenden wächst damit in seiner Bedeutung ebenso wie das Wissen um die Begrenztheit der computationellen Verfahren, die mit anderen Zugängen zum Datenmaterial ergänzt werden müssen.

Für die DH ist dies eine Chance nicht nur für eine weitere Öffnung hin zu multimodalen Datengrundlagen und kombinierenden Verfahren, sondern auch für eine stärkere Verzahnung mit den CSS und damit einer Bündelung der Kompetenzen. Die Herausforderungen und Chancen sind hier vergleichbar: es wird in beiden Bereichen herausgearbeitet, welche neuen und alten Quellengattungen mit welchen neuen, digitalen Verfahren bearbeitet werden können. Dabei liegt ein besonderes Potential der kulturwissenschaftlichen Perspektive in der stärkeren Einbindung auch qualitativer Forschungsmethoden, die sich mit quantitativen Ansätzen ganz im Sinne klassischer Mixed-Methods-Ansätze (Kelle 2019) digital erweitert hin zu neuen Zugriffen auf Welt entwickeln können. Zudem kann die Erweiterung der Digital Humanities um kulturwissenschaftliche Perspektiven in der Zukunft ein noch stärkeres Reflexionsbewusstsein in der Community schaffen, das neben dem konkreten Anwenden und Weiterentwickeln von Verfahren zentral für die wissenschaftstheoretische ebenso wie die wissenschaftspolitische Verankerung der Transformationswissenschaft DH ist und bleibt.

Bibliographie

- Adelmann, Benedikt; Franken, Lina** (2020): "Thematic Web Crawling and Scraping as a Way to form Focussed Web Archives". In: Sharon Healy, Michael Kurzmeier, Helena La Pina und Patricia Duffe (Hg.): *Book of Abstracts: #EWAVirtual 2020*, S. 35–37.
- Adelmann, Benedikt; Franken, Lina; Gius, Evelyn; Krüger, Katharina; Vauth, Michael** (2019): "Die Generierung von Wortfeldern und ihre Nutzung als Findeheuristik. Ein Erfahrungsbericht zum Wortfeld 'medizinisches Personal'". In: Patrick Sahle (Hg.): *6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019). Digital Humanities: multimedial & multimodal*. Universitäten zu Mainz und Frankfurt 25. bis 29. März 2019. Book of Abstracts. Frankfurt a.M., S. 114–116.
- Baur, Nina; Graeff, Peter** (2020): "Datenqualität und Selektivitäten digitaler Daten. Alte und neue digitale und analoge Daten-sorten im Vergleich." Unveröffentlichter Vortrag im Rahmen des 40. Kongresses der Deutschen Gesellschaft für Soziologie.
- Beck, Stefan** (2019 [2015]): "Von Praxistheorie 1.0 zu 3.0. Oder: wie analoge und digitale Praxen relationiert werden sollten". In: *Berliner Blätter. Ethnographische und ethnologische Beiträge* (81), S. 9–27.
- Bendix, Regina** (2007): "Kulturelles Erbe zwischen Wirtschaft und Politik. Ein Ausblick". In: Dorothee Hemme, Markus Tauschek und Regina Bendix (Hg.): *Prädikat Heritage. Wertschöpfung aus kulturellen Ressourcen*. Berlin, S. 337–356.
- Bowker, Geoffrey C.** (2013): "Data Flakes. An Afterword to 'Raw Data' is an Oxymoron". In: Lisa Gitelman (Hg.): *"Raw Data" is an Oxymoron*. Cambridge, Massachusetts, S. 167–171.
- boyd, danah; Crawford, Kate** (2012): "Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon". In: *Information, Communication & Society* 15 (5), S. 662–679. DOI: 10.1080/1369118X.2012.678878.
- Faubion, James D.; Marcus, George E. (Hg.)** (2009): *Fieldwork is not what it used to be. Learning Anthropology's Method in a Time of Transition*. Ithaca.
- Fortun, Kim et al.** (2014): "Experimental Ethnography Online". In: *Cultural Studies* 28 (4), S. 632–642. DOI: 10.1080/09502386.2014.888923.
- Franken, Lina** (2020a): "Methodologie der Zukunft? Automatisierungspotentiale in kulturwissenschaftlicher Forschung". In:

Dagmar Hänel et al. (Hg.): *Planen. Hoffen. Fürchten. Zur Gegenwart der Zukunft im Alltag*. Münster/New York, S. 217–233.

Franken, Lina (2020b): "Kulturwissenschaftliches digitales Arbeiten. Qualitative Forschung als 'digitale Handarbeit'?" In: *Berliner Blätter. Ethnographische und ethnologische Beiträge* 82, S. 107–118. Online verfügbar unter <https://www2.hu-berlin.de/ifeeojs/index.php/blaetter/article/view/1069/16>.

Geertz, Clifford (1983): *Dichte Beschreibung. Beiträge zum Verstehen kultureller Systeme*. Frankfurt a.M.

Glaser, Barney G.; Strauss, Anselm L. (2010 [1967]): *Grounded Theory. Strategien qualitativer Forschung*. Bern.

Hess, Sabine; Schwertl, Maria (2013): "Vom 'Feld' zur 'Assemblage'? Perspektiven europäisch-ethnologischer Methodenentwicklung - eine Hinleitung". In: Sabine Hess, Johannes Moser und Maria Schwertl (Hg.): *Europäisch-ethnologisches Forschen. Neue Methoden und Konzepte*. Berlin, S. 13–37.

Hine, Christine (2015): *Ethnography for the Internet. Embedded, Embodied and Everyday*. London et al.

Hirschfelder, Gunther (2012): "Europäischer Alltag im Fokus der Kulturanthropologie/Volkskunde". In: Stephan Conermann (Hg.): *Was ist Kulturwissenschaft? Zehn Antworten aus den „Kleinen Fächern“*. Bielefeld, S. 135–173.

Hoffmeister, Anouk; Marguin, Séverine; Schendzielorz, Cornelia (2018): "Feldnotizen 2.0. Über Digitalität in der ethnografischen Beobachtungspraxis". In: *Zeitschrift für digitale Geisteswissenschaften Sonderband 3*. DOI: 10.17175/SB003_007.

Introna, Lucas D. (2016): "Algorithms, Governance, and Governmentality". In: *Science, Technology, & Human Values* 41 (1), S. 17–49. DOI: 10.1177/0162243915587360.

Jannidis, Fotis; Kohle, Hubertus; Rehbein, Malte (Hg.) (2017): *Digital Humanities. Eine Einführung*. Stuttgart.

Kelle, Udo (2019): "Mixed Methods". In: Nina Baur und Jörg Blasius (Hg.): *Handbuch Methoden der empirischen Sozialforschung*. Wiesbaden, S. 159–172.

Kirschenblatt-Gimblett, Barbara (2014): "Intangible Heritage as Metacultural Production". In: *Museum International* 66, S. 163–174.

Kitchin, Rob; Lauriault, Tracey P. (2018): "Toward Critical Data Studies. Charting and Unpacking Data Assemblages and Their Work". In: Jim Thatcher, Andrew Shears und Josef Eckert (Hg.): *Thinking Big Data in Geography. New Regimes, New Research*. Lincoln/London, S. 3–20.

Koch, Gertraud; Franken, Lina (2019): "Automatisierungspotenziale in der qualitativen Diskursanalyse. Das Prinzip des 'Filtorns'". In: Patrick Sahle (Hg.): *6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019). Digital Humanities: multimedial & multimodal*. Universitäten zu Mainz und Frankfurt 25. bis 29. März 2019. Book of Abstracts. Frankfurt a.M., S. 89–91.

Lipp, Carola (2013): "Perspektiven der historischen Forschung und Probleme der kulturhistorischen Hermeneutik". In: Sabine Hess, Johannes Moser und Maria Schwertl (Hg.): *Europäisch-ethnologisches Forschen. Neue Methoden und Konzepte*. Berlin, S. 205–246.

Mackenzie, Adrian (2005): "The Performativity of Code. Software and Cultures of Circulation". In: *Theory, Culture & Society* 22 (1), S. 71–92. DOI: 10.1177/0263276405048436.

Manovich, Lev (2020): *Cultural Analytics*. Cambridge.

Marino, Mark C. (2020): *Critical Code Studies*. Cambridge.

Pink, Sarah et al. (2016): *Digital Ethnography. Principles and Practice*. Los Angeles et al.

Pink, Sarah et al. (2018): "Broken Data. Conceptualising Data in an Emerging World". In: *Big Data & Society* 5 (1), DOI: 10.1177/2053951717753228.

Reckwitz, Andreas (2000): *Die Transformation der Kulturtheorien. Zur Entwicklung eines Theorieprogramms*. Weilerswist.

Sahle, Patrick (2016): "Digital Humanities? Gibt's doch gar nicht!". In: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/SB001_004.

Salganik, Matthew J. (2018): *Bit by Bit. Social Research in the Digital Age*. Princeton.

Stützer, Cathleen; Welker, Martin; Egger, Marc (Hg.) (2018): *Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications*. Köln.

Svensson, Patrik (2016): *Big Digital Humanities*. Michigan.

Tauschek, Markus (2013): *Kulturerbe. Eine Einführung*. Berlin.

Tolbert, Jeffrey A.; Johnson, Eric D. M. (2019): "Digital Folkloristics". In: *Western Folklore* 78, S. 327–356.

Wiedemann, Gregor (2016): *Text Mining for Qualitative Data Analysis in the Social Sciences. A Study on Democratic Discourse in Germany*. Wiesbaden.

Williams, Raymond (1960): *Culture and Society 1780–1950*. New York.

Evaluating Hyperparameter Alpha of LDA Topic Modeling

Du, Keli

duk@uni-trier.de

Universität Trier, Germany

Introduction

As a quantitative text analysis method, Latent Dirichlet Allocation (LDA), also often referred to as topic modeling (Blei 2012), has been widely used in Digital Humanities in recent years to explore numerous unstructured text data. When topic modeling is used, one has to deal with many parameters that could influence the result, such as the hyperparameter Alpha and Beta, the topic number, document length, or the number of iterations when updating the model. To understand the impact of these parameters, they must be systematically evaluated. In the last few years, there have been several studies evaluating LDA topic modeling in Digital Humanities or Computational Literary Studies (e.g., Jockers 2013; Schöch 2017; Du 2020; Uglanova & Gius 2020) and the presented paper focuses on evaluating the impact of hyperparameter Alpha on LDA topic models.

Hyperparameter Alpha can refer to two different types of parameters in the context of LDA topic modeling: LDA model parameter and inference algorithm parameter. As a parameter of the LDA model, Alpha determines the properties of a Dirichlet distribution, which is the prior probability distribution of the topic-document distribution. Together, the hyperparameter Alpha and the prior probability distribution determine which topics we expect to occur more frequently in the corpus and how confident we are about them. In practice, when we employ Gibbs Sampling to train our topic model, Alpha is the parameter, which has the smoothing effect on the topic-document distribution and ensures that the pro-

bability of each topic in each document is not 0 throughout the entire inference procedure. More importantly, Alpha represents the assumption about the data on how topics are distributed in documents before inferencing the topic model. In other words, the hyperparameter Alpha affects how often each topic occurs in each document. When the alpha value of a topic is set larger in a document, it means that the topic has a greater chance of appearing in that document. And vice versa. For this reason, the setting of Alpha can affect the quality of the inferred topic model. Therefore, this paper focuses on evaluating the impact of inference algorithm parameter alpha systematically.

According to Griffiths & Steyvers (2004), the topic model has the best quality when the sum of Alphas of all topics is equal to 50. This is probably the reason that in MALLET 2.0.7, the default value of the sum of Alphas was set to 50, while in MALLET 2.0.8, the value is reduced to 5. According to the supervisor of MALLET, David Mimno: “The general experience was that 50 was too large, and that 5 is a better default.”¹ Since there are different opinions on this issue, it is interesting to test how Alpha affects LDA topic modeling, especially on different types of text collections that are not in English. Therefore, this paper presents a study on evaluating Alpha on two German text collections and aims to understand the influence of hyperparameter Alpha from two perspectives: topic modeling based single-label document classification and topic coherence, representing the quality of the topic model and the quality of the topics, respectively.

Method

Two collections of German texts were built for the study. The first corpus is a collection of 2000 newspaper articles published between 2001 and 2014. The articles belong to ten different thematic classes, and each class contains 200 articles. The ten classes are “Digital”, “Society”, “Career”, “Culture”, “Lifestyle”, “Politics”, “Travel”, “Sports”, “Study” and “Economy”. The corpus contains over 3.4 million words in total, and the average text length is about 1800 words. The second corpus consists of 439 dime novels published between 1961 and 2016, and they belong to five subgenres, namely 100 fantasy novels, 51 horror novels, 88 crime novels, 100 romance novels, and 100 science fiction stories. The corpus contains about 13.4 million words, and the average text length is about 30,000 words. All texts are lemmatized. Since the average document length of the newspaper articles is 1800 words, the novels are also split into 1800 words segments. Thus, the document length is no longer a confounding factor when comparing the test results on the newspaper corpus with the results on the novel corpus.

The goal of the following tests is to explore the influence of the hyperparameter Alpha. While training topic models, the setting is varied by the value of Alpha $\in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 20, 30, 40, 50, 100\}$ and number of topics $\in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500\}$. For all other parameter settings, the default values of the topic modeling software were taken. All models were trained without applying hyperparameter optimization, which means that if Alpha is set to 0.1, the Alpha value for each topic is set to 0.1 during the whole training process. Common stop words were removed from both corpora. For technical reasons, namely random initialization in the topic assignment and Gibbs sampling, two topic models from one corpus are not completely identical even if the parameter settings during training are the same. Therefore, ten models were trained for each setting to balance the randomness from the technical side.

The topic models were trained using MALLET (McCallum 2002). As a result, a topic-document distribution and the topics are obtained for each topic model. In a topic-document distribution, each document is represented by an N-dimensional vector, while N is the number of topics of the topic model. Based on the topic-document distribution, the document classification was performed, and the classification was done as a 10-fold cross-validation with a linear SVM classifier. For the newspaper corpus, the articles were classified according to their thematic classes. For the novel corpus, the novel segments were classified according to their subgenre. The topic coherence was automatically calculated by the Java program Palmetto (Röder et al. 2015), and the first ten most important words of each topic were taken for the calculation. The reference corpus for the calculation of the topic coherence is the lemmatized German Wikipedia. Several topic coherence measures have been implemented in Palmetto. For this work, the Normalized Pointwise Mutual Information (NPMI) based coherence measure proposed in Aletas & Stevenson (2013) was taken. The theoretical range of NPMI based coherence measure is between -1 and 1. The higher the score, the better the topic.

By performing Bag-of-Words (BoW) model-based classification, a baseline of document classification has been defined for both corpora. The tests were also done as a 10-fold cross-validation with a linear SVM classifier. The F1(macro) score for the newspaper articles and for the novel segments was 0.758 and 0.993, respectively. A baseline of the NPMI value was also defined for each corpus. With only one iteration, 14 topic models were first trained on each corpus, containing 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 topics, respectively. In this way, 1,950 “topics before topic modeling” have been trained for each corpus. The NPMI scores of these topics were then calculated, and the average NPMI score is the NPMI baseline, which is -0.0619 for the newspaper corpus and -0.1153 for the novel corpus. A black line represents the baselines in Figure 3 and Figure 4.

Results

Document classification : Figure 1 and Figure 2 show the classification results based on topic models of newspaper articles and novel segments, respectively. It can be seen in both figures that the classification results gradually become worse with the increase of the setting of Alpha, regardless of how many topics have been trained. Especially if Alpha is set to greater than 1, the classification results based on topic models with more topics (the blue lines) show a stronger decreasing trend than the results based on topic models with fewer topics (the red lines). In comparison, most F1 scores change less when Alpha is set to a value smaller than 1. However, we can still see that the blue lines start to decrease when the Alpha is raised from 0.5 to 1. The highest F1-score of classifying newspaper articles and novel segments in this test are 0.752 and 0.998, respectively, which do not differ much from the pre-defined baseline based on BoW-model.

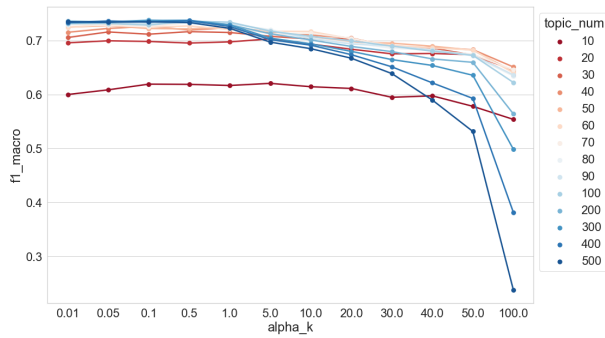


Fig. 1: Average F1(macro)-scores of topic modeling based classification of newspaper articles

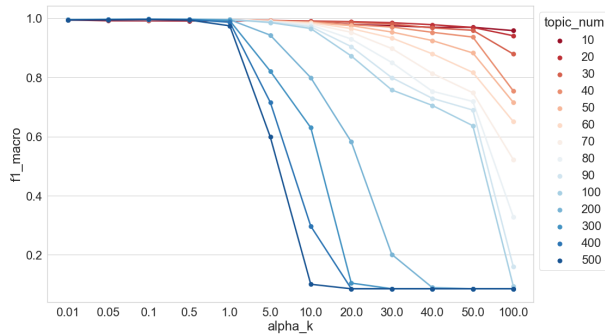


Fig. 2: Average F1(macro)-scores of topic modeling based classification of novel segments

Topic coherence : Compared to the classification results, the evaluation from the perspective of topic coherence shows some differences between the two corpora. Firstly, it can be observed in Figure 3 that the maximum of the NPMI-score distributions decreases with the increase of Alpha from almost 0.3 to about 0.12. In addition, the median of the distributions also shows a decreasing trend. At Alpha = 0.01, the median is lower than the NPMI baseline if the number of topics is set higher than 100. However, at Alpha = 100, the median is already lower than the NPMI baseline if the number of topics is set to 70. Apart from that, we can observe that the topic models with a higher number of topics contain more topics with low NPMI scores, regardless of the setting of Alpha. Compared to the test on the newspaper corpus, the test results on the novel corpus are slightly different. When Alpha is set smaller than 1, the NPMI-score distributions do not show an evident change as Alpha increases, and the range of distribution is often broader when the number of topics is set between 60 and 300. Starting from Alpha being raised to greater than 1, the distributions of the NPMI-scores clearly change, and the results then are similar to the previous test on the newspaper corpus: the maximum of the NPMI-score distributions decreases with the increase of the Alpha, and topic models with a higher number of topics contain more topics with low NPMI scores.

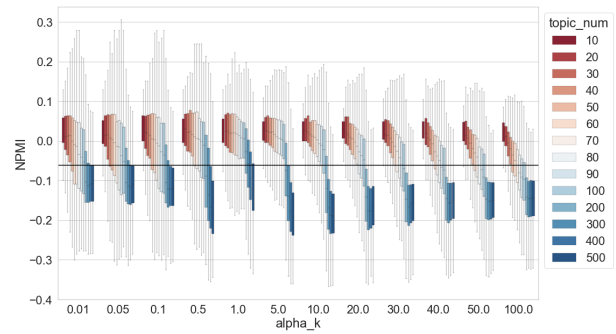


Fig. 3: NPMI-score distributions of topics from newspaper articles

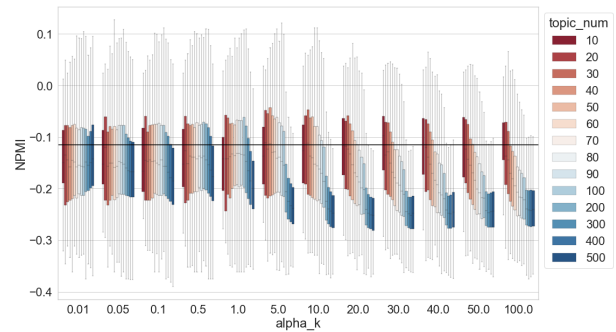


Fig. 4: NPMI-score distributions of topics from novel segments

Conclusion

The presented research evaluates the influence of hyperparameter Alpha in topic modeling on a German newspaper corpus and a German literary text corpus from two perspectives, single-label document classification, and topic coherence. Based on the results of the presented investigation, it can be stated that one should avoid training topic models with a setting of the Alpha of each topic to greater than 1 in order to ensure better topic modeling based document classification results and to get more coherent topics. In addition to that, LDA topic models with many topics are more vulnerable to changes in Alpha. Therefore, with the result of the presented investigations in this study, one can confirm the explanation of Mimno mentioned earlier that a smaller Alpha is better suitable for LDA Topic Modeling.

Footnotes

1. <https://stackoverflow.com/questions/45162186/mallet-topic-modeling-topic-keys-output-parameter> (15.07.2021)

Bibliography

- Aletras, Nikolaos / Stevenson, Mark (2013): "Evaluating topic coherence using distributional semantics". In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* (pp. 13-22).

Blei, David M. (2012): "Probabilistic topic models", in: *Communications of the ACM*, 55(4), 77-84.

Du, Keli (2020): „Der Spielraum zwischen "zu wenig" und "zu viel"". Presented at the DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2020), Paderborn: Zenodo. <http://doi.org/10.5281/zenodo.4621770>.

Griffiths, Thomas L. / Steyvers, Mark (2004): "Finding scientific topics", in: *Proceedings of the National Academy of Sciences*, 101 (Supplement 1), 5228–5235. <https://doi.org/10.1073/pnas.0307752101>.

Jockers, Matthew L. (2013): *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit*. <http://mallet.cs.umass.edu>.

Röder, Michael / Both, Andreas / Hinneburg, Alexander (2015): "Exploring the space of topic coherence measures", in: *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Schöch, Christof (2017): "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.", in: *Digital Humanities Quarterly* 11, no. 2. §1-53. <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>.

Uglanova, Inna / Gius, Evelyn (2020): "The Order of Things. A Study on Topic Modelling of Literary Texts", in: *Online Workshop on Computational Humanities Research, Proceedings*. <http://ceur-ws.org/Vol-2723/long7.pdf>.

Evaluation computergestützter Verfahren der Emotionsklassifikation für deutschsprachige Dramen um 1800

Schmidt, Thomas

thomas.schmidt@ur.de

Lehrstuhl für Medieninformatik, Universität Regensburg

Dennerlein, Katrin

katrin.dennerlein@uni-wuerzburg.de

Institut für Deutsche Philologie, JMU Würzburg

Wolff, Christian

christian.wolff@ur.de

Lehrstuhl für Medieninformatik, Universität Regensburg

Einleitung

Transformerbasierte Sprachmodelle wie BERT (Devlin et al. 2018) und ELECTRA (Clark et al. 2020) gelten als state-of-the-art und Ausgangspunkt für zahlreiche Aufgaben des Natural Language Processing (NLP) (Shmueli / Ku 2019; Munikar et al. 2019; Cao et al. 2020; Dang et al. 2020; Gonzáles-Carvajal et al. 2021; Cortiz 2021). Als ein entscheidender Vorteil dieser Modelle hat sich die dynamische Repräsentation von Tokens in Abhängigkeit von ihrem Kontext herausgestellt. Der Großteil dieser Modelle wird jedoch mit zeitgenössischer Sprache, vor allem mit Sach- und Fachtexten aus dem Web (z.B. *Wikipedia*) trainiert. Dies stellt ein Problem für Forschungsbereiche wie die Digital Humanities (DH) dar, die mit literarischen Texten arbeiten. Literarische Texte unterscheiden sich entscheidend von Textsorten wie Wikipedia-Artikeln, weil sie fiktional sind und Sprache kreativ und ästhetisch motiviert verwenden. Mit literarischen Texten wird zudem häufig nicht explizit, sondern indirekt durch Bilder kommuniziert. Entwicklungen im Bereich der Domänenadaptation ermöglichen jedoch auch die Optimierung transformerbasierter Modelle auf spezielle Domänen, was Projekte auch im deutschsprachigen Bereich bereits gewinnbringend nutzen konnten (Labusch et al. 2019; Schweter / Baiter 2019; Brunner et al. 2020; Schweter / März 2020). Für die Aufgabe der Emotionsklassifikation findet man im englischsprachigen Bereich Studien, die derartige Methoden für zeitgenössische Texte explorieren (Shmueli / Ku 2019; Acheampong et al. 2020; Cao et al. 2020).

In den Digital Humanities (DH) werden Sentiment-Analyse (die Einteilung, ob ein Text eher positiv/negativ konnotiert ist) und Emotionsklassifikation (die Erkennung bzw. Zuordnung distinkter Emotionskonzepte in Texten) in den letzten Jahren immer populärer. Sie werden verwendet, um moderne Textsorten wie Songtexte (Schmidt et al. 2020a), Filmtexte (Schmidt et al. 2020b) und Texte aus den sozialen Medien zu analysieren (Moßburger et al. 2020; Schmidt et al. 2020c; 2020d) finden aber auch Einsatz für literarische Genres wie beispielsweise Märchen (Alm / Sproat 2005; Mohammad 2011), Romane (Kakkonen / Kakkonen 2011; Mohammad et al. 2011; Reagan et al. 2016; Zehe et al. 2016) oder Dramen (Mohammad 2011; Schmidt / Burghardt 2018; Schmidt et al. 2018a; 2018b; Schmidt 2019; Schmidt et al. 2019a; 2019b; 2019c; Yavuz 2020; Schmidt et al. 2021). Die Ziele variieren dabei von der Exploration von Sentiment- und Emotionsverläufen in einzelnen Werken bis zu Gruppenvergleichen (siehe Kim / Klinger 2019). Die steigende Popularität ist wenig überraschend, da die hermeneutische Analyse von Emotionen eine lange Tradition in der Literaturwissenschaft hat, z.B. in der Dramenanalyse (Pikulik 1965; Wiegmann 1987; Anz 2011; Schonlau 2017).

Im folgenden Proposal präsentieren wir eine Studie aus dem DFG-Projekt *Emotions in Drama*¹ zur Evaluation von Methoden transformerbasierter Emotionsklassifikation für ein annotiertes Korpus historischer deutschsprachiger Dramentexte. Unser Ziel ist, es die Leistung verschiedener Verfahren zu vergleichen und Impulse für Optimierungen auf dieser Textsorte zu sammeln. Im nächsten Kapitel wird dazu zunächst in das verwendete Annotationsschema sowie das annotierte Goldstandard-Korpus eingeführt.² Danach werden die verwendeten Klassifikationsverfahren erläutert. Aktuelle Verfahren werden dabei mit bekannten Baseline-Methoden verglichen und für verschiedene Kategorienmodelle evaluiert. Abschließend werden die Ergebnisse der Evaluation präsentiert.

Annotation und Goldstandard-Erstellung

Zur Evaluation und zum Training von Algorithmen wurde ein Goldstandard für ein Sub-Korpus unseres Gesamtkorpus‘ annotiert.

Definitionen und Annotationsschema

Emotion wird definiert als der Bewusstseinszustand einer Figur, wie sie sich auch in Text ausdrückt. Annotiert wird die eigene oder zugeschriebene Emotion von Figuren in Abhängigkeit von Kontext und Interpretation. Das Schema hebt sich von üblichen Schemata, die meist von der Psychologie inspiriert sind (Wood et al. 2018a; 2018b) ab, um literarische Interessen zu integrieren. Es besteht aus 13 *Sub-Emotionen*, die sich in sechs *Hauptklassen* unterteilen lassen und weiter in die *Polarität* (positiv/negativ) auf höchster Ebene. Abbildung 1 (Kapitel *Annotationsergebnisse*) illustriert die einzelnen Konzepte.

Ein Sonderfall des Schemas ist *emotionale Bewegtheit*, die verwendet wird, um unspezifische emotionale Erregungen zu markieren. Zusammen mit den Klassen *negativ/ positiv* bezeichnen wir diese Sammlung an Oberkategorien als *Dreifach-Polarität*. Es werden sowohl Repliken (einzelne Sprechakte von Figuren) als auch Regieanweisungen annotiert, sofern Annotator*innen dort Emotionen erkennen. Annotator*innen können variable Textlängen pro Einheit annotieren, also einzelne Wörter, Satzteile und mehrere Sätze. Annotationen können sich zudem überlappen. Obwohl es Vorteile hat, feste Annotationseinheiten festzulegen, wurde dieser variable Annotationsstil basierend auf der Erfahrung von Pilotstudien bestimmt.

Annotiertes Teilkorpus

Das zu analysierende Hauptkorpus unseres Gesamtprojektes setzt sich aus unterschiedlichen Dramenkollektionen für die Jahre 1650-1815 aus TextGrid³, GerDracor (Fischer et al. 2019) und anderen Quellen zusammen. Für die vorliegende Studie wurde eine repräsentative Menge von Dramen, gemessen an Sprache und Genre für die Zeit um 1800, gewählt: *Minna von Barnhelm* (1767, Lessing, Komödie), *Kabale und Liebe* (1784, Schiller, Tragödie), *Kasperl' der Mandolettikrämer* (1789, Eberl, Komödie), *Menschenhass und Reue* (1790, Kotezbue, Komödie), *Faust. Eine Tragödie* (1807, Goethe, Tragödie).

Annotationsprozess

Für die Annotation wurde das Tool CATMA (Gius et al. 2020) verwendet. Die Dramen wurden vollständig von Anfang bis Ende annotiert. Die Lektüre des gesamten Dramas ist notwendig, da kontextabhängig annotiert wird. Je zwei studentische Hilfskräfte haben jedes Werk unabhängig voneinander annotiert. Die Hilfskräfte wurden vor der Annotation mittels Pilotstudien von einer Expertenannotatorin trainiert und hatten Zugriff auf eine Annotationsanleitung. Je nach Länge des Textes hatten die Annotator*innen 1-2 Wochen Zeit pro Drama.

Annotationsergebnisse

Der Goldstandard besteht insgesamt aus 6.596 Emotionsannotationen (Abbildung 1).

Hauptklassen und Sub-Emotionen	absolut	%	Avg: tokens	Min: tokens	Max: tokens	Std: tokens
HK: Emotionen der Zuneigung	1 266	19	24,05	1	326	28,61
Lust (-)	50	1	23,22	4	83	16,49
Liebe (+)	783	12	26,16	1	326	33,67
Freundschaft (+)	127	2	22	1	120	18,66
Verehrung (+)	306	5	19,63	1	96	16,36
HK: Emotionen der Freude	1 051	16	23,21	1	223	23,86
Freude (+)	850	13	22,78	1	223	24,3
Schadenfreude (+)	201	3	25,02	1	121	21,89
HK: Emotionen der Angst	706	11	22,42	1	206	24,32
Angst (-)	424	7	16,87	1	173	17,45
Verzweiflung (-)	282	4	30,78	1	206	30,15
HK: Emotionen des Leids	2 196	33	23,87	1	302	26,27
Leid (-)	998	15	26,12	1	302	28,91
Mitleid (-)	318	5	21,61	1	156	21,87
Ärger (-)	880	13	22,14	1	261	24,35
HK: Abscheu (-)	614	9	25,05	1	167	26,19
HK: Emotionale Bewegtheit	763	12	24,4	1	313	32,74
Gesamt	6 596	100	23,82	1	326	24,08

Abb. 1: Verteilung der Annotationsklassen. Nach den jeweiligen Hauptklassen (HK) folgen die Sub-Emotionen. + markiert positive Polarität, - negative Polarität (Avg=Mittelwert, Std=Standardabweichung). Die Aufteilung für die Polarität ist: 3.566 absolut, 54% für negativ, 2.267, 34% positiv und 763, 12% Emotionale Bewegtheit. Alle Prozentangaben sind gerundet.

Auf Polaritätsebene sind die meisten Annotationen negativ (56%), 34% positiv und 11% mit der Klasse „emotionale Bewegtheit“ markiert. Einige Kategorien (z.B. Lust und Freundschaft) wurden selten markiert. Die Token-Statistiken verdeutlichen die Varianz in den Annotationslängen: im Schnitt besteht eine Annotation aber aus 25 Tokens für alle Kategorien.

Da Texteinheiten von variabler Länge und überlappende Texteinheiten annotiert werden können, muss zur Berechnung von Übereinstimmungsmetriken eine Festlegung auf eine Texteinheit getroffen werden. Dazu wird folgende Heuristik angewendet: Für jede Replik oder Regieanweisung wird pro Annotator*in diejenige Annotation markiert, die am meisten (gemessen an der Zahl an annotierten Token) markiert wurde. Keine Annotation pro Replik/Regieanweisung wird als zusätzliche Klasse markiert und dann replikenweise Übereinstimmungen kalkuliert (vgl. Abbildung 2).

Drama	Polarität (κ)	Polarität (%)	Hauptklasse (κ)	Hauptklasse (%)	sub-Emotion (κ)	sub-Emotion (%)
Faust	0,44	67,853	0,345	59,399	0,342	58,064
Kabale und Liebe	0,382	58,908	0,325	50,313	0,312	47,992
Menschenhass und Reue	0,402	75,28	0,347	72,331	0,347	71,91
Minna von Barnhelm	0,406	74,619	0,377	72,752	0,356	71,23
Kasperl' der Mandolettikrämer	0,42	70,83	0,344	65,34	0,312	62,72
Gesamt	0,41	69,498	0,3476	64,027	0,333	62,383

Abb. 2: Übereinstimmungsmetriken für jedes Drama und insgesamt (κ=Cohen's κ; % =prozentuelle Übereinstimmung der Annotator*innen).

Zur Interpretation von Cohen's κ werden im Folgenden in Klammern die Wertebereiche für einzelne Intervalle gemäß Landis und Koch (1977) mitangegeben. Im Schnitt kann man für die Polarität eine moderate Übereinstimmung (laut Landis und Koch gilt moderat für $0,4 < \kappa \leq 0,6$) und für die anderen Kategorien eine schwache Übereinstimmung ($0,1 < \kappa \leq 0,4$) feststellen. Im Ver-

gleich zu anderen Textsorten ist dies eine geringe Übereinstimmung (Wood et al. 2018a; 2018b), die jedoch vergleichbar mit anderen Sentiment- und Emotionsannotationsprojekten mit literarischen und/oder historischen Texten ist (Alm / Sproat 2005; Sprugnoli et al. 2016; Schmidt et al. 2018b; Schmidt et al. 2019b; 2019d). Mehr Erläuterungen und Ergebnisse zur Annotation findet man bei Schmidt et al. (2021c).

Trainings- und Evaluationsmaterial

Im Folgenden werden die Ergebnisse für denjenigen Fall präsentiert, bei dem als Trainings- und Evaluationsmaterial („Goldstandard“) alle Annotationen des obigen Annotationskorpus⁴ verwendet werden (also je zwei Annotationssätze pro Drama). Dadurch liegt folgende Besonderheit vor: Eindeutige und partielle Annotationswidersprüche werden nicht aufgelöst, sondern dem Modell mit als Trainingsmaterial übergeben. Je nach kategorialem System gibt es eine unterschiedliche Menge an partiellen und absoluten Widersprüchen (ca. 16% für Polarität, 14% für Dreifach-Polarität, 28% für Hauptklassen, 47% für Sub-Emotionen). Dieses Verfahren wurde dennoch gewählt, da aufgrund der variablen Annotationspraxis die Auflösung eindeutiger Annotationswidersprüche schwerfällt (siehe Kapitel *Diskussion* mit Anregungen, wie mit diesem Problem in künftigen Studien umzugehen ist). Für weitere Evaluationen mit anderen Korpusinstanzen siehe Schmidt et al. (2021b). Insgesamt besteht der „Goldstandard“ aus 6.596 annotierten Textsequenzen variabler Länge. Nicht-annotiertes Textmaterial wurde dem Goldstandard nicht hinzugefügt. Auch diese Limitation wird in der Diskussion besprochen.

Verfahren der Emotionsklassifikation

Wir definieren die Emotionsklassifikation als single-label-Klassifikationsaufgabe für Textsequenzen variabler Länge für folgende Klassengruppen:

- Polarität (zwei Klassen: positiv vs. negativ; Emotionale Bewegtheit wird hierbei entfernt)
- Dreifach-Polarität (drei Klassen)
- Hauptklassen (sechs Klassen)
- Sub-Emotionen (13 Klassen)

Alle Verfahren wurden in *Python* implementiert. Für die Evaluation und klassische Methoden des maschinellen Lernens wurde *scikit-learn* (Pedregosa et al. 2011) verwendet, für die transformerbasierten Modelle die *Hugging-Face* library (Wolf et al. 2019) und *simpletransformers*⁴.

Baseline-Methoden

Obschon die Leistung lexikonbasierter Sentiment-Analyse meist von *Machine Learning*-Verfahren übertroffen wird, wird sie in den DH häufig angewendet, da keine vorannotierten Trainingskorpora notwendig sind (siehe Kim / Klinger 2019 und Schmidt et al. 2021a). Das Verfahren ist regelbasiert und wird bei Taboada et al. (2011) beschrieben. Wir evaluieren zwei Ansätze: (1) das Lexikon *SentiWortschatz* (SentiWS) (Remus et al. 2010) ohne Vorverarbeitung (im Folgenden als *lb-sentiws* bezeichnet), (2) SentiWS kombiniert mit Methoden wie Lemmatisierung und

Lexikonerweiterung (Schmidt / Burghardt 2018) (*lb-sentiws-optimized*). Letztere Methodik erzielte gute Ergebnisse in historischen deutschsprachigen Dramen (Schmidt / Burghardt 2018). Die gewählten Ansätze können nur für die Polarität angewendet werden, da keine differenzierten Emotionsannotationen in SentiWS vorhanden sind.

Wir evaluieren zudem zwei klassische Methoden des maschinellen Lernens: (1) Repräsentation über Termfrequenzen in einem bag-of-words-Modell und dem Lern-Algorithmus *Multinomial Naive Bayes* (*bow-mnb*) sowie (2) *Support Vector Machines* (SVM) (*bow-svm*) als Lern-Algorithmus. Methode (2) wurde mit dem rbf-kernel der SVC-Klasse von *scikit-learn* umgesetzt.⁵ Für mehr Informationen über bag-of-words-Ansätze siehe Gonz  les-Carvajal et al. (2021). Die Algorithmen wurden in einer stratifizierten 5x5 Kreuzevaluation trainiert und evaluiert.

fastText

Statische Sprachmodelle repr  sentieren W  rter als Vektoren in Vektorr  umen, so dass geometrische Verh  ltnisse der jeweiligen Semantik entsprechen. Diese Repr  sentationen (*word embeddings*) k  nnen als Input f  r neuronale Netze genutzt werden. Wir evaluieren das *word embedding fastText* (Bojanowski et al. 2017), da es im Vergleich zu anderen statischen Modellen gute Ergebnisse f  r deutsche Sprache erzielt (Schmitt et al. 2018). Wir nutzen deutschsprachige *fastText embeddings*⁶ trainiert auf der deutschsprachigen Wikipedia sowie ein rekurrentes neuronales Netzwerk (RNN) zur Klassifikation (Cho et al. 2014). Bez  glich der Hyperparameter wird der empfohlene Default des FLAIR-frameworks gew  hlt (Akbik et al. 2019)⁷ und je ein Modell in einem stratifizierten 5x5-Setting f  r 12 Epochen trainiert und evaluiert. F  r alle Evaluationsmetriken wird der Mittelwert aus den Ergebnissen der f  nf Modelle gebildet.

Transformerbasierte Sprachmodelle (zeitgen  ssische Sprache)

Als transformerbasierte Sprachmodelle werden dynamische *word embeddings* wie BERT (Devlin et al. 2018) oder ELECTRA (Clark et al. 2020) bezeichnet, die in Erweiterung zu statischen Modellen den Kontext eines Wortes in seiner Umgebung. Wir evaluieren einige der wichtigsten und (  ber die *Hugging Face*-Plattform⁸) frei verf  gbaren Modelle, die auf zeitgen  ssischer Sprache trainiert wurden (Abbildung 3). Die gew  hlten Modelle erreichen state-of-the-art-Ergebnisse in standardisierten Evaluationen auf deutscher Sprache (Chan et al. 2020).

ID	Texte f��r das Vortraining	Zugeh��riges Paper (wenn verf��gbar) und Provider
<i>bert-base-german-cased</i>	Wikipedia, juristische Texte, News (~ 12 GB)	Deepset
<i>dbmdz-bert-base-german-cased</i>	Wikipedia, B��cher, Untertitel, Web-Texte, News (~ 16 GB)	MDZ Digital Library
<i>electra-base-german-uncased</i>	Wikipedia, Untertitel, News (~ 73 GB)	German-NLP-Group
<i>gbert-large</i>	Web-Texte, Wikipedia, Untertitel, B��cher, juristische Texte (~ 161 GB)	Deepset (Chan et al., 2020)
<i>gelectra-large</i>	Web-Texte, Wikipedia, Untertitel, B��cher, juristische Texte (~ 161 GB)	Deepset (Chan et al., 2020)

Abb. 3: Evaluierte transformerbasierte Modelle (vortrainiert mit zeitgen  ssischer Sprache).

Für die Klassifikationsaufgabe werden die Modelle in einem „Fine-Tuning“-Schritt mit dem Goldstandard trainiert. Für die konkrete Implementierung folgen wir den jeweiligen Empfehlungen für die gewählte Architektur (Devlin et al. 2018; Clark et al. 2020)⁹ und nutzen die *Hugging Face*-Bibliothek (Wolf et al. 2020). Pro Sprachmodell und Klassifikationstask werden fünf Klassifikationsverfahren in einem stratifizierten 5x5-setting für je vier Epochen trainiert und Mittelwerte gebildet.

Transformerbasierte Sprachmodelle (historische/poetische Sprache)

Die Performanz von Klassifikations-Aufgaben kann verbessert werden, indem Texte der gleichen Domäne zum Vortraining von transformerbasierten Modellen genutzt werden (siehe Rietzler et al. 2020; Gururangan et al. 2020). Man kann entweder (1) selbst ein Modell von Grund auf mit domänennahen Texten erstellen oder (2) Modelle zeitgenössischer Sprache mit domänenspezifischen historischen Texten nachtrainieren. Beide Methoden wurden bereits erfolgreich im Kontext deutscher, historischer Sprache angewendet (Labusch et al. 2019; Schweter / Baiter 2019; Schweter / März 2020; Brunner et al. 2020).

ID	Vortrainierte Texte	Zeitraum	Zugehöriges Paper (wenn verfügbar) und Provider
<i>bert-base-german-europeana-cased</i>	Europeana-Zeitungen (51 GB)	18.-20. Jahrhundert	MDZ Digital Library (Schweter, 2020)
<i>electra-base-german-europeana-cased-discriminator</i>	Europeana-Zeitungen (51 GB)	18.-20. Jahrhundert	MDZ Digital Library (Schweter, 2020)
<i>literary-german-bert</i>	Basiert auf <i>bert-base-german-dbmz-cased</i> weiter vortrainiert mit <i>Corpus of German-Language-Fiction (CGLF)</i> (~ 1 GB)	CGLF: hauptsächlich 1840-1930	Severin Simmler
<i>bert-base-historical-german-rw-cased</i>	Märchen, historische Zeitungen, narrative Texte, Texte von Projekt Gutenberg (genaue Größe unbekannt)	1840-1920	Brunner et al. (2020)

Abb. 4: Evaluierbare transformerbasierte Modelle vortrainiert mit historischer Sprache.¹⁰

Auch hier evaluieren wir etablierte vortrainierte Modelle, die über die *Hugging Face*-Plattform frei verfügbar sind. Abbildung 4 fasst die Daten der Modelle zusammen. Alle Modelle nähern sich dem Kontext unserer Dramen-Texte auf historischer Ebene oder dadurch, dass narrative/poetische Texte genutzt werden, an. Des Weiteren wurde das Modell *bert-base-german-cased* noch mit den Texten des eigenen Korpus nachtrainiert, zum einen mit unserem Hauptkorpus GerDracor (*bert-base-german-cased-main-corpus*) und in einem zweiten Ansatz lediglich mit den annotierten Dramen (*bert-base-german-cased-annotated-texts*). Das Nachtraining wurde für 4 Epochen mit den default-settings der *simpletransformer*-library durchgeführt.¹¹ Das Implementierungs-, Trainings- und Evaluationsverfahren sowie die gewählten Hyperparameter für die Emotionsprädiktion sind äquivalent zum vorigen Kapitel.

Ergebnisse

Hauptmetrik zur Interpretation der Ergebnisse ist die *accuracy*, also der Anteil an korrekt erkannten Annotationen an allen Annotationen (siehe Abbildung 5). Weitere Details und Informationen zu den Ergebnissen der Studie findet man bei Schmidt et al. (2021d).

Methodengruppe	Methode / accuracy	Polarität	Dreifach-Polarität	Hauptkategorie	Sub-Emotion
Baseline-Methoden	random baseline	.500	.333	.167	.077
	majority baseline	.612	.541	.333	.151
	lb-sentisw	.445	-	-	-
	lb-sentisw-optimized	.588	-	-	-
	bow-mnb	.742	.659	.451	.348
	bow-svm	.685	.603	.392	.284
Statisches Sprachmodell	fasttext	.714	.647	.404	.289
Zeitgenössische Transformer-Modelle	bert-base-german-cased	.804	.711	.512	.428
	dbmdz-bert-base-german-cased	.804	.716	.517	.430
	electra-base-german-uncased	.776	.690	.474	.358
	gbert-large	.821	.740	.545	.467
	gelectra-large	.825	.748	.564	.460
	bert-base-german-europeana-cased	.798	.718	.528	.420
Historische Transformer-Modelle	electra-base-german-europeana-cased-discriminator	.808	.722	.525	.416
	literary-german-bert	.799	.718	-	-
	bert-base-historical-german-rw-cased	.813	.723	.524	.444
Transformer-Modelle trainiert mit eigenem Korpus	bert-base-german-cased-main-corpus	.796	.714	.492	.379
	bert-base-german-cased-annotated-texts	.809	.709	.505	.425

Abb. 5: Klassifikationsergebnisse für alle Methoden (die drei besten Ergebnisse je Kategorie sind hervorgehoben).

Alle gewählten Methoden übertreffen in den einzelnen Settings die *random* und *majority*-baseline. Die Ergebnisse der lexikonbasierten Sentiment-Analyse bewegen sich auf einem ähnlichem Niveau für Evaluationen auf unterschiedlichen literarischen Texten (Fehle et al. 2021). Die beste Erkennungsrate für Polarität beträgt 83% und wird vom Modell *gelectra-large* erreicht. Gleiches gilt für die Dreifach-Polarität mit 75% sowie die Hauptklassen (55%). Das beste Modell für die Sub-Emotionen ist *gbert-large* mit jedoch lediglich 47% Erkennungsrate. Transformerbasierte Modelle erreichen im Schnitt wesentliche bessere Erkennungsraten als alle Baseline-Methoden oder fastText. Mit zunehmender Klassenzahl werden die Ergebnisse (trivialerweise) schlechter. Auch die Abstände zwischen bester und schlechtester Methode werden geringer. Die drei besten Modelle sind konsistent die zwei größten Modelle zeitgenössischer Sprache *gbert-large* und *gelectra-large* sowie das auf historische und narrative Sprache optimierte Modell *bert-base-historical-german-rw-cased*.

Diskussion

Obschon die Menge an annotiertem Material im Vergleich zu Studien auf der Basis anderer Textsorten limitiert ist, konnten wir erste Erkenntnisse für die Optimierung computergestützter Methoden sammeln. Für Polarität und Dreifach-Polarität erreichen die besten Modelle in ihren Default-Settings bereits Ergebnisse, die durchaus vergleichbar sind mit state-of-the-art-Resultaten für Sentiment- und Emotionsklassifikation in anderen Bereichen (Yang et al. 2019; Munikar et al. 2019; Cao et al. 2020; Dang et al. 2020). Die besten Ergebnisse erzielen grundsätzlich die derzeit größten transformerbasierten Modelle für die deutsche Sprache. Die Optimierung für historische oder poetische Sprache hat lediglich geringfügige Verbesserungen gegenüber den äquivalenten kontemporären Modellen aufgezeigt. Ein Grund dafür ist möglicherweise, dass die gewählten historischen Modelle noch zu viele Texte aus dem 19. und 20. Jahrhundert enthalten, die doch zu weit entfernt von unserer Zeitepoche sind. Wir befinden uns momentan im Prozess der Akquise großer Textmengen aus dem

entsprechenden Zeitraum, um vortrainierte Modelle zu evaluieren, die noch stärker an unsere Domäne angepasst sind.

Für die mehrklassigen Kategoriensysteme können keine zufriedenstellenden Ergebnisse erzielt werden. Dies ist ohne größere Optimierung für derartige Klassifikationsverfahren nicht ungewöhnlich. Wir planen sowohl die Anwendung verschiedener empfohlener Verfahren, um mit dem Klassenungleichgewicht umzugehen (Buda et al. 2018) und die Optimierung von Hyperparametern als auch die Exploration des Einsatzes einer neutralen „Nicht-annotiert“-Klasse. Im Bereich der Annotation soll eine Expertenannotation eingefügt werden, welche die Entscheidungen der ersten beiden Annotationen berücksichtigt, aber eine eigenständig verwendbare, widerspruchsfreie Annotationsschicht darstellt. Evaluationsergebnisse mittels der Anwendung von manuellen Widerspruchsaufösungen findet man bei Schmidt et al. (2021b). Wir lassen derzeit weitere Texte annotieren und explorieren historische *word embeddings*, um akzeptable Ergebnisse für die Hauptkategorien zu erreichen und Emotionen in größeren Mengen unseres Korpus vorhersagen zu können.

Fußnoten

1. Das Projekt wird von der Deutschen Forschungsgemeinschaft (DFG) im Rahmen des Schwerpunktprogramms Computational Literary Studies (SPP 2207/1) gefördert https://dfg-spp-cls.github.io/projects_en/2020/01/24/TP-Emotions_in_Drama/ (Sachbeihilfen DE 2188/3-1 und WO 835/4-1, Projektnummer: 424207618).
2. Zur Definition des Emotionsbegriffs und zur Emotionsauswahl vgl. Dennerlein et al. 2022
3. <https://textgrid.de>
4. <https://simpletransformers.ai/>
5. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC>
6. <https://fasttext.cc/docs/en/crawl-vectors.html>
7. Lern-Rate: 0.1, Batch-Größe: 32
8. <https://huggingface.co/>
9. Lern-Rate: 0.00004; Batch-Größe: 32, maximale Sequenzlänge: 128; Adam als Optimizer
10. Aus Architektur-Gründen wird das Modell literary-german-bert nur für Polarität/Dreifach-Polarität evaluiert
11. Siehe <https://simpletransformers.ai/docs/lm-specifics/>

Bibliographie

- Acheampong, Francisca Adoma / Wenyu, Chen / Nunoo-Mensah, Henry (2020): "Text-based emotion detection: Advances, challenges, and opportunities.", in: *Engineering Reports* 2.7: e12189.
- Akbik, Alan / Bergmann, Tanja / Blythe, Duncan / Rasul, Kashif / Schweter, Stefan / Vollgraf, Roland (2019): "FLAIR: An easy-to-use framework for state-of-the-art NLP.", in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.
- Alm, Cecilia Ovesdotter / Sproat, Richard (2005): "Emotional sequencing and development in fairy tales.", in: *International Conference on Affective Computing and Intelligent Interaction*. Springer, Berlin, Heidelberg.
- Anz, Thomas (2011): "Todesszenarien: literarische Techniken zur Evokation von Angst, Trauer und anderen Gefühlen.", in: *Emotionale Grenzgänge*. Würzburg, pp. 54–59.
- Bojanowski, Piotr / Grave, Edouard / Joulin, Armand / Mikolov, Tomas (2017): "Enriching word vectors with subword information.", in: *Transactions of the Association for Computational Linguistics* 5: 135–146.
- Brunner, Annalen / Duyen Tanja Tu, Ngoc / Weimer, Lukas / Jannidis, Fotis (2020): "To BERT or not to BERT-Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation.", in: *SwissText/KONVENS*.
- Buda, Mateusz / Maki, Atsuto / Mazurowski, Maciej A. (2018): "A systematic study of the class imbalance problem in convolutional neural networks.", in: *Neural Networks* 106: 249–259.
- Cao, Lihong / Peng, Sancheng / Yin, Pengfei / Zhou, Yongmei / Yang, Aimin / Li, Xinguang (2020): "A Survey of Emotion Analysis in Text Based on Deep Learning.", in: *2020 IEEE 8th International Conference on Smart City and Informatization (iSCI)*. IEEE.
- Chan, Branden / Schweter, Stefan / Möller, Timo (2020): "German's Next Language Model.", in: *arXiv preprint arXiv:2010.10906*
- Cho, Kyunghyun / Merriënboer, Bart van / Bahdanau, Dzmitry / Bengio, Yoshua (2014): "On the properties of neural machine translation: Encoder-decoder approaches.", in: *arXiv preprint arXiv:1409.1259*
- Clark, Kevin / Luong, Minh-Thang / Le, Quoc V. / Manning, Christopher D. (2020): "Electra: Pre-training text encoders as discriminators rather than generators.", in: *arXiv preprint arXiv:2003.10555*
- Cortiz, Diogo (2021): "Exploring Transformers in Emotion Recognition: a comparison of BERT, DistilBERT, RoBERTa, XLNet and ELECTRA." *arXiv preprint arXiv:2104.02041*
- Dang, Nhan Cach / Moreno-García, María N. / De la Prieta, Fernando (2020): "Sentiment-Analyse based on deep learning: A comparative study.", in: *Electronics* 9.3 (2020): 483.
- Dennerlein, Katrin / Schmidt, Thomas / Wolff, Christian (2022): "Emotionen im kulturellen Gedächtnis bewahren.", in: *Book of Abstracts, DHd2022*.
- Devlin, Jacob / Chang, Ming-Wei / Lee, Kenton / Toutanova, Kristina (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding.", in: *arXiv preprint arXiv:1810.04805*.
- Fehle, Jakob / Schmidt, Thomas / Wolff, Christian (2021): "Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques", in: *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, 86–103.
- Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hecht, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama." Zenodo. <<https://doi.org/10.5281/zenodo.4284002>>
- Evelyn, Gius / Meister, Jan Christoph / Petris, Marco / Meister, Malte / Bruck, Christian / Jacke, Janina / Schumacher, Mareike / Flüh, Marie / Horstmann, Jan (2020): "CATMA." Zenodo. <<https://doi.org/10.5281/zenodo.4353618>>
- González-Carvajal, Santiago / Garrido-Merchán, Eduardo C. (2020): "Comparing BERT against traditional machine learning text classification.", in: *arXiv preprint arXiv:2005.13012*
- Gururangan, Suchin / Marasović, Ana / Swamydiptra, Swabha / Lo Kyle / Beltagy, Iz / Downey, Doug et al. (2020): "Don't stop pretraining: adapt language models to domains and tasks.", in: *arXiv preprint arXiv:2004.10964*

- Kakkonen, Tuomo / Kakkonen, Gordana Galić** (2011): "SentiProfiler: creating comparable visual profiles of sentimental content in texts.", in: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*.
- Kim, Evgeny / Klinger, Roman** (2019): "A survey on sentiment and emotion analysis for computational literary studies.", in: *Zeitschrift für digitale Geisteswissenschaften*.
- Labusch, Kai / Neudecker, Clemens / Zellhofer, David** (2019): "BERT for Named Entity Recognition in Contemporary and Historical German.", in: *Proceedings of the 15th Conference on Natural Language Processing, Erlangen, Germany*.
- Landis, J. Richard / Koch, Gary G.** (1977): "The measurement of observer agreement for categorical data.", in: *biometrics* (1977): 159-174.
- Mohammad, Saif** (2011): "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales.", in: *arXiv preprint arXiv:1309.5909*
- Moßburger, Luis / Wende, Felix / Brinkmann, Kay / Schmidt, Thomas** (2020): "Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum", in: *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, 70-81.
- Munika, Manish / Shakya, Sushil / Shrestha, Aakash** (2019): "Fine-grained sentiment classification using bert." *2019 Artificial Intelligence for Transforming Business and Society (AITB)*. Vol. 1. IEEE.
- Pedregosa, Fabian et al.** (2011): "Scikit-learn: Machine learning in Python", in: *Journal of machine Learning research*, 12, 2825-2830.
- Pikulik, Lothar** (1966): *"Bürgerliches Trauerspiel" und Empfindsamkeit*. Köln, Graz.
- Reagan, Andrew J. / Mitchell, Lewis / Kiley, Dilan / Danforth, Christopher M. / Dodds, Peter Sheridan** (2016): "The emotional arcs of stories are dominated by six basic shapes." *EPJ Data Science* 5.1: 1-12.
- Remus, Robert / Quasthoff, Uwe / Heyer, Gerhard** (2010): "SentiWS-A Publicly Available German-language Resource for Sentiment-Analyse.", in: *LREC*.
- Rietzler, Alexander / Stabinger, Sebastian / Opitz, Paul / Engl, Stefan** (2019): "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification.", in: *arXiv preprint arXiv:1908.11860*
- Schmidt, Thomas** (2019): "Distant Reading Sentiments and Emotions in Historic German Plays", in: *Abstract Booklet, DH_Budapest_2019*. Budapest, Hungary, 57-60.
- Schmidt, Thomas / Burghardt, Manuel** (2018): "An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing", in: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, 139-149.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin** (2018a): "'Kann man denn auch nicht lachend sehr ernsthaft sein?' – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen", in: *Book of Abstracts, DHd 2018*.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin** (2018b): "Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior.", in: Sandra Kübler, Heike Zinsmeister (eds.), *Proceedings of the Workshop on Annotation in Digital Humanities (annDH 2018)*. Sofia, Bulgaria, 47-52.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin / Wolff, Christian** (2019a): "Katharsis - A Tool for Computational Drametrics", in: *Book of Abstracts, Digital Humanities Conference 2019 (DH 2019)*. Utrecht, Netherlands.
- Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin / Wolff, Christian** (2019b): "Sentiment Annotation in Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts", in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. LDK Posters. Leipzig, Germany.
- Schmidt, Thomas / Burghardt, Manuel / Wolff, Christian** (2019c): "Towards Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti.", in: *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*. Copenhagen, Denmark, 405-414.
- Schmidt, Thomas / Winterl, Brigitte / Maul, Milena / Schark, Alina / Vlad, Andrea / Wolff, Christian** (2019d): "Inter-Rater Agreement and Usability: A Comparative Evaluation of Annotation Tools for Sentiment Annotation", in: Draude, C., Lange, M. & Sick, B. (Hrsg.), *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*. Bonn: Gesellschaft für Informatik e.V., 121-133. DOI: 10.18420/inf2019_ws12
- Schmidt, Thomas / Bauer, Marlene / Habler, Florian / Heuberger, Hannes / Pils, Florian / Wolff, Christian** (2020a): "Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte", in *Book of Abstracts, DHd 2020*, 296-299.
- Schmidt, Thomas / Engl, Isabella / Halbhuber, David / Wolff, Christian** (2020b): "Comparing Live Sentiment Annotation of Movies via Arduino and a Slider with Textual Annotation of Subtitles", in: *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*, 212-223.
- Schmidt, Thomas / Hartl, Philipp / Ramsauer, Dominik / Fischer, Thomas / Hilzenthaller, Andreas / Wolff, Christian** (2020c): "Acquisition and Analysis of a Meme Corpus to Investigate Web Culture", in: *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Conference Abstracts*. Ottawa, Canada.
- Schmidt, Thomas / Kaindl, Florian / Wolff, Christian** (2020d): "Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit", in: *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*. Riga, Latvia.
- Schmidt, Thomas / Dangel, Johanna / Wolff, Christian** (2021a): "SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities", in: Schmidt, Thomas / Wolff, Christian (Eds.), *Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)*. Glückstadt: Verlag Werner Hülsbusch, 156-172. DOI: 10.5283/epub.44943
- Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian** (2021b): "Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language", in: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 67-79.
- Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian** (2021c): "Towards a Corpus of Historical German Plays with Emotion Annotations", in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021d): "Using Deep Neural Networks for Emotion Analysis of 18th and 19th century German Plays", in: *Fabrikation von Erkenntnis: Experimente in den Digital Humanities*. Melusina Press. DOI:10.26298/melusina.8f8w-y749-udlf

Schmitt, Martin / Steinheber, Simon / Schreiber, Konrad / Roth, Benjamin (2018): "Joint aspect and polarity classification for aspect-based Sentiment-Analysis with end-to-end neural networks.", in: *arXiv preprint arXiv:1808.09238*

Schonlau, Anja (2017): *Emotionen im Dramentext: eine methodische Grundlegung mit exemplarischer Analyse zu Neid und Intrige 1750-1800*. De Gruyter.

Schweter, Stefan (2020): *Europeana BERT and ELECTRA models*. <<https://doi.org/10.5281/zenodo.4275044>>

Schweter, Stefan / Baiter, Johannes (2019): "Towards robust named entity recognition for historic german.", in: *arXiv preprint arXiv:1906.07592* (2019).

Schweter, Stefan / März, Luisa (2020): "Triple E-Effective Ensembling of Embeddings and Language Models for NER of Historical German.", in: *CLEF (Working Notes)*.

Shmueli, Boaz / Lun-Wei Ku (2019): "Socialnlp emotionx 2019 challenge overview: Predicting emotions in spoken dialogues and chats.", in: *arXiv preprint arXiv:1909.07734*

Sprugnoli, Rachele / Tonelli, Sara / Marchetti, Alessandro / Moretti, Giovanni (2016): "Towards Sentiment-Analysis for historical texts.", in: *Digital Scholarship in the Humanities* 31.4: 762-772.

Wiegmann, Hermann (Hrsg.) (1987): *Die ästhetische Leidenschaft: Texte zur Affektenlehre im 17. und 18. Jahrhundert*.

Wolf, Thomas / Debut, Lysandre / Sanh, Victor / Chaumond, Julien / Delangue, Clement / Moi, Anthony et al. (2019): "Huggingface's transformers: State-of-the-art natural language processing.", *arXiv preprint arXiv:1910.03771*

Wood, Ian / McCrae, John / Andryushechkin, Vladimir / Buitelaar, Paul (2018a): "A comparison of emotion annotation schemes and a new annotated data set.", in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wood, Ian / McCrae, John / Andryushechkin, Vladimir / Buitelaar, Paul (2018b): "A comparison of emotion annotation approaches for text.", in *Information* 9.5: 117.

Yang, Kisu / Lee, Dongyub / Whang, Taesun / Lee, Seolhwa / Lim, Heuiseok (2019): "Emotionx-ku: Bert-max based contextual emotion classifier.", in: *arXiv preprint arXiv:1906.11565*

Yavuz, Mehmet Can (202) "Analyses of Character Emotions in Dramatic Works by Using EmoLex Unigrams.", in: *CLiC-it*.

Zehe, Albin / Becker, Martin / Hettlinger, Lena / Hotho, Andreas / Reger, Isabella / Jannidis, Fotis (2016): "Prediction of happy endings in German novels based on sentiment information.", in: *3rd Workshop on Interactions between Data Mining and Natural Language Processing, Riva del Garda, Italy*.

Executable Papers in den Computational Humanities Von technischen Herausforderungen und erkenntnistheoretischen Mehrwerten

Walkowski, Niels-Oliver

niels-oliver.walkowski@uni.lu
Universität Luxemburg

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de
Universität Leipzig

Einleitung

Die Gegenüberstellung von Programmcode und narrativem Text bei gleichzeitiger Evaluation ihrer Beziehungen, Ähnlichkeiten sowie Modi der gegenseitigen Bezugnahme ist so alt wie der Computer selbst. Bereits vor 40 Jahren appellierte der Informatiker Donald Knuth, man solle Programme doch besser als "literarische Werke" betrachten (Knuth 1984, 97). Knuths Ansatz bildete die Grundlage vieler gegenwärtiger Softwaredokumentationssysteme wie z.B. Pythons *docutils*, *reStructuredText* und *sphinx*. Der durch Knuth sprichwörtlich gewordene Begriff des *literate programming* weist jedoch eindeutig darauf hin, dass diese Idee sehr viel weitreichender gemeint war als das zum quasi Standard gewordene formatierungs-fähige Dokumentieren eines Programms im Programmcode selbst. Legt der Begriff der Dokumentation immer noch eine Priorisierung nahe – das Eigentliche ist der Programmcode – so eröffnen *Executable Papers* (ExP) unter dem Vorzeichen des wissenschaftlichen Publizierens ein Feld in dem beide Gegenstandsbereiche sehr viel egalitärer und vielfältiger miteinander interagieren als im Bereich der Softwareentwicklung. ExP ist ein Sammelbegriff für einen Diskurs in den (digitalen) Wissenschaften sowie einer Reihe von Aktivitäten, die die gleichzeitige Veröffentlichung von Forschungsartikeln mit dem der Forschung zu Grunde liegendem, ausführbaren Programmcode vorsieht. Diese Aktivitäten sind unter anderem eng verbunden mit *Elseviers Executable Papers Grand Challenge* im Jahr 2011 (Gabriel 2011). Die Genealogie der dahinter liegenden Idee beginnt aber auch hier wesentlich früher. Erste Beispiele lassen sich bereits zum Ende des letzten Jahrtausends entdecken (vgl. Singh et al. 1998; Burg et al. 2000).

Die Realisierung von ExP variiert allerdings sehr stark. Dies betrifft sowohl die Art der Integration von Text und Programmcode als auch den technischen Ansatz, mit dem eine stabile, möglichst kontext-ungebundene Ausführbarkeit des Programmcodes gesichert werden soll. Gerade frühere Ansätze setzten dabei auf Komplettvirtualisierungen von Betriebssystemen, die ein lokal installierbares Image oder einen Remote-Desktop zur Verfügung stellen innerhalb dessen Leser:innen das ExP konsumieren konnten (Brammer et al. 2011; van Gorp and Mazanek 2011). Mittlerweile werden dafür nunmehr eher "leichtgewichtige" Virtualisierungsverfahren wie Docker (Cito, Ferme, und Gall 2016;

Boettiger 2015) genutzt. Ein anderer Ansatz arbeitet mit deklarativen Beschreibungen der für die Ausführbarkeit von ExP aufzulösenden Abhängigkeitsbäume, die wie eine Art Manifest mit dem ExP zusammen veröffentlicht werden (Pebesma et al. 2012; Nüst et al. 2016). Auch im Kontext von PDFs haben *Springer Nature* und der PDF Reader *ReadCube* vor einigen Jahren mit der Integration live ausführbarer, interaktiver Elemente experimentiert.

ExP arrangieren den Programmcode in den Randbereichen des Textes, vergleichbar mit einer Fußnote (Ciepiela et al. 2013; siehe Abbildung 1) oder aber auch auf Tokenebene, eng mit dem Text verwoben (Maciocci et al. 2019). Manchmal bilden Text und Programmcode zwei separate Bestandteile einer übergreifenden Publikation (Agnone 2020; Kray et al. 2019), ein anderes mal wird der vollständige Text zusammen mit den ausführbaren Bestandteilen überhaupt erst generiert (Smith et al. 2013).

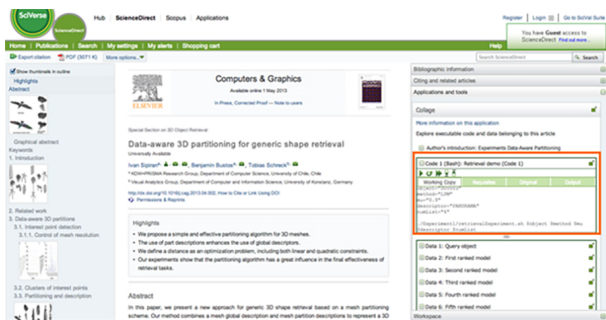


Abb. 1: Einer der Gewinner des Elseviers Executable Papers Grand Challenge mit ausführbaren Codefragmenten im Randbereich des Artikels (Hervorhebung)

Eine der größeren Herausforderungen von ExP betrifft die Möglichkeiten und Mittel ihrer Erstellung. Vereinzelt hat dies, wie im Fall des *Collage Authoring Environment* (Ciepiela et al. 2013) oder *Stencila* (Aufreiter et al. 2018) zur Entwicklung genuiner ExP-"Editoren" geführt. Eine andere Perspektive basiert auf der Weiterentwicklung eines bestehenden Ökosystems aus Werkzeugen, Infrastrukturen und Initiativen. Gemeint sind hier Projekte wie das *Jupyter Notebook*, *Git* und *Docker* u. ä. Gerade das *Jupyter Notebook* ist seit Jahren ein integraler Bestandteil bestimmter Bereiche der Wissenschaftskommunikation. Doch der potenzielle Wert für die Anfertigung vollwertiger Wissenschaftspublikationen wird ebenfalls seit einiger Zeit erkannt (Kluyver et al. 2016; Chandre et al. 2021), und das nicht nur durch die Integration von *Jupyter* in den Publikationsworkflow des O'Reilly-Verlags (Odeh 2015). Mit dem *Jupyter Book Framework* hat das *Executable Book Project* kürzlich einen weiteren interessanten Schritt auf diesem Weg vorgelegt.

Executable Publications in den Humanities

Die eben zusammengefasste Entwicklung wird zu großen Teilen von der Informatik sowie auch den Lebens- und Naturwissenschaften getragen. Beispiele aus den Geisteswissenschaften, wie etwa die *Executable Music Documents* (De Roure et al. 2014), Melanie Walshs *Introduction to Cultural Analytics & Python* (2021) oder mit Einschränkungen innerhalb von Alan Lius *WhatEveryISays*-Projekt (Liue et al. 2017), sind selten – und das obwohl *Jupyter Notebooks* auch in den Computational Humanities längst zum Grundlagenwerkzeug geworden sind. Schaut man je-

doch auf die etablierte Zeitschriftenlandschaft oder in das akademische Verlagswesen, so sieht die Situation disziplinübergreifend nicht viel anders aus. Zweifellos hat dies mit den großen Herausforderungen zu tun, denen man sich stellen muss, sollen ExP eine nachhaltige und vollwertige Form wissenschaftlichen Publizierens werden. Dabei stehen die technischen Herausforderungen nicht einmal unbedingt im Vordergrund. Publizieren ist mehr als eine Ressource online verfügbar machen, es ist der Übergang von einer klandestinen, informellen und damit weniger verpflichtenden Wissenschaftskommunikation in eine wesentlich komplexere Kommunikationsökologie mit stärker kodifizierten Normen, Erwartungen, Bewertungsmaßstäben, Rollen und Infrastrukturen – ein Umstand der bei innovativen Publikationsformaten häufig unterschätzt wird (Walkowski 2019). Auf der anderen Seite bieten *Jupyter Notebooks* gerade in den Geisteswissenschaften einen idealen Ausgangspunkt für die Etablierung von ExP, stellen sie doch die bislang weitestgehende Entsprechung mit Knuths Vision des Programmierens als eine literarische Tätigkeit dar. Sie sind daher am stärksten in der Lage, den häufig narrativen Logiken geisteswissenschaftlicher Erkenntnis- und Darstellungsformen zu entsprechen.

Lessons learned aus einem aktuellen ExP-Projekt

Vor diesem Hintergrund hat sich die *Melusina Press Luxembourg* und die Computational Humanities-Gruppe Leipzig im Rahmen des vDHd Bandes "Fabrikation von Erkenntnis" dazu entschlossen, die Realisierbarkeit von ExP als vollwertige Publikationen der Computational Humanities zu erproben. Der Vortrag möchte sowohl die Kontextualisierung und die Potenziale dieses Publikationsformats im zuvor angedeuteten Sinne vorstellen, als auch detailliert die relevanten Fragen und Problemfelder in der Perspektive einer funktionierenden Publikationsökologie evaluieren. Nachfolgend stellen wir einige vorläufige Erkenntnisse zu drei unterschiedlichen Problemfeldern aus dem aktuellen vDHd-Publikationsprojekt vor.

Zum Verhältnis von epistemischen Mehrwerten und (technischen) Aufwänden

Es hat sich herausgestellt, dass die Aufbereitung eines *Jupyter Notebooks* zu einem ExP trotz der Tatsache, dass ersteres meist bereits während des Forschungsprozesses entsteht, einen Arbeitsaufwand mit sich bringt, der den eines herkömmlichen Artikels deutlich übersteigt. Gleichzeitig bleiben häufig propagierte Mehrwerte, wie die bessere Nachvollziehbarkeit von Forschungsmethoden sowie die Reproduzierbarkeit von Forschungsergebnissen (Lasser 2020), schwer empirisch evaluierbar. Inwieweit ist zum Beispiel die Eleganz der Möglichkeit Code während des Lesens innerhalb des Artikels ausführen zu können ein entscheidender Vorteil gegenüber dem Lesen desselben Codes in einem statischen Export des Notebooks oder dem beiläufigen Verfügbarmachen von Scripten in einem Git-Repository. Solche Fragen lassen sich nur für konkrete Szenarien beantworten. Mit unserer Sektion im vDHd-Band wollten wir einen Beitrag für die bessere Identifizierbarkeit dieser Szenarien liefern. Da geisteswissenschaftliche Forschung sowieso sehr viel seltener einer rein empirischen Evidenzlogik folgt, scheint uns der größte Mehrwert für die Geisteswissenschaften eher in den epistemischen Potenzialen dieses Formats zu liegen. Vor diesem Hintergrund ist es jedoch bemer-

kenswert, dass die meisten Einreichungen das ExP-Format eher in Richtung eines Methodenpapiers bzw. Tutorials interpretiert haben. Diese Beobachtung zeigt, dass ExP, verstanden als eine Form des *literate programming* im weitestmöglichen Sinne, auch auf Autor:innenseite noch Raum für eine kreative Aneignung zulässt

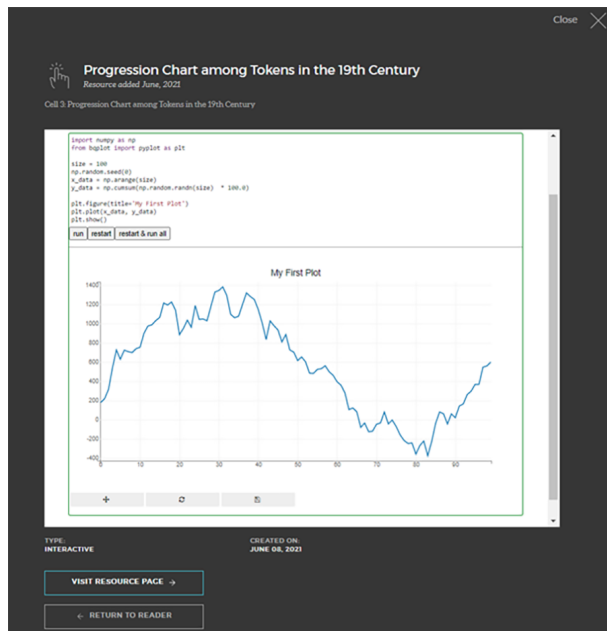


Abb. 2: Artikeloverlay mit ausgeführter und editierbarer Codezelle auf Melusina Press

Zu technischen Anforderungen und nachhaltiger Infrastruktur

Technische und infrastrukturelle Herausforderungen bestehen am augenscheinlichsten zunächst erst einmal in der situativen Bereitstellung jeder Zeit abrufbarer Rechnerressourcen, die für das Ausführen des Codes notwendig sind, und bei der verlässlichen Auflösung von Abhängigkeiten. Hier hat sich mit dem Projekt *myBinder* ein Verfahren etabliert, das unter Einhaltung bestimmter Vorgaben aus einem Git-Repository mit einem *Jupyter Notebook* ein lauffähiges *Docker Image* generiert und dieses mittels Link über virtuelle Server kooperierender Institutionen im Browser verfügbar macht. Zwar erfreut sich dieses Verfahren aufgrund seiner Einfachheit großer Beliebtheit, allerdings zeigten die Einreichungen zu unserer vDHD-Sektion, dass die auf diese Weise zur Verfügung stehenden Ressourcen den Bedarfen, z. B. im Bereich des statistischen Lernens, schnell nicht mehr gerecht werden. Gleichzeitig steht der Nutzung institutioneller Kapazitäten aus dem High Performance Computing-Bereich (HPC) wie sich in unserem Falle erwiesen hat – nicht selten eine andere Dienstphilosophie entgegen. HPCs arbeiten mit definierbaren Aufgabenbeschreibungen, Ressourcenallokation und geplanten Abarbeitungen und sehen daher keine Benutzerinteraktion zum Zeitpunkt der Abarbeitung vor. Aus der Perspektive der Nachhaltigkeit gibt es nicht zuletzt auch immer wieder generelle Bedenken gegen die Verwendung von *Docker* und anderen Containerisierungsverfahren (Nüst et al. 2016). Festzuhalten bleibt auf jeden Fall, dass die Perspektive der Langzeitverfügbarkeit von ExP andere Lösungen erfordert, als die zur Zeit für die unmittelbare Verfügbarkeit gängigen Mittel. Weitere Probleme dieses Verfahrens sind

unzufriedenstellende Mittel der Absicherung copyright-geschützten Materials vor Zugriffen aus dem ExP, der Einbettung ausführbarer Bestandteile aus einem Notebook, geeignete Leseumgebungen und Webseiten von Verlagen (siehe auch Abbildung 2) sowie damit in Beziehung stehend, die Absicherung der Verlagsinfrastruktur vor Attacken, etwa durch *Code Injections*. Der Vortrag wird einen Überblick über die Melusina Press-Infrastruktur für die Realisierung von ExP im Rahmen des vDHD-Bandes geben und Stärken und Schwächen im Kontext der zuvor genannten Problemfelder diskutieren. Die Infrastruktur besteht im Wesentlichen aus der Nutzung der GESIS *myBinder*-Instanz, einem virtuellen Server zur technischen Isolierung der ausführbaren Publikationsabschnitte, einem virtuellen Server für die Präsentationsschicht, einer GitLab-Instanz der Universität Luxemburg und der *thebe*-Bibliothek (Abbildung 3).

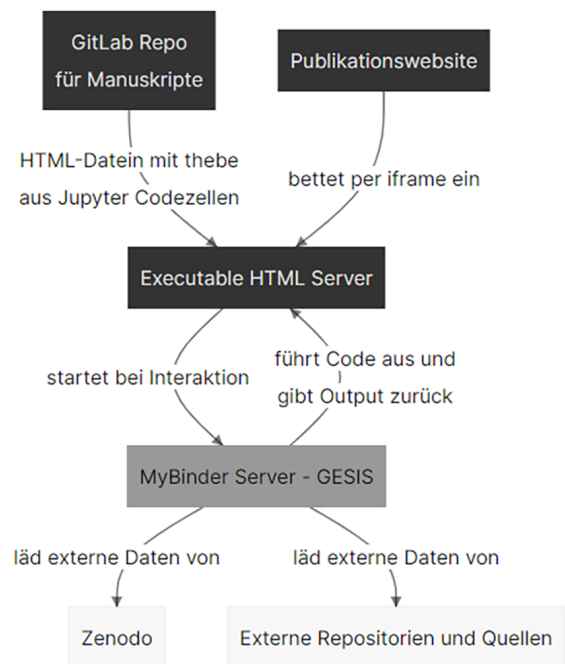


Abb. 3: Kerninfrastruktur von ExP auf Melusina

Zu Publikationsworkflows, Rechten, Versionierung und Reviewkriterien

Während es für viele technische Fragestellungen von ExP immerhin schon erste Lösungsvorschläge gibt und die Wahrnehmung potenzieller Mehrwerte der Entwicklung eher vorausgeht, wurde das Problemfeld der Publikationsworkflows bisher am wenigsten bearbeitet. Ob diesem Umstand mangelndes Bewusstsein gegenüber der Notwendigkeit ExP anders "bewerten" und "behandeln" zu müssen als traditionelle Forschungsartikel zugrunde liegt, oder ob sich Schwierigkeiten und Ausmaß dieser Perspektiven erst in konkreten Publikationssituationen von ExP so richtig identifizieren lassen, bleibt offen. Nichtsdestotrotz hat sich während der Konzeptualisierung des vDHD-Bandes und bei der Gestaltung von Autor:innenhandreichung zur Anfertigung der ExP schnell gezeigt, dass nicht nur mehrere Dimensionen, wie etwa Code, Interaktivität, multimodale Komposition, Organisation aller für die Lauffähigkeit notwendiger Ressourcen usw., zu bewer-

ten sind, sondern das in dem Moment wo man diese Mittel ermöglicht ebenfalls Kriterien für Einschränkung ihres Gebrauchs entwickelt werden müssen. Wie Lasser (2020) für das Element der Reproduktionsfähigkeit beschriebener Ergebnisse zeigt, ist mehr nicht immer besser – insbesondere für die Frage der "Lesbarkeit" von ExP. Daran schließt sich sogleich die Fragestellung an, was Lesbarkeit im Kontext von ExP überhaupt bedeuten soll: was ist eine lesbare multimodale Komposition, wie ist Interaktivität unter dem Gesichtspunkt der Lesbarkeit zu fassen, sind die Ansprüche an lesbaren Code dieselben wie in der traditionellen Softwareentwicklung? Wahrscheinlich sind diese und andere Fragen nur im Kontext von spezifischen Disziplinen und konkreten Forschungsbereichen beantwortbar. Da es bisher aber kaum ExP gibt, konnte ein solcher Entwicklungsprozess auch noch nicht wirklich stattfinden. Folgen hat dies sowohl für die Bewertbarkeit der Qualität von ExP, d. h. der Durchführbarkeit transparenter Reviewverfahren, als auch für die Infrastruktur und die Workflows der beteiligten Verlage, die die Gestalt von ExP in der Praxis antizipieren können müssen. Es ergibt sich aus der zuvor beschriebenen Situation von selbst, dass ein Standardisierungsprozess und die Entwicklung von Best Practices in Teilbereichen von ExP, wo dies sinnvoll und möglich erscheint, noch nicht eingesetzt hat. Es darf nicht unterschätzt werden, dass es sich hierbei eben nicht nur um eine Bedingung der Möglichkeit ihrer Realisierung handelt, sondern ebenso um die Grundlage dafür ExP zu einer vertrauensvollen Ressource, zu einer Währung mit Wert innerhalb der Wissenschaftskommunikation, werden zu lassen. Angrenzende Aktivitäten (Leipzig 2019; Sackmann 2020) finden bisher eher in informellen Kontexten statt, inkludieren meist nur ein Bruchteil der Akteursgruppen, die an solch einem Prozess beteiligt sein müssten, und versuchen zunächst eher das Feld zu kartographieren als es zu organisieren. Hervorgehoben werden kann "A guide to reproducible code in ecology and evolution" (Cooper & Hsing 2017), doch wie der Name anzeigt bewegen sich diese Aktivitäten stärker im *Reproducibility*-Bereich, der zwar viele Schnittmengen mit ExP aufweist, aber eben nicht mit ihnen deckungsgleich ist.

Durch die Diskussion von Entscheidungen und Vorgaben aus den Autor:innenhandreichung sowie der Illustration des Publikationsworkflows von ExP bei Melusina Press soll der Vortrag einen Beitrag zu der Entwicklung künftiger Best Practices und Standards leisten.

Bibliographie

- Agnone, Anthony** (2020): "Papers with Code + ArXiv = Reproducible, Organized" in: *Towards Data Science*. <https://towardsdatascience.com/papers-with-code-arxiv-reproducible-organized-research-f5404eb6a22e> [letzter Zugriff 9. Juli 2021]
- Aufreiter, Michael / Pawlik, Aleksandra / Bentley, Nokome** (2018): "Stencila – an Office Suite for Reproducible Research" in: *eLife*, July 2, 2018 https://elifesciences.org/labs/c496b8bb/stencila-an-office-suite-for-reproducible-research?utm_source=labworm&utm_medium=feed&utm_campaign=stencila.
- Brammer, Grant / Crosby, Ralph / Matthews, Suzanne / Williams, Tiffani** (2011): "Paper Mâché: Creating Dynamic Reproducible Science" in: *Procedia Computer Science, Proceedings of the International Conference on Computational Science* 4: 658–67 10.1016/j.procs.2011.04.069.
- Burg, Jennifer / Wong, Yue-Ling / Yip, Ching-Wan / Boyle, Anne** (2000): "The state of the art in interactive multimedia journals for academia" in: *Proceedings of EdMedia: World Conference on Educational Media and Technology 2000* 2: 37–42. Montréal: Association for the Advancement of Computing in Education <http://www.editlib.org/p/16036>.
- Ciepiela, Eryk / Harezlak, Daniel / Kasztelnik, Marek / Meizner, Jan / Dyk, Grzegorz / Nowakowski, Piotr / Bubak, Marian** (2013): "The Collage Authoring Environment: From Proof-of-Concept Prototype to Pilot Service" in: *Procedia Computer Science* 18: 769–78 10.1016/j.procs.2013.05.241.
- Chandre, Cristel / Dubois, Jonathan** (2021): "Notebook Articles: Towards a Transformative Publishing Experience in Non-linear Science" in: *Communications in Nonlinear Science and Numerical Simulation* 97: 105753 10.1016/j.cnsns.2021.105753.
- Cooper, Natalie / Hsing, Pen-Yuan** (2017): *A guide to reproducible code in ecology and evolution*. London: British Ecological Society.
- Gabriel, Ann / Capone, Rebecca** (2011): "Executable Paper Grand Challenge Workshop" in: *Procedia Computer Science* 4: 577–78 10.1016/j.procs.2011.04.060.
- Gorp, Pieter van / Mazanek, Steffen** (2011): „SHARE: a web portal for creating and sharing executable research papers" in: *Procedia Computer Science, Proceedings of the International Conference on Computational Science* 4: 589–97 10.1016/j.procs.2011.04.062.
- Thomas, Kluyver / Benjamin, Ragan-Kelley / Fernando, Pérez / Brian, Granger / Matthias, Bussonnier / Jonathan, Frederic / Kyle, Kelley** (2016): "Jupyter Notebooks – a publishing format for reproducible computational workflows" in: Loizides, Fernando / Schmidt, Birgit (eds.): *Positioning and Power in Academic Publishing: Players, Agents and Agendas* 87–90. Amsterdam: IOS Press 10.3233/978-1-61499-649-1-87.
- Knuth, Donald E.** (1984): "Literate Programming" in: *The Computer Journal* 27,2: 97–111 10.1093/comjnl/27.2.97.
- Kray, Christian, / Pebesma, Edzer / Konkol, Markus / Nüst, Daniel** (2019): "Reproducible Research in Geoinformatics: Concepts, Challenges and Benefits" in: Timpf, Sabine / Schlieder, Christoph / Kattenbeck, Markus / Ludwig, Bernd / Stewart, Kathleen (eds.): *COSIT2019* 142: 8:1–8:13 10.4230/lipics.cosit.2019.8.
- Lasser, Jana** (2020): „Creating an Executable Paper Is a Journey through Open Science" in: *Communications Physics* 3,1: 1–5 10.1038/s42005-020-00403-4.
- Leipzig, Jeremy** (2019): *Awesome Reproducible Research* 10.5281/ZENODO.3564746.
- Liu, Alan / Kleinman, Scott / Douglass, Jeremy / Thomas, Lindsay / Champagne, Ashley / Russell, Jamal** (2017): „Open, Shareable, Reproducible Workflows for the Digital Humanities: The Case of the 4Humanities.Org „WhatEvery1Says“ Project" in: *Digital Humanities 2017*. Conference Abstracts. <https://alanyliu.org/research/talks/2017-dh2017/we1s-dh2017-panel-abstract.pdf> [letzter Zugriff 7. Juli 2021].
- Maciocci, Giuliano / Aufreiter, Michael / Bentley, Nokome** (2019): "Introducing ELife's First Computationally Reproducible Article" in: *eLife* <https://elifesciences.org/labs/ad58f08d/introducing-elifes-first-computationally-reproducible-article> [letzter Zugriff 7. Juli 2021].
- Nüst, Daniel / Konkol, Markus / Pebesma, Edzer / Kray, Christian / Klötgen, Stephanie / Schutzzeichel, Marc / Lorenz, Jörg / Przibytzin, Holger / Kussmann, Dirk** (2016): „Opening Reproducible Research" in: *EGU General Assembly Conference Abstracts* <http://adsabs.harvard.edu/abs/2016EGUGA..18.7396N> [letzter Zugriff 7. Juli 2021].

Odehahn, Andrew (2015): „Embracing Jupyter Notebooks at O'Reilly" in: O'Reilly Media Blog <https://beta.oreilly.com/ideas/jupyter-at-oreilly> [letzter Zugriff 7. Juli 2021].

Pebesma, Edzer / Nüst, Daniel / Bivand, Roger (2012): „The R Software Environment in Reproducible Geoscientific Research" in: *Eos, Transactions American Geophysical Union* 93,16: 163–163 10.1029/2012EO160003 .

Roure, David De (2014): "Executable Music Documents" in: *Proceedings of the 1st International Workshop on Digital Libraries for Musicology - DLFM '14*, 2014, 1–3 10.1145/2660168.2660183.

Singh, Ripudaman / Chudoba, Rostislav / Gopal, K. / Koenke, Carsten (1998): „IMMJ: interactive multi-media journals in science and technology" in: *Ejournal* 8, 2.

Smith, Vincent / Georgiev, Teodor / Stoev, Pavel / Biserkov, Jordan / Miller, Jeremy / Livermore, Laurence / Baker, Edward (2013): „Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal" in: *Biodiversity data journal* 1: e995 10.3897/BDJ.1.e995.

Walsh, Melanie (2021): Introduction to Cultural Analytics & Python. Ithaca, NY 10.5281/zenodo.4411250.

Fluch und Segen der Visualisierung

Unterschiedliche Zielfunktionen im Forschungsprozess der historischen Netzwerkanalyse

Balck, Sandra

balcksaa@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Menzel, Sina

menzel@ub.fu-berlin.de
Freie Universität Berlin, Germany

Petras, Vivien

vivien.petras@ibi.hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Schnaitter, Hannes

hannes.schnaitter.1@ibi.hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Zinck, Josefine

josefine.zinck.1@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

In einer Interviewstudie mit sieben Forschenden der historischen Netzwerkanalyse (HNA) wurden die wichtigsten Einsatzgebiete von Visualisierungen im Forschungsprozess identifiziert: Theorieentwicklung und Datenexploration, Datenqualitätsüberprüfung, Analyse sowie Präsentation der Ergebnisse. Die Diskussion der Visualisierungen zeigt ein zwiespältiges Verhältnis

der Community: sie werden von den Forschenden sehr differenziert betrachtet, sowohl wenn es um den Zeitpunkt ihres Einsatzes im Forschungsprozess geht und wer für die Entwicklung der Visualisierungen zuständig sein sollte, aber auch, ob diese der Präsentation der Forschungsergebnisse in der historischen Forschungsgemeinschaft zuträglich sind oder nicht. Dabei werden zwei Zielfunktionen unterschieden: Visualisierungen für die Exploration bzw. Analyse und erläuternde Visualisierungen. Entscheidend für die Akzeptanz von Visualisierungen ist deren Dokumentation und Kontextualisierung einschließlich der in ihnen enthaltenen Daten, um eine verlässliche Grundlage für die Forschung zu gewährleisten.

Einleitung

Visualisierungen können ein wichtiges Werkzeug in der Analyse von Daten sowie in der Kommunikation von Forschungsergebnissen sein. Einerseits bieten sie die Möglichkeit, schnell einen Überblick über die gesammelten Daten zu erlangen und Fehler oder interessante Fälle zu finden. Sie ermöglichen es den Auswertenden, große Datenmengen ganzheitlich zu betrachten und zu interpretieren. Sie sind aber auch schnell unübersichtlich, ihre Interpretation unterliegt kulturell gelernten Biases und verständliche Visualisierungen sind aufwendig zu erstellen. Diese Dichotomie spiegelt sich auch in der Bewertung von Visualisierungen im Forschungsprozess wider: sie werden in vielen Schritten des Forschungsprozesses genutzt und können dort Mehrwert liefern. Gleichzeitig wird die Nutzung oft als problematisch angesehen und insbesondere die Erstellung guter Visualisierungen oft nicht als eigenständige Forschungsleistung gewürdigt.

In diesem Beitrag berichten wir über die Ergebnisse einer von uns im Rahmen des DFG-Projekts SoNAR (IDH) (siehe Bludau et al. 2020) durchgeführten Interviewstudie mit Forschenden aus der historischen Netzwerkanalyseforschung (HNA). Eine der Leitfragen, welche auf der Zielstellung des Projekts basiert, das Netzwerkvisualisierungen für historische Daten entwickeln soll, ist: welchen Wert haben Visualisierungen als Werkzeug für Analyse und Präsentation von Informationen in der Forschungsgemeinschaft der HNA? In der Beschreibung der Interviewstudie und deren Ergebnisse gehen wir besonders auf die Sicht der Forschenden auf Visualisierungen ein, die ein zwiespältiges Verhältnis aufzeigt.

Interviewstudie zur Historischen Netzwerkanalyse

Ziel der Interviewstudie war es, Forschungsprozesse und -fragen der Historischen Netzwerkanalyse zu beschreiben und Anforderungen an eine digitale Informationsinfrastruktur zu identifizieren. Dafür wurden Expert:innen zum Vorgehen bei eigenen Forschungsarbeiten im Bereich der HNA in Form qualitativer Leitfadeninterviews befragt.

Es wurden sieben Interviews mit Expert:innen der HNA durchgeführt, davon waren zwei männlich und fünf weiblich. Die Interviewten kamen aus den Bereichen Sozialwissenschaft, Geschichtswissenschaft und Wirtschaftswissenschaft, was die Interdisziplinarität des zu untersuchenden Gebiets untermauert. Die Interviews wurden im Zeitraum Juli-August 2020 durchgeführt und dauerten zwischen 37:08 und 51:28 Minuten. Als Expert:innen gelten Personen, die wissenschaftliche Arbeiten im Bereich HNA vorweisen.

Die Befragungen wurden remote über die Videokonferenzlösung Zoom durchgeführt, mit Einverständnis der Proband:innen aufgezeichnet und anschließend mittels MAXQDA transkribiert und codiert. Es folgte eine thematische Analyse, hierbei gebildete Kategorien dienten als Grundlage für die Auswertung. Insgesamt wurden 126 Kategorien gebildet, die in sechs Oberkategorien zusammengefasst wurden:

- **Community:** Disziplinäres Selbstverständnis bezogen auf wissenschaftliche und disziplinäre Infrastrukturen, Wissenskommunikation sowie die Einbindung digitaler Methoden in historische Curricula.
- **Netzwerke:** Einschätzungen über die Verbindung der Netzwerkanalyse mit der historischen Forschung im allgemeinen und Definitionsversuche/Einordnungen der HNA durch die Interviewten.
- **Forschungsprozess:** Äußerungen zum Vorgehen bei der Bearbeitung eines Forschungsvorhabens mittels historischer Netzwerkanalyse.
- **Werkzeuge:** Für die Datenaufbereitung und -speicherung verwendete Software, ihre Vor- und Nachteile sowie Anforderungen an zukünftige Systeme.
- **Datenqualität:** Datenqualität ist als besondere Herausforderung der HNA zu verstehen, der Fokus liegt hier auf Anforderungen und Probleme bezogen auf historische Daten.
- **Visualisierung:** Bezieht sich auf die Anwendung von Visualisierungen und dem daraus resultierenden Erkenntnisgewinn sowie Kritik und Herausforderungen im Umgang mit historischen Daten.

Die Auswertung der Interviews gibt einen Überblick über exemplarische Forschungsprozesse und deren Aufbereitung in einzelne Teilschritte sowie erste Erkenntnisse zu Anforderungen an Visualisierungen für die HNA in verschiedenen Disziplinen, auf die wir uns im Folgenden besonders fokussieren.

Der Forschungsprozess in der Historischen Netzwerkanalyse

“Die Geschichtswissenschaft beschäftigt sich mit der Analyse des menschlichen Zusammenlebens in der Vergangenheit und bemüht sich, Ereignisse und Entwicklungen aus der jeweiligen Zeit heraus zu verstehen und zu deuten. [...] Historische Akteure werden [...] immer auch als Kontext der sie umgebenden Strukturen betrachtet.“ (Düring 2015: 337). Diese Strukturen manifestieren sich in sozialen Beziehungen. Die ursprünglich aus den Sozialwissenschaften kommende formale Netzwerkanalyse erlaubt es, diese nicht immer offensichtlichen Strukturen zu analysieren, indem sie es ermöglicht, diese “präzise zu beschreiben, zu verstehen, wie sie geschaffen wurden und welche Folgen sie haben.” (Lemerrier 2012: 21). Die HNA ist eine Methode, die die Interpretation von historischen Strukturen unterstützt und somit als interdisziplinäres Zusammenspiel von historischer Erzählung und formaler Netzwerkanalyse beschrieben werden kann.

Die HNA setzt sich aus Komponenten der Sozialwissenschaft und Geschichtswissenschaft und damit verbunden den Digital Humanities zusammen. Der Forschungsprozess dieser Disziplinen ähnelt sich in vielen Bereichen. Bhattacharjee (2012: 20) unterteilt den sozialwissenschaftlichen Forschungsprozess in drei Phasen, welche wiederum in neun Unterphasen aufgeteilt werden. Im Vergleich dazu nennen Burghardt et al. (2014: 2) für die Geisteswissenschaften acht Phasen (siehe Tab. 1). Ausgenommen von “Kommunikation und Kollaboration”, was im weitesten Sinne

auch unter Theoriebildung gefasst werden kann, sind keine Unterschiede zum sozialwissenschaftlichen Forschungsprozess zu erkennen.

Tab. 1: Modelle des Forschungsprozesses in angrenzenden Disziplinen

Bhattacharjee (2012) - Sozialwissenschaften	Burghardt et al. (2014) - Geisteswissenschaften
exploration • research question, literature review, theory research design • operationalization, research method, sampling strategy research execution • pilot testing, data collection, data analysis + research report	• Kommunikation und Kollaboration • Recherche • Konzeptualisierung • Datenerhebung • Datenaufbereitung • Datenauswertung • Verschriftlichung • Veröffentlichung

Der mittels Aussagen der Proband:innen beschriebene Forschungsprozess der HNA kann insgesamt in sechs Phasen unterteilt werden, wobei Datenauswahl- und -erhebung sowie Datenaustausch- und -publikation zusammengefasst wurden (siehe Abb. 1): Theorieentwicklung, Datenauswahl- und -erhebung, Qualitätssicherung, Datenanalyse, Dokumentation, Datenaustausch und -publikation.



Abb. 1: Forschungsprozess in der Historischen Netzwerkanalyse

Hierbei lassen sich viele Parallelen zu den oben beschriebenen allgemeinen Forschungsprozessen der Geistes- sowie Sozialwissenschaften erkennen. Im Unterschied zu den anderen Studien sind zwei zusätzliche Prozessabschnitte stärker herausgehoben: Qualitätssicherung und Dokumentation. Diese zwei Phasen des Forschungsprozesses sind stark miteinander verbunden und weisen auf die Wichtigkeit von Transparenz innerhalb des Forschungsprozesses der HNA sowie die Schwierigkeiten im Umgang mit historischen Daten, bezogen auf die Datenqualität, hin. Allen vorgestellten Modellen des Forschungsprozesses ist gemein, dass dieser nicht in streng linearen Pfaden verläuft, sondern durch ständige Feedbackschleifen charakterisiert ist. Die wurden in Abbildung 1 nicht durch Pfeile dargestellt, in den Interviews aber immer wieder charakterisiert.

Visualisierung im HNA Forschungsprozess

Innerhalb des HNA-Forschungsprozesses nehmen Visualisierungen eine Sonderstellung ein: sie können am Anfang des Forschungsprozesses eingesetzt werden, um auf neue Forschungsfragen aufmerksam zu werden, sie können bei der Analyse hilfreich sein oder auf Fehler aufmerksam machen. Allgemein werden

Visualisierungen im Zusammenhang mit folgenden Prozessabschnitten erwähnt:

Hypothesenbildung:

Forschungsfragen werden von HNA-Forschenden sowohl theorie- als auch datengeleitet entwickelt. Im Laufe des Forschungsprozesses können sich Forschungsfragen durch neue Impulse der Datenanalyse verändern. Visualisierungen, hier als Teil des Analyseprozesses, können durch explorative Vorgänge dazu beitragen, neue Thematiken zu entdecken: “[...] es [ist] auch ok, da mit einer nicht klaren Fragestellung ranzugehen, sondern auch explorativ zu sagen, ich schau mal, was da drin ist” (Interview 5). Strukturen werden so sichtbar und können Ausgangspunkt für konkrete Forschungsfragen bzw. die Hypothesenentwicklung sein.

Qualitätssicherung:

Die Qualitätssicherung läuft begleitend zur Datenerhebung und -analyse und kann sich im Laufe des Prozesses wiederholen. Neben intellektueller Durchsicht und statistischen Berechnungen spielen auch Visualisierungen eine Rolle bei dem Aufspüren von Fehlern. Diese können auch erst nach der eigentlichen Qualitätssicherung, im Prozess der Datenanalyse/Visualisierung, sichtbar werden: “[...] es sind immer noch Fehler drinnen, das ist das Schöne, das fällt einem dann bei der Netzwerkanalyse, wenn man es dann visualisiert, teilweise erst auf, dass da ein Knoten absolut unverbunden ist [...]” (Interview 3).

Datenanalyse:

Die meisten Interviewten erwähnen Visualisierungen im Zusammenhang mit der Datenanalyse. Die Netzwerkanalyse macht durch die Gewichtung von Beziehungen Strukturen innerhalb eines Netzwerks sichtbar: “ich habe durch eine Visualisierung zwei Knoten gefunden, die relevant sind und zwar nur durch die Visualisierung” (Interview 6). Visualisierungen werden genutzt, um übergeordnete Muster zu erkennen und so die Interpretation zu unterstützen. In den meisten Fällen dienen sie dazu, zeitliche und geografische Entwicklungen sozialer Netzwerke sichtbar zu machen und dazu verschiedene Schichten – betreffend Person, Ort und Zeit – in Verbindung miteinander darzustellen. Erwähnt werden in diesem Zusammenhang die Identifikation von Netzwerkeigenschaften wie Zentralität, Dichte und Hubs.

Ergebnispräsentation:

Visualisierungen werden eingesetzt, um Forschungsergebnisse in Vorträgen oder Artikeln zu vermitteln: “das ist halt das Schöne an den Netzwerken, dass man da manchmal doch sehr komplizierte Sachverhalte ein bisschen unterbrechen kann und sich ein Bild häufig sehr sehr viel leichter erklären lässt, als ne umständliche Erklärung über acht Seiten, die manchmal dieses Bild bräuchte, um es in den Köpfen der Leser entstehen zu lassen” (Interview 3). Visualisierungen werden hierbei im Anschluss an den eigentlichen Forschungsprozess entwickelt und dienen lediglich der visuellen Untermalung von Analysen.

Abbildung 2 fasst die Funktionen von Visualisierungen im HNA-Forschungsprozess visuell zusammen.



Abb. 2: Verwendung von Visualisierungen innerhalb des Forschungsprozesses der Historischen Netzwerkanalyse

Fluch oder Segen der Visualisierung?

Visualisierungen werden von den Forschenden sehr differenziert betrachtet, sowohl wenn es um den Zeitpunkt ihres Einsatzes im Forschungsprozess geht und wer für die Entwicklung der Visualisierungen zuständig sein sollte, aber auch, ob diese der Präsentation der Forschungsergebnisse in der historischen Forschungscommunity zuträglich ist oder nicht.

Visualisierungen werden von einer eher quantitativ sozialwissenschaftlich arbeitenden Gruppe bevorzugt selbst gebaut und erst im Anschluss an die Analyse verwendet, um eine Beeinflussung durch bildliche Verzerrungen zu verhindern: “also Netzwerke würde ich mir immer selber bauen [...]” (Interview 3) oder auch: “für mich sind die Netzwerkvisualisierungen erst das Ende des ganzen Analyseprozesses. Also ich weiß, dass man mit Visualisierungen das menschliche Auge sehr verwirren kann und in Richtungen lenken will, die vielleicht die Daten gar nicht hergeben [...]” (Interview 5).

Die andere Gruppe nutzt Visualisierungen dagegen bevorzugt während des Forschungsprozesses und erwartet von automatisch erstellten Datenrepräsentationen bessere Einsichten, die nur durch die Zusammenstellung überhaupt aufgedeckt werden können: “um etwas zu sehen, visualisiert zu bekommen, was ich vorher noch nicht wusste und so noch nicht ... mir noch nicht zugänglich war durch irgendwelche Einzelinformationen, sondern diese Kumulation, dieses Aggregat quasi nur sichtbar werden kann” (Interview 4). Diese Gruppe verwendet Visualisierungen auch für die Darstellung von Forschungsergebnissen in Veröffentlichungen.

Andere verwenden hingegen Visualisierungen lediglich für den eigenen Erkenntnisgewinn und nicht zur Darstellung von Forschungsergebnissen: “[...] immer, wenn man ein Bild also so eine Grafik zeigt, muss man die erstmal eine halbe Stunde erklären und sagen, was sie jetzt nicht zeigt und was sie zeigt und warum der Algorithmus jetzt den einen Punkt in die Mitte macht und der andere am Rand, aber das eigentlich keine Wertung ist. Bilder zeigen ist sehr komplex und ich bin eher davon abgekommen” (Interview 6).

In mehreren Interviews wird eine Skepsis gegenüber Visualisierungen und der manipulativen Kraft von Bildern deutlich. Genau wie der Datenbestand, der der Visualisierung zugrunde liegt, ist auch die Visualisierung selbst nur eine Repräsentation und keine Replikation der Realität und immer mit einer Reduktion der Komplexität verbunden. (vgl. Freyberg 2020: 1; Drucker, 2014; Kasunic & Sweetapple, 2014) Genau diese Reduktion wird kritisiert, da es Visualisierungen zu „höchst interpretierbedürftige[n] Quel-

len“ (Interview 7) macht und besonders in der öffentlichen Diskussion intensiver Erklärung bedarf. Kritische Reaktionen sind auch aus der “klassischen” geschichtswissenschaftlichen Community zu vernehmen: „[...] es kam dann immer so der Einwand: Ja, das ist ja spannend und sieht irgendwie cool aus, aber was hast du davon?“ (Interview 7).

Zusammenfassung und Ausblick

Für eine angemessene Nutzung in der Forschung muss die Rolle der Visualisierung klar sein. Dabei können zwei Zielfunktionen unterschieden werden: Visualisierungen für die Exploration bzw. Analyse und erläuternde Visualisierungen. Visualisierungen zur Exploration bieten den Betrachter:innen durch die Wahl passend gewählter Abstraktionsebenen und graphischer Elemente die Möglichkeit der Interpretation des Ganzen und des Auffindens interessanter Teilaspekte. Sie unterstützen die Analyse durch die Aggregation von individuellen Merkmalen in den Daten. Eine erläuternde Visualisierung soll den Betrachter:innen einen bestimmten Sachverhalt darstellen und eine bestimmte Erkenntnis kommunizieren. Sie unterstützen insbesondere die kondensierte Darstellung von komplexen Sachverhalten für die Repräsentation und den Austausch mit anderen Forschenden.

Wichtig für beide Zielfunktionen ist die Kontextualisierung von Visualisierungen und den in ihnen enthaltenen Daten sowie die Transparenz durch Dokumentation der zugrundeliegenden Datenbankstrukturen, um eine verlässliche Grundlage für die Forschung zu gewährleisten.

Basierend auf sieben Expert:inneninterviews kann kein allgemeines Bild für die gesamte Forschungscommunity gezeichnet werden. In der vorliegenden Studie konnten jedoch interessante Muster zwischen verschiedenen Gruppen entdeckt werden, die in einer späteren Nutzer:innenstudie zum SoNAR-Prototyp bestätigt wurden (vgl. Schnaitter et al., 2021). Weitere ausführliche Studien müssen jedoch folgen, um diese Ergebnisse zu validieren.

Die HNA Community sieht zwar grundsätzlich nützliche Unterstützungsmöglichkeiten durch die Visualisierung, zweifelt aber gleichzeitig an deren Interpretationskraft sowie deren Bedeutung als Erklärungswerkzeuge in der Forschungskommunikation. Dies wurde mittlerweile in vielen Studien gezeigt und auch in dieser bestätigt. Dies mag einerseits an fehlenden Standards für Darstellungsformen und Datendokumentation liegen, andererseits aber auch der gebotenen Skepsis gegenüber der Interpretationsfähigkeit von historischen Quellen geschuldet sein, die der historischen Forschung unterliegen muss. Es ist eine interessante Frage, ob Fortschritte in der Datendokumentation und -aufbereitung diese Zwiespältigkeit auflösen können.

Bibliographie

Bhattacharjee, Anol (2012): *Social Science Research: Principles, Methods, and Practices*, Textbooks Collection 3: 20 https://scholarcommons.usf.edu/oa_textbooks/3.

Bludau, Mark-Jan / Dörk, Marian / Halling, Thorsten / Leitner, Elena / Menzel, Sina / Müller, Gerhard / Petras, Vivien / Rehm, Georg / Neudecker, Clemens / Zellhöfer, David / Moreno-Schneider, Julian (2020): "SoNAR (IDH): Datenschnittstellen für Historische Netzwerkanalyse", in: *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. 7. Tagung des Verbands "Digital Humanities im

deutschsprachigen Raum" (DHd 2020), Paderborn <https://doi.org/10.5281/zenodo.4621862>.

Burghardt, M. / Schubert, A. / Traber, M. / Wolff, C. (2014): "Empirische Untersuchung zu digitalen, geisteswissenschaftlichen Arbeitspraktiken an der Universität Regensburg", in: *Online-Proceedings der 1. Jahrestagung der "Digital Humanities im deutschsprachigen Raum"* <http://doi.org/10.5283/epub.35713>.

Drucker, Johanna (2014): "*Graphesis: Visual forms of knowledge production*" Cambridge, MA: Harvard University Press.

Düring, Marten / von Keyserlingk, Linda (2015): "Netzwerkanalyse in den Geschichtswissenschaften. Historische Netzwerkanalyse als Methode für die Erforschung von historischen Prozessen", in: Schützeichel, Rainer / Jordan, Stefan (Hrsg.): *Prozesse - Formen, Dynamiken, Erklärungen*. Berlin: Springer 337-350 https://doi.org/10.1007/978-3-531-93458-7_15.

Freyberg, Linda (2020): "Ikonizität als Erkenntnismittel – Vollständigkeit, Verständlichkeit und Kontextualisierung als Grundprinzipien der Visualisierung", in: *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. 7. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2020), Paderborn <https://doi.org/10.5281/zenodo.4621742>.

Kasunic, Jacqueline Lorber / Sweetapple, Kate (2014): "Visualizing Texts A design practice approach to humanities data." In: *DRHA2014 Conference Digital Research in the Humanities and Arts Theme: Communication Futures: Connecting interdisciplinary design practices in arts/culture* <https://core.ac.uk/download/pdf/42391357.pdf#page=88>

Lemercier, Claire (2012): "Formale Methoden der Netzwerkanalyse in den Geschichtswissenschaften: Warum und Wie?", in: *Österreichische Zeitschrift für Geschichtswissenschaften* 23(1): 16-41 <https://doi.org/10.25365/oezg-2012-23-1-2>.

Schnaitter, Hannes / Balck, Sandra / Zinck, Josefine / Petras, Vivien (2021): „SoNAR (IDH) AP4-5 Evaluierung IV: Nutzer:innenstudie“ *Interner Bericht*. <https://github.com/sonar-idh/reports/blob/main/AP4-HU-4-5-3-Evaluierung-IV.pdf>

Forschendes Lernen digital

Bläß, Sandra

sandra.blaess@uni-hamburg.de
Universität Hamburg, Germany

Flüh, Marie

marie.flueh@uni-hamburg.de
Universität Hamburg, Germany

Gerstorfer, Dominik

gerstorfer@linglit.tu-darmstadt.de
Technische Universität Darmstadt

Gius, Evelyn

evelyn.gius@tu-darmstadt.de
Technische Universität Darmstadt

Meister, Malte

meister@linglit.tu-darmstadt.de
Technische Universität Darmstadt

Nantke, Julia

julia.nantke@uni-hamburg.de
Universität Hamburg, Germany

Schumacher, Mareike

schumacher@linglit.tu-darmstadt.de
Technische Universität Darmstadt

Gerade Projekte aus dem Bereich der digitalen Literaturwissenschaften bringen vielseitige Ergebnisse und Forschungsdaten ganz unterschiedlicher Art hervor, die sich in die Lehre integrieren lassen und so einen Brückenschlag zwischen geisteswissenschaftlicher Forschung und Lehre ermöglichen. Während im angloamerikanischen Sprachraum unter dem Oberbegriff „DH-Pedagogy“ eine lebendige Debatte über didaktisch fundierte Lehr-Lern-Szenarien in den Digital Humanities geführt wird, (Hirsch 2012) sind systematische Überlegungen dieser Art im deutschen Sprachraum noch eher selten zu finden.

In unserem Beitrag schildern wir aus der Erfahrung der konkreten Forschungs- und Lehrpraxis, auf welche Art und Weise sich das DH-Forschungs- und Editionsprojekt *Dehmel digital* sowie die Lern- und Disseminationsplattform *forTEXT* in der DH-Lehre fruchtbar kombinieren lassen, um ein digitales Lehren und Lernen zu ermöglichen, das Grundideen des Forschenden Lernens aufgreift und weiterdenkt. Wir stellen vor, wie durch die Dokumentation und Aufbereitung der Forschungsprozesse für die Plattform *forTEXT* gleichzeitig eine nachhaltige Nutzung der etablierten Verfahren und produzierten Forschungsdaten in anderen Lehr-Lern-Szenarien ermöglicht wird.

Forschendes Lernen als Leitgedanke

Im hochschuldidaktischen Diskurs kommen die Begriffe „Forschendes“, „Forschungsbasiertes“, „Forschungsorientiertes“ oder „Forschungsnahes Lernen“ nach wie vor vielseitig und häufig zum Einsatz. In der internationalen Diskussion um Forschendes Lernen haben sich vor allem Systematisierungsansätze nach Hearley und Jenkins (2009) und Huber (2009 und 2014) etabliert, die in aktuellen und auf die Hochschulbildung im deutschsprachigen Raum bezogenen Publikationen aufgegriffen werden. Beide Ansätze definieren zwei Dimensionen des Forschenden Lernens: die inhaltliche Ausrichtung des Forschungsangebots (ergebnisorientiert und prozessorientiert) und die Form der Studierendenbeteiligung (passives Rezipieren und aktive Gestaltung). Darauf aufbauend lassen sich drei Kategorien unterscheiden, auf die der Lernfokus bei Ansätzen des Forschenden Lernens gelegt werden kann: auf den Forschungsprozess, die Forschungsmethode und die Forschungsergebnisse. Die Studierenden werden je nach Fokus rezeptiv aktiv, wenden erworbenes Wissen an und/oder werden selbst forschend tätig (Wulf et al. 2020).

Diverse Universitäten oder Fachbereiche bestimmen Forschendes Lernen als didaktisches Leitprinzip für die Lehre. Konsens besteht bisher vor allem darüber, dass es im Lehrprofil von Hochschulen einen festen Platz einnehmen sollte. Ob aus den Zielen konkrete Maßnahmen folgen, hängt schlussendlich am Engagement von Einzelpersonen. Eine curriculare Verankerung ist selten der Fall (Huber 2020). Dieser Widerspruch aus proklamiertem Leitprinzip und eher vereinzelter praktischer Umsetzung kann auch im vorliegenden Beitrag nicht aufgelöst werden. Stattdessen zeigen wir ein beispielhaftes Lehr- und Lernszenario, das For-

schungsnahes Lernen im Rahmen der Digital Humanities praktisch austestet.

Das Erlernen digitaler Verfahren erfolgt am effektivsten möglichst praxisnah. Gerade digitale Tools und Methoden bieten deshalb Möglichkeiten, um Forschendes Lernen erfolgreich umzusetzen (Schirmer und Martin 2020). Beiträge, in denen digitale Literaturwissenschaft und Ansätze dieser Lernform systematisch zusammengebracht werden, sind allerdings selten. Eine fachspezifische Form des Forschenden Lernens ist aufgrund unterschiedlicher Forschungsformen, -begriffe und -gegenstände (trotz der domänenübergreifenden Forschungstätigkeiten des Beobachtens, Beurteilens, Modellierens und Konstruierens) allerdings wichtig. Gerade für die in geisteswissenschaftlichen Forschungs- und Lernszenarien entscheidende Begriffs- und Theoriebildung sowie die Fähigkeit zur hermeneutischen Interpretation gilt ein Zugang über Forschendes Lernen bislang als schwierig; im domänenübergreifenden Diskurs über Forschendes Lernen sind Beiträge aus den Geisteswissenschaften unterrepräsentiert (Huber 2017). Hier finden sich mittlerweile aber auch gelungene Gegenbeispiele (Hethy und Struve 2017, Mieg 2020), an die dieser Beitrag direkt anschließt. Die wechselseitige Kooperation und die gemeinsame Integration von Ansätzen des Forschenden Lernens in den Projekten *Dehmel digital* und *forTEXT* hat sich dabei als besonders fruchtbar erwiesen.

Projekt Dehmel digital

Das Forschungs- und Editionsprojekt *Dehmel digital* hat die sukzessive materielle Erschließung und inhaltliche Erforschung des Korrespondenznetzwerks von Richard und Ida Dehmel zum Ziel, die um 1900 das Zentrum eines europaweiten Netzwerks von Künstler:innen und Kulturschaffenden bildeten. Der umfangreiche Briefnachlass des Ehepaars (insgesamt ca. 35.000 Briefe) liegt hauptsächlich in handschriftlicher Form vor. Im Rahmen des Projekts werden die Briefe digitalisiert und mit quantitativen computationalen Verfahren erschlossen. Die digitalisierten Briefe werden auf einer digitalen Plattform für unterschiedliche Zielgruppen (Wissenschaftler:innen, Studierende und interessierte Lai:innen) aufbereitet und zur Nachnutzung zur Verfügung gestellt. Im Rahmen des Projekts kommen verschiedene Verfahren der digitalen Manuskriptanalyse und quantitativen inhaltlichen Texterschließung zum Einsatz, die bereits in *forTEXT* als Methoden und/oder Lerneinheiten integriert sind. Gleichzeitig ist das Thema der digitalen Erschließung in *forTEXT* bislang zwar in Methoden-, Tool- und Lerneinheitsbeiträgen zur Anwendung von *Transkribus* aufgegriffen, aber noch kaum auf den Bereich der Edition bezogen, der aktuell noch nicht explizit repräsentiert ist.

Eine Verknüpfung zwischen dem Forschungsprojekt *Dehmel digital* und der universitären Lehre findet an der Universität Hamburg seit 2019 in verschiedenen Seminarformaten statt. Die Integration in die Lehre enthält im Sinne des Forschenden Lernens ergebnisorientierte sowie prozessorientierte Aspekte und sowohl passiv-rezipierende als auch aktiv-gestaltende Elemente. In den Lehrveranstaltungen lernen die Studierenden unterschiedliche Arbeitsphasen des Erschließungsprojekts kennen. Auch die *forTEXT*-Lehr- und Lernmaterialien werden seit Projektbeginn im Jahr 2018 in der universitären Lehre erprobt. Auf diese Weise kann das Feedback der Studierenden kontinuierlich in die Weiterentwicklung der Disseminationsplattform einbezogen werden. Die enge Zusammenarbeit mit dem Projekt *Dehmel digital* ermöglicht darüber hinaus erstmalig die Verbindung mit einem laufenden Forschungsprojekt und damit die Umsetzung eines sitzungsübergreifenden Lehr- und Lernszenarios des Forschenden Lernens.

Gerade die Kombination aus konkretem Forschungsprojekt und generischer Disseminationsplattform ermöglicht dabei eine strukturierte und nachhaltige Gestaltung der verschiedenen Komponenten und ist dazu geeignet, bisherige Konzepte des Forschenden Lernens um literaturwissenschaftliche Anwendungsszenarien zu erweitern.

Das forTEXT-Projekt

forTEXT ist ein Disseminationsprojekt, das Interessierten einen Einstieg in die Digital Humanities ermöglicht. Über die Homepage fortext.net (vgl. Gius et al. 2021) werden zitierfähige Methodenbeschreibungen, Textsammlungen und Tools verfügbar gemacht, die niedrighschwellige Einführungen vor dem Hintergrund der nicht-digitalen Geisteswissenschaften geben. Thematisch reichen diese von Digitalisierung über Annotation zu Interpretation und Visualisierung. Die präsentierten Materialien sind unterteilt in Routinen, Ressourcen und Tools. Sie werden aus literaturwissenschaftlicher Perspektive bewertet und – zum Teil mit der Hilfe von Videos – erklärt.

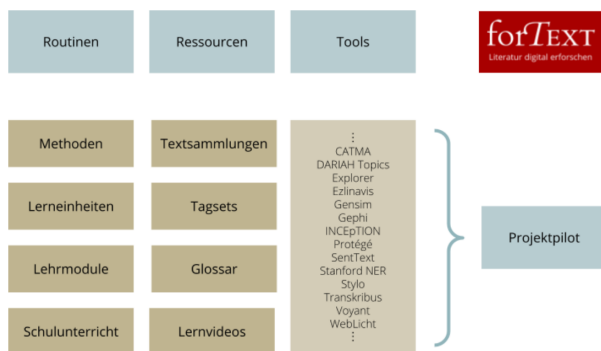


Abb. 1: *forTEXT*-Inhalte im Überblick

Die Sammlung der *forTEXT*-Lehr- und Lernmaterialien unterstützt einen aktiven Forschungsprozess in jeder Phase (vgl. Abb. 1). Die Inhalte der Plattform haben einen starken methodischen Fokus und sind darum besonders geeignet, um prozessorientierte und methodische Aspekte des Forschenden Lernens abzudecken. Die Beitragskategorien sind entweder passiv-informativ oder für das aktive Lernen im ‚Hands-on‘-Modus konzipiert, sodass auch unterschiedliche Formen der Beteiligung sich ergänzen. Unter *Routinen* finden sich z.B. Methodeneinträge und Lerneinheiten, mithilfe derer Forschende und Studierende mit einem Interesse für bestimmte Methoden der Digital Humanities sich diese theoretisch sowie praktisch aneignen können. Sind Methodenkenntnisse und Kompetenzen der Toolnutzung vorhanden, rücken *Ressourcen* ins Zentrum des Interesses. Je nach gewählter Methode werden z.B. kleinere oder größere Korpora benötigt, die in einschlägigen *Textsammlungen* zu finden sind, die jeweils in einem eigenen Beitrag vorgestellt werden. Während Nutzer:innen sich durch die kombinierte Rezeption von Methodenbeiträgen, Lerneinheiten und Ressourcen die Basiskompetenzen zur Nutzung einer Methode aneignen, werden im Bereich *Tools* nützliche und in den Digitalen Geisteswissenschaften häufig verwendete Softwares ausführlich vorgestellt und hinsichtlich der Eignung für DH-Neulinge bewertet.

Gerade für Einsteiger:innen in die Digital Humanities ist die Frage, welche *Routinen*, *Ressourcen* und *Tools* überhaupt für die

eigene Forschung geeignet wären, nicht leicht zu beantworten. Darum gibt es den *forTEXT-Projektpiloten* – einen interaktiven Fragebogen, der den Weg in die eigenständige Projektarbeit bereitet. Studierende, die im Seminarkontext das Forschungsprojekt *Dehmel digital* kennengelernt haben, gehören auch zur primären Zielgruppe von *forTEXT*, da sie (teilweise) schon darin geübt sind, eigene Forschungsprojekte zu planen, aber oft noch nicht darin, digitale Methoden darin einzusetzen. Über den Projektpiloten, also in einem aktiven Frage-Modus, oder über die theoretischen Einträge der *forTEXT*-Webseite und damit eher passiv-rezeptiv – die Studierenden können sich bedarfsgenau und auf ihr konkretes Forschungsvorhaben konzentriert in die digitalen Geisteswissenschaften einarbeiten. Dabei wird die Erschließung und Erforschung des Dehmel-Korrespondenznetzwerks didaktisch mit der Vermittlung von Methodenwissen und -kritik sowie der Verwendung von Tools verschränkt und auch durch die direkte Anwendung der Methoden im Seminar können Studierende Inspiration für ihr Vorgehen in eigenen Forschungsprojekten finden.

Erkenntnisse aus der Praxis

Im Rahmen der Integration des Forschungsprojekts *Dehmel digital* in die universitäre Lehre kommen vor allem Methoden der digitalen Textanalyse zum Einsatz. Dabei werden zunächst in einer eher passiv-rezeptiven Phase auf Grundlage der *forTEXT*-Inhalte Prozesse grundlegende Funktionen ausgewählter Tools und Methoden zur digitalen Textanalyse vermittelt (prozessorientiert, passiv). Die Lektüre von Methodeneinträgen, Tool- und Ressourcenvorstellungen (z.B. zum Thema NER oder HTR) und deren Diskussion im Seminar vermitteln theoretisch-methodologisches Hintergrundwissen und bestimmen die inhaltliche Dimension des Forschenden Lernens (nach Hearley und Jenkins 2009). In einer zweiten Phase steht die aktive Mitarbeit der Studierenden im Projektkontext im Vordergrund. Hierbei bildet die theoretische Vorarbeit den Ausgangspunkt für die teilautomatisierte Transkription sowie das Training eigener Modelle zur Erkennung von Handschriften oder Named Entities in Briefen. Auf diese Weise erwerben die Studierenden zum einen die Kompetenz, mittels digitaler Verfahren Daten zu produzieren. Zum anderen lernen sie prozessorientiert, die angewendeten Verfahren kritisch zu hinterfragen. Sie gewinnen also unter anderem einen Einblick in die Praxis der Entstehung von Quellen, die eine Grundlage der literaturwissenschaftlichen Lehre und Forschung sind, und üben den kritischen, fruchtbaren Umgang mit digitalen Quellen. Ihr konkretes Feedback besitzt auch eine Relevanz für Dritte und kann sowohl auf die Ausgestaltung des Forschungsprojekts *Dehmel digital* als auch auf die Präsentation von Inhalten auf *forTEXT.net* rückwirken: Bei *Dehmel digital* können mithilfe der Resonanz beispielsweise Veränderungen im Transkriptionsvorgehen vorgenommen werden, *forTEXT* kann z.B. inhaltliche Erweiterungen vornehmen, indem neue Methoden und Tools getestet und beschrieben werden, die es ermöglichen, die von den Studierenden beschriebenen Grenzen auszuweiten.

Beispielsweise wurde in einem Seminar 2019/20 mit Studierenden eine Korrespondenz des Dehmekorpus transkribiert. Durch Diskussionsrunden angeregt, konnten manche Metadatenkategorisierungen im Dehmelprojekt und Teile der Editionsrichtlinien optimiert werden. Das Feedback der Studierenden war positiv, vor allem unter dem Gesichtspunkt, dass sie sich anhand der für das literaturwissenschaftliche Studium eher unüblichen praktischen Arbeit im Seminar eine konkrete Vorstellung von editorischer Praxis machen konnten. Aus den Rückmeldungen konnten vonseiten der Seminarleitung zudem Ideen gewonnen werden, wie sich der

Seminarplan für das nächste Mal verständlicher und effizienter aufbauen ließe, und Vorstellungen, für welche theoretischen und methodischen Konzepte die Vermittlung noch verbessert werden könnte, wie z.B. für das Training eines HTR-Modells.

Die im Seminar erhobenen Daten und trainierten Modelle werden ebenfalls ins Projekt *Dehmel digital* zurückgespielt. Die Studierenden lernen ein konkretes Forschungsprojekt kennen und vollziehen einzelne Schritte im Forschungsprozess nach. Gleichzeitig bekommen sie einen breit angelegten Überblick über Methoden der digitalen Geisteswissenschaften und die Forschungstraditionen, an die sie anknüpfen.

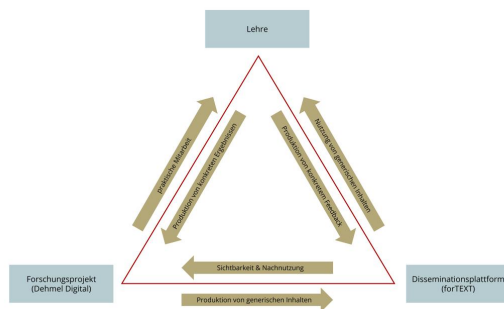


Abb. 2: Didaktisches Dreieck zwischen Dehmel digital, Lehre und forTEXT

Forschung – Lehre – Dissemination

Abbildung 2 macht anhand des beschriebenen Use Cases prototypisch das Zusammenwirken von passiv-rezipierenden, aktiv-forschenden und produzierenden Anteilen der Lehrveranstaltungen sowie von konkreten und generischen Inhalten deutlich und zeigt die Reintegration der Ergebnisse Forschenden Lernens in das Forschungsprojekt und die Disseminationsplattform. Ausgangspunkt ist das Projekt *Dehmel digital*, das die beiden Dimensionen Forschenden Lernens nach Healey und Jenkins – inhaltlicher Fokus und Studierendenbeteiligung – als auch Wulfs Kategorien von Prozess, Methode und Ergebnis voll abdeckt. Studierende arbeiten aktiv in einem Forschungssetting und können Ergebnisse zum Projekt beitragen. *forTEXT* verstärkt den methodischen Fokus und dient der Vermittlung zwischen digitalen Geisteswissenschaften und den in diesem Falle ansonsten eher analog-traditionell ausgebildeten Studierenden. Über das Projekt *Dehmel digital* können bisher auf der *forTEXT.net*-Plattform bestehende Desiderate im Bereich der digitalen Editionswissenschaften ausgemacht und in Form konkreter Inhalte ergänzt werden. Dort wird die Sichtbarkeit des Dehmel-Projekts sowie der in dessen Kontext erarbeiteten Workflows erhöht und neben den Studierenden des eigenen Seminars können auch andere Forschende und Lernende auf die Methodenkompetenzvermittlung zugreifen. Gerade die Verknüpfung von konkreter Projekt-Perspektive und generischer Sicht auf das Feld der Digital Humanities macht eine übergeordnete hermeneutische Reflexion möglich, die dem Forschenden Lernen entsprechend sowohl theorie- als auch erfahrungsbasiert ist und somit Teil eines nachhaltigen Lernprozesses werden kann. Durch diese Kombination können also Elemente des Forschenden Lernens gewinnbringend in literaturwissenschaftliche Lernszenarien integriert werden.

Bibliographie

Gius, Evelyn / Gerstorfer, Dominik / Meister, Malte / Schumacher, Mareike et al. (2021): *forTEXT. Literatur digital erforschen*. <https://fortext.net> [letzter Zugriff: 6. Juli 2021].

Healey, Mitch / Jenkins, Alan (2009): *Developing undergraduate research and inquiry*. York.

Hethy, Meike / Struve, Karen (2017): „MitLesen. Forschen des Lernen in den Literaturwissenschaften“, in: Kaufmann, Margrit E. / Satilmis, Ayla / Mieg, Harald A. (eds.): *Forschendes Lernen in den Geisteswissenschaften. Konzepte, Praktiken und Perspektiven hermeneutischer Fächer*. Wiesbaden: Springer VS 141 -166.

Hirsch, Brett D. (2012): „</Parenteses>: Digital Humanities and the Place of Pedagogy“, in: Hirsch, Brett D. (ed.): *Digital Humanities Pedagogy. Practices, Principles and Politics*. Cambridge: Open Book Publishers 3 -30.

Huber, Ludwig (2009): „Warum Forschendes Lernen nötig und möglich ist“, in: Huber, Ludwig / Hellmer, Julia / Schneider, Friederike (eds.): *Forschendes Lernen im Studium. Aktuelle Konzepte und Erfahrungen*. Bielefeld: Universitäts Verlag Weblar 9 -35.

Huber, Ludwig (2014): „Forschungsbasiertes, Forschungsorientiertes, Forschendes Lernen: Alles dasselbe? Ein Plädoyer für eine Verständigung über Begriffe und Unterscheidungen im Feld des forschungsnahen Lehrens und Lernens“, in: *Das Hochschulwesen HSW* 36 (1/2): 22 -29.

Huber, Ludwig (2017): „Forschendes Lernen in den Geisteswissenschaften. Fernes Echo seiner historischen Ursprünge“, in: Kaufmann, Margrit E. / Satilmis, Ayla / Mieg, Harald A. (eds.): *Forschendes Lernen in den Geisteswissenschaften. Konzepte, Praktiken und Perspektiven hermeneutischer Fächer*. Wiesbaden: Springer VS 21 -34.

Huber, Ludwig (2020): „Curriculare Verankerung des forschungsnahen Lernens“, in: Wulf, Carmen / Haberstroh, Susanne / Petersen, Maren (eds.): *Forschendes Lernen. Theorie, Empirie, Praxis*. Wiesbaden: Springer VS 3 -20. 10.1007/978-3-658-31489-7.

Mieg, Harald A. (2020): „Eine Systematik der Forschungsformen und ihre Eignung für Forschendes Lernen“, in: Wulf, Carmen / Haberstroh, Susanne / Petersen, Maren (eds.): *Forschendes Lernen. Theorie, Empirie, Praxis*. Wiesbaden: Springer VS 21 -34. 10.1007/978-3-658-31489-7.

Schirmer, Carola / Martin, Victoria (2020): „Die Gestaltung Forschenden Lernens mit digitalen Medien“, in: Wulf, Carmen / Haberstroh, Susanne / Petersen, Maren (eds.): *Forschendes Lernen. Theorie, Empirie, Praxis*. Wiesbaden: Springer VS 285 -297. 10.1007/978-3-658-31489-7_24.

Wulf, Carmen / Haberstroh, Susanne / Petersen, Maren (2020): *Forschendes Lernen. Theorie, Empirie, Praxis*. Wiesbaden: Springer VS. 10.1007/978-3-658-31489-7.

Gedächtnis digitaler Kulturen und digitaler Geisteswissenschaften

Plädoyer für eine Wissenschaftsgeschichte der DH

Bernhart, Toni

toni.bernhart@ilw.uni-stuttgart.de
Universität Stuttgart, Germany

Ansatz und Anspruch

Digitale Kulturen konstituieren Gedächtnis. Sie sammeln, speichern, überliefern, analysieren, verknüpfen und interpretieren Information. Doch verhalten sie sich bislang ahistorisch, indem sie sich für die Entwicklung und Geschichte ihrer Ansätze, Ansprüche und Methoden kaum interessieren und das Gedächtnis und die Reflexion ihrer selbst vernachlässigen. Die Forderung nach einer Fach- oder Wissenschaftsgeschichte der digitalen Geisteswissenschaften wurde im Verlauf der vergangenen Jahre punktuell artikuliert (Hoover 2007, Kelih 2008, Cortelazzo / Tuzzi 2008, Jannidis / Lauer 2014, Jannidis 2015, Twellmann 2015, Weitin 2015, Thaller 2017, Schöch 2017, Bernhart 2018, Fischer / Akimova / Orekhov 2019, Fischer / Gramelsberger / Hoffmann / Hofmann / Rheinberger / Rickli 2020, Bernhart im Druck), zuletzt auch deutlich in etlichen Diskussionen bei der 7. Jahrestagung des Verbands „Digital Humanities im deutschsprachigen Raum e.V.“ 2020 in Paderborn. In institutionellen Positionsbestimmungen findet das Desideratum bisher allerdings kaum programmatischen Niederschlag.

Der Beitrag möchte für die Etablierung einer Wissenschaftsgeschichte der digitalen Geisteswissenschaften sensibilisieren.¹ Eigentlich müsste das Ansinnen selbsterklärend sein, schon allein deshalb, weil digitale Geisteswissenschaften maßgeblich aus geisteswissenschaftlichen Kontexten resultieren, in denen diachrone und synchrone Perspektivierungen von Wissens- und Theoriebildung selbstverständlich sein sollten. Doch die bislang nur sehr verhaltenen Bemühungsbestrebungen in diese Richtung deuten darauf hin, dass das Bewusstsein für Sinn und Dringlichkeit einer Wissenschaftsgeschichte der digitalen Geisteswissenschaften auch in einem geisteswissenschaftlichen Kontext kaum gegeben zu sein scheint.

Kontinuum und Ähnlichkeiten

Algorithmische und digitale Kulturen und mit ihnen ihre Geschichte beginnen entgegen gängiger Lehrmeinung (vgl. etwa Ciotti / Crupi 2012, Nyhan / Flinn 2016) nicht erst mit der Erfindung des Computers. Sie beginnen dort, wo Enumeration, Indizierung, Serialisierung, Quantifizierung und Formalisierung semantischer Entitäten und kultureller Artefakte sichtbar werden. Solche Ansätze sind nicht eindeutig in eine bestimmte Epoche oder eine begrenzbare zeitliche Phase datierbar, sondern sie stellen grundlegende kulturelle Praktiken dar. Bezogen auf Literaturen, wies dar-

auf zuletzt Eva von Contzen in ihren listologischen Forschungsarbeiten hin, die Personen-, Sach- und Eigenschaftslisten von den antiken über die mittelalterlichen bis in die Literaturen der Gegenwart untersuchen (Contzen 2016, Contzen 2018). Auf einen Vorlauf, der länger ist als bislang angenommen, macht mit Blick auf digitale Gesellschaften und Kulturen auch Armin Nassehi aufmerksam, wenn er behauptet, „dass die moderne Gesellschaft bereits ohne die digitale Technik in einer bestimmten Weise *digital* ist“ (Nassehi 2019: 11; Kursivierung im Original).

Deutlich sichtbar werden Operationalisierungen und Algorithmisierungen unter dem Eindruck der Empirismen der Frühen Neuzeit. Erste Ballungen sind im Zuge der Ausdifferenzierung der Einzelwissenschaften im 19. Jahrhundert erkennbar. Erstaunlich und überraschend an diesen frühen Ansätzen sind Ähnlichkeiten mit den Zielen und Ansprüchen der digitalen Geisteswissenschaften.

Einander ähnlich sind frühe operationalisierende Ansätze und die späteren digitalen Geisteswissenschaften in ihren Ansprüchen, schneller und effizienter zu Ergebnissen zu kommen als mit der Methode des menschlichen und sinnkonstituierenden Lesens, auf der Grundlage gewonnener Daten intersubjektive und interoperable Vergleichs- und Austauschmöglichkeiten herzustellen, einen verborgenen oder nicht-bewussten Informationsgehalt literarischer oder künstlerischer Werke offenzulegen und darüber hinaus neue und ungewöhnliche Forschungsfragen zu generieren. Charakteristisch für diese Ansprüche ist darüber hinaus ein disziplinenübergreifender Ansatz, der schon sehr früh zu beobachten ist. Belege für solche Ähnlichkeiten finden sich im 19., 20. und 21. Jahrhundert. Anhand von Beispielen, die sich in ihren Praktiken und Zielsetzungen im Rückblick als (proto-)typisch erweisen, möchte dieser Beitrag die genannten Ansprüche erläutern.

Effizienz und Offenlegung verborgener Ästhetiken

Als ein Beispiel dafür, dass von operationalisierter, quantitativer oder algorithmischer Arbeitsweise raschere Ergebnisse zu erwarten seien, kann Thomas C. Mendenhall (1841–1924) gelten. Zur Datenerhebung für seine Stilanalysen verwendete er mechanische Zählmaschinen, mit denen es ihm – wie er berechnete – möglich war, in einem Viertel der Zeit, die bei traditionellem Vorgehen erforderlich gewesen wäre, zu Ergebnissen zu kommen; bedient wurden die Zählmaschinen bezeichnenderweise von Frauen (Mendenhall 1901: 102). Sein methodisches Vorgehen beschreibt Mendenhall folgendermaßen:

Nearly twenty years ago I devised a method for exhibiting graphically such peculiarities of style in composition as seemed to be almost purely mechanical and of which an author would usually be absolutely unconscious. The chief merit of the method consisted in the fact that its application required no exercise of judgment, accurate enumeration being all that was necessary, and by displaying one or more phases of the mere mechanism of composition characteristics might be revealed which the author could make no attempt to conceal, being himself unaware of their existence. (Mendenhall 1901: 97)

Kennzeichnend für die entwickelte Methode ist der Weg der Wissensgenerierung: Stilistische Besonderheiten würden „rein mechanisch“ („purely mechanical“) sichtbar und ein Autor sei sich der in seinem Text wirksamen Muster „absolut nicht bewusst“ („absolutely unconscious“). Die Hauptleistung der Methode bestehe darin, dass eine forschende Person über keiner-

lei fachwissenschaftlich geschultes Urteilsvermögen verfügen müsse, allein die akkurate Darstellung der Messwerte genüge, um die Kompositionsprinzipien offenzulegen, die ein Autor nicht verbergen könne, ja, deren Existenz ihm weder bewusst noch bekannt sei. Die hier sehr früh formulierte Vorstellung einer mechanischen, unbewussten oder verborgenen Poetik ist kennzeichnend für zahlreiche quantitative und algorithmische Analyseverfahren und findet sich in abgewandelter Form in Theorien der Digital Humanities wieder (Burrows 1987: 2, Moretti 2005: 54, Jockers 2013: 106–108, Jannidis 2014: 169).

Intersubjektivität und neue Fragestellungen

Beispielhaft ist etwa der wenig beachtete Psychologe Karl Groos (1861–1946), der umfassende statistische Analysen zu Farbworthäufigkeiten in den Werken von William Shakespeare, Johann Wolfgang Goethe, Friedrich Schiller, Richard Wagner u.a. unternommen hat. In einer kurzen autobiografischen Schrift, in der er seine Arbeitsweise reflektiert, hält er fest:

Erstens können wir auf Grund von zahlenmäßigen Feststellungen in einer objektiveren Weise vergleichen, sei es nun, daß es sich um die Eigenart verschiedener Individuen, sei es, daß es sich um verschiedene Perioden in der Entwicklung derselben Persönlichkeit handelt. Und zweitens treibt das Fixieren von Zahlen sozusagen automatisch neue Fragestellungen aus sich heraus, auf die eine andere Methode gar nicht verfallen würde. (Groos 1921: 109–110)

Groos weist hier auf die Möglichkeiten hin, dass „zahlenmäßige[] Feststellungen“ zu „objektiveren“ Arbeitsweisen führen können und dass solche Verfahren „sozusagen automatisch“ neue Forschungsfragen generierten. „[E]ine andere Methode“ – er meint damit wohl erprobte und etablierte Hermeneutiken – würde zu solchen neuen Fragen gar nicht erst führen. Eine vergleichbare Argumentation spielt in den Digital Humanities eine wichtige Rolle, wenn darauf verwiesen wird, dass Algorithmen und digitale Modellierungen in nicht unwesentlichem Maße dazu beitragen, weiterführende und erweiternde Fragen aufzuwerfen. Die Methoden und Werkzeuge der Digital Humanities lieferten demnach nicht nur Ergebnisse im Sinne von Daten und Erkenntnissen, sondern dienten wesentlich auch der Heuristik auf dem Weg zur Interpretation. Das Verfahren, das Groos beschreibt, zeigt überraschende Ähnlichkeit mit dem später von Franco Moretti modellierten Konzept des *Distant reading*.

Disziplinenübergreifende Arbeitsweise

Es fällt auf, dass operationalisierende, quantifizierende und algorithmisierende Verfahren schon sehr früh oft disziplinenübergreifend operierten. Entsprechende Ansätze zur Bearbeitung geisteswissenschaftlich virulenter Fachfragen wurden häufig nicht innerhalb der dafür zuständigen Fächer entwickelt, sondern in anderen Disziplinen wie der Medizin, der Physik oder der Mathematik und von diesen an die zuständigen Disziplinen herangetragen oder ihnen angeboten. Nicht selten hatte dies dort Abwehrreaktionen zur Folge.

Ein frühes Beispiel dafür ist Sir Thomas Young (1773–1829), Arzt und Professor für Naturphilosophie an der Royal Institution

in London. In seiner Schrift mit dem Titel *Remarks on the probabilities of error in physical observations, and on the density of the earth, considered, especially with regard to the reduction of experiments on the pendulum* (Young 1819) geht es um Fragen der physikalischen Dichte der Erde, die Young mit den Methoden der Wahrscheinlichkeitsrechnung behandelt. Doch auch Fragen der sogenannten Sprachenverwandtschaft, zur realgeschichtlichen Aussagekraft literarischer Quellen und zur Entschlüsselung der Hieroglyphenschrift greift er auf, für die er Lösungen mithilfe der Wahrscheinlichkeitsrechnung vorschlägt. Die Frage der Ähnlichkeit von Sprachen und die noch nicht entschlüsselten Hieroglyphen waren drängende wissenschaftliche Fragen der Zeit.

Ausblick

Indem Praktiken neben Objekten das Fundament für kulturelles Gedächtnis bilden, bilden digitale, auch proto-digitale Praktiken umso deutlicher das Fundament für das Gedächtnis digitaler Kulturen. Die Geisteswissenschaften sind dazu berufen und befähigt, solche Gedächtnisse zu kartografieren und systematisch und historisch zu perspektivieren.

Digitale Praktiken in einem weiten Sinn gibt es nicht erst, seitdem von Digital Humanities gesprochen wird, auch nicht erst seit der Erfindung des Computers. Sie emergieren vielmehr dann, wenn zählende, messende, rechnende, operationalisierende und algorithmisierende Ansätze und Verfahren bei der Beschreibung, Speicherung und Analyse insbesondere kultureller Artefakte Anwendung finden.

Konkret stellt dies die Digital Humanities – nach ihrer Emanzipierung innerhalb der Geisteswissenschaften, um die sie lange ringen mussten – vor einen erweiterten kulturellen Horizont, der zu ihrer historisch informierten Verortung, zum Verständnis ihrer Leistungsfähigkeit und ihrer Grenzen und zur Systematisierung (proto-)typischer Denk- und Argumentationsmuster beitragen kann. Auch die Digital Humanities haben eine Geschichte und ein Anrecht darauf, dass sie erzählt – und geschrieben – wird.

Fußnoten

1. Der Beitrag fußt auf den Forschungsergebnissen des DFG-geförderten Forschungsprojekts „Quantitative Literaturwissenschaft“ (Projektnummer 259167649).

Bibliographie

Bernhart, Toni (2018): „Quantitative Literaturwissenschaft: Ein Fach mit langer Tradition?“, in: Bernhart, Toni / Willand, Marcus / Richter, Sandra / Albrecht, Andrea (Hg.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin, Boston: Walter de Gruyter 207–219.

Bernhart, Toni (im Druck): „Algorithmische Wissenskulturen in den Geisteswissenschaften und ihr Vorlauf im 19. Jahrhundert“, in: Hashagen, Ulf / Seising, Rudolf (Hg.): *Algorithmische Wissenskulturen? Der Einfluss des Computers auf die Wissenschaftsentwicklung*. Cham: Springer.

Burrows, John F. (1987): *Computation into Criticism. A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Ciotti, Fabio / Crupi, Gianfranco (Hg.) (2012): *Dall'informatica umanistica alle culture digitali. Atti del convegno di studi (Roma, 27-28 ottobre 2011) in memoria di Giuseppe Gigliozzi*. Roma: Sapienza Università Editrice.

Contzen, Eva von (2016): „The Limits of Narration: Lists and Literary History“, in: *Style* 50/3: 241–260.

Contzen, Eva von (2018): „Experience, Affect, and Literary Lists“, in: *Partial Answers* 16/2: 315–327.

Cortelazzo, Manlio / Tuzzi, Arjuna (2008): *Metodi statistici applicati all'italiano*. Bologna: Zanichelli.

Fischer, Frank / Akimova, Marina / Orekhov, Boris (Hg.) (2019): *Moscow Formalism and Literary History* (Sonderband der Zeitschrift *Journal of Literary Theory*). Berlin: Walter de Gruyter.

Fischer, Philipp / Gramelsberger, Gabriele / Hoffmann, Christoph / Hofmann, Hans / Rheinberger, Hans-Jörg / Rickli, Hannes (2020): *Datennaturen. Ein Gespräch zwischen Biologie, Kunst, Wissenschaftstheorie und -geschichte*. Zürich: Diaphanes.

Groos, Karl (1921): [Ohne Titel], in: Schmidt, Raymund (Hg.): *Die Philosophie der Gegenwart in Selbstdarstellungen*, Band 2. Leipzig: Meiner 101–115.

Hoover, David L. (2007): „Quantitative Analysis and Literary Studies“, in: Siemens, Ray / Schreibman, Susan (Hg.): *A Companion to Digital Literary Studies*. Malden, Mass. u.a.: Blackwell 517–533.

Jannidis, Fotis (2014): „Der Autor ganz nah. Autorstil in Stilistik und Stilometrie“, in: Schaffrick, Matthias / Willand, Marcus (Hg.): *Theorien und Praktiken der Autorschaft*. Berlin, Boston: Walter de Gruyter 169–195.

Jannidis, Fotis (2015): „Perspektiven empirisch-quantitativer Methoden in der Literaturwissenschaft. Ein Essay“, in: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte (DVJs)* 89/4: 657–661.

Jannidis, Fotis / Lauer, Gerhard (2014): „Burrows Delta and its Use in German Literary History“, in: Erlin, Matt / Tatlock, Lynne (Hg.): *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*. Suffolk: Boydell & Brewer 29–54.

Jockers, Matthew L. (2013): *Macroanalysis. Digital Methods and Literary History*. Urbana, Chicago, Springfield: University of Illinois Press.

Kelih, Emmerich (2008): *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft*. Hamburg: Kovač.

Mendenhall, Thomas C. (1901): „A Mechanical Solution of a Literary Problem“, in: *Popular Science Monthly* 60: 97–105.

Moretti, Franco (2005): *Graphs, Maps, Trees. Abstract Models for a Literary History*. London, New York: Verso.

Nassehi, Armin (2019): *Muster. Theorie der digitalen Gesellschaft*. München: Beck.

Nyhan, Julianne / Flinn, Andrew (2016): *Computation and the Humanities. Towards an Oral History of Digital Humanities*. Heidelberg: Springer.

Schöch, Christof (2017): „Quantitative Analyse“, in: Jannidis, Fotis / Kohle, Hubertus / Rehbein Malte (Hg.): *Digital Humanities. Eine Einführung. Mit Abbildungen und Grafiken*. Stuttgart: Metzler 279–298.

Thaller, Manfred (2017): „Geschichte der Digital Humanities; Digital Humanities als Wissenschaft“, in: Jannidis, Fotis / Kohle, Hubertus / Rehbein Malte (Hg.): *Digital Humanities. Eine Einführung. Mit Abbildungen und Grafiken*. Stuttgart: Metzler 3–18.

Twellmann, Marcus (2015): „„Gedankenstatistik“. Vorschlag zur Archäologie der Digital Humanities“, in: *Merkur* 69: 19–30.

Weitin, Thomas (2015): „Digitale Literaturwissenschaft“, in: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte (DVJs)* 89/4: 651–656.

Genitivmetaphern in der Lyrik des Realismus und der frühen Moderne

Kröncke, Merten

merten.kroencke@uni-goettingen.de
Universität Göttingen, Germany

Konle, Leonard

leonard.konle@uni-wuerzburg.de
Universität Würzburg, Germany

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Germany

Winko, Simone

simone.winko@phil.uni-goettingen.de
Universität Göttingen, Germany

Einleitung

Ein wichtiger Aspekt der sprachlichen Gestaltung literarischer Texte besteht im Einsatz von Metaphern, Metonymien und Tropen im Allgemeinen. Einzelnen Werken, aber auch ganzen Gattungen oder Epochen wird zugeschrieben, dass ihre Spezifik nicht zuletzt in einer jeweils charakteristischen Verwendungsweise uneigentlicher Rede gründe. Unter anderem betrifft das die Geschichte der Lyrik, das heißt die Geschichte einer Gattung, die laut Benjamin Specht „in Bezug auf die Verwendung von Metaphern die weitesten Lizenzen besitzt“ (Specht, 2017: 90).

Das Ziel dieses Beitrags besteht darin, den Gebrauch von Metaphern in der deutschsprachigen Lyrik des Realismus und der frühen Moderne¹ mithilfe automatisierter, quantitativer Methoden zu identifizieren und zu analysieren. Damit sollen mehrere literarhistorische Forschungsthesen auf einer breiten Datengrundlage geprüft und gegebenenfalls differenziert werden.

Die literaturwissenschaftliche Forschung macht die Unterscheidung von realistischer und moderner Lyrik unter anderem am Aufkommen neuer, innovativer Formen uneigentlichen Sprechens fest. In der modernen Lyrik treten Metaphern auf, die als „Radikalisierung, Komplizierung und Steigerung“ lyrischer Bildlichkeit (Hiebel 2005: 28), als Beitrag zur sprachlichen „Verfremdung“ (Lamping, 2010: 148; vgl. auch Lamping, 2008: 25f), als „assoziativ-hermetische“ Muster (Specht, 2014: 5) oder auch als „Blume[n] ohne Stiel auf der Oberfläche des Gedichts“ (Neumann, 1970: 195f) zu charakterisieren seien. Die Forschung dürfte sich einig sein, dass die moderne Metaphorik gegenüber der vorherigen, traditionellen Bildlichkeit zu größerer Individualität und Heterogenität tendiert (vgl. zur Homogenität der realistischen (Massen-)Lyrik und ihrer Sprachbilder z. B. Stockinger, 2010: 88). Doch durch welche Textmerkmale sich die neuen, modernen Metaphern im Einzelnen auszeichnen, wird unterschiedlich und zum Teil sogar gegensätzlich konzeptualisiert.

1. *Große vs. kleine Bildspanne*: Hugo Friedrich ist der Auffassung, dass die Bildspanne, also etwa im Fall der Metapher ‚die Asche der Schande‘ der semantische Abstand zwischen den Begriffen ‚Asche‘ und ‚Schande‘, bei modernen Metaphern besonders groß sei (Friedrich, 1992: 17f, 207). Harald Weinrich geht vom Gegenteil aus: „Gerade die Nahmetaphern sind befremdend und verfremdend und erscheinen uns kühn. Fernmetaphern sind ungefährlich.“ (Weinrich, 1963: 335)
2. *Abstraktion vs. Konkretisierung*: Eine weitere These Friedrichs lautet, dass in der Lyrik der Moderne besonders häufig Konkretes mit Abstraktem verbunden werde, etwa in Metaphern wie ‚der Schnee des Vergessens‘ (Friedrich, 1992: 205; Andreotti, 2014: 317). Gerhard Neumann konstatiert am Beispiel der Metaphorik Mallarmés ein „Zunehmen des konkreten Wortmaterials“ und ein „Untergehen der abstrakten Substantive“ (Neumann, 1970: 199), während Martin Anderle annimmt, dass gerade der Realismus in seiner Bildlichkeit zum „Gegenständliche[n]“ tendiere, die Moderne hingegen zum „Metaphysischen“ neige (Anderle, 1979: 77-79, Zitat 79).

Unser Beitrag untersucht nur Genitivmetaphern (‚Das Lächeln der Natur‘ usw.); andere Formen, zum Beispiel Adjektivmetaphern (‚Die lächelnde Natur‘ usw.), bleiben unberücksichtigt. Zumindest Hugo Friedrich ist allerdings der Auffassung, dass es sich bei Genitivmetaphern ohnehin um den häufigsten Typ von Metaphern in der (modernen) Lyrik handelt (Friedrich, 1992: 205).

Im Normalfall der Genitivmetapher bezieht sich das *Kopfnomen* auf den Ursprungsbereich der Metapher, in dem der Ausdruck im nicht-übertragenen Sinn verwendet wird, während das *Genitivnomen* sich auf den Zielbereich der übertragenen Rede bezieht. Das erste Nomen wird metaphorisch, das im Genitiv wörtlich verstanden (Skirl und Schwarz-Friesel, 2013: 22), z.B. „Lächeln der Natur“, „Feuermeer der Melodie“. Es können auch beide Nomen metaphorisch verwendet werden und gemeinsam auf einen Ursprungsbereich verweisen. Dies ist z.B. der Fall in „Und wenn es Abend ist, / Empfangen sie [die Menschen] den Tau der Gnadensonne“ (Franz Evers: Der Künstler), wo die Genitivmetapher „Tau der Gnadensonne“ sich auf das Alter der Menschen bezieht. Die Verse sind zugleich ein Beispiel dafür, dass in Gedichten Genitivmetaphern auch Teil einer umfassenderen Passage mit uneigentlicher Rede sein können, in diesem Fall einer Allegorie.

Die hier untersuchten Phänomene (Metaphern in Genitivkonstruktionen) sind mithin nicht exakt identisch mit dem Gegenstand der literaturwissenschaftlichen Forschungsthesen (Metaphern im Allgemeinen), auch wenn man davon ausgehen darf, dass Aussagen über den einen Bereich ebenfalls relevant für den anderen Bereich sind. Eine weitere Relativierung betrifft das Untersuchungskorpus: Während sich die Forschungsaussagen in der Regel auf kanonisch-moderne sowie des Öfteren auf deutlich nach 1900 erschienene Gedichte beziehen, enthält das hier zu analysierende Korpus lediglich Texte der Jahrhundertwende um 1900 und damit der *frühen* Moderne, zudem umfasst es auch Texte von nicht-kanonischen Autoren und Autorinnen dieser Zeit.

Ressourcen

Die analysierten Gedichte stammen aus den 7 Anthologien des Realismus und den 13 Anthologien der Lyrik um 1900 mit insgesamt 6249 Texten, die wir im Rahmen des Projekts „The beginnings of modern poetry - Modeling literary history with text similarities“ untersuchen. Bei den Anthologien um 1900 handelt es

sich um Sammlungen, deren Herausgeber Gedichte aufgrund ihrer Modernität ausgewählt haben (siehe Tabelle 1).

Tab. 1: Korpus Statistik.

	Anzahl Gedichte	Anzahl Wörter
Realismus	3367	400k
Moderne	2882	320k
Insgesamt	6249	720k

Aus diesem Korpus sind unter Verwendung von spaCy (Montani et al., 2021) 4300 Genitivkonstruktionen² extrahiert worden.

Drei Annotatoren haben Genitivkonstruktionen als metaphorisch oder nicht-metaphorisch annotiert. Annotiert wurden die beiden oben beschriebenen Formen von Genitivmetaphern, aber nur wenn sie in dem Muster *Kopfnomen---Artikel---Nomen_im_Genitiv* vorkommen (wir haben das Muster *Nomen_im_Genitiv---Kopfnomen* ignoriert). Während des Annotationsprozesses sind verschiedene Problemfälle aufgetreten. Zum Beispiel konnte die Abgrenzung der Metaphern von einer anderen Form uneigentlichen Sprechens, der Metonymie, herausfordernd sein. Zudem musste entschieden werden, ob eine Metapher im Untersuchungszeitraum über wiederholte Verwendung bereits so etabliert war, dass sie den Charakter des Metaphorischen im Sinn einer Übertragung verloren hatte. Potentiell lexikalisierte Genitivmetaphern (z.B. „Fuß des Berges“) haben wir manuell in Wörterbüchern der Zeit (Sanders, 1865; Heyne, 1895) nachgeschlagen und als metaphorisch markiert, wenn sie dort als Übertragung gekennzeichnet wurden.

Tab. 2: Zusammensetzung der Trainingsdaten.

	Genitivkonstruktionen	davon Metaphern	davon keine Metaphern
Annotationen Realismus	317	176	141
Annotationen Moderne	308	132	176
Vorstudie	285	173	112
Insgesamt	910	442	468

Der annotierte Datensatz umfasst 625 Genitivkonstruktionen mit einem Agreement von 0.53³ und 285 weitere Beispiele aus einer Vorstudie (siehe Tabelle 2). Da für diese keine Zweitannotation zur Validierung vorliegt, werden sie ausschließlich zum Training, nicht zur Evaluation verwendet.

Für die automatische Metaphererkennung verwenden wir neben den Annotationen ein deutsches Bert Model⁴ (Chan et al., 2020), ein FastText (Bojanowski et al., 2016) Embedding⁵, trainiert auf dem deutschen Oscar Subkorpus (Open Super-large Crawled Aggregated coRpus, Ortiz Suárez et al., 2020), Super-sense Wortlisten aus GermaNet⁶ und die in Köper und Schulte im Walde (2016) beschriebene Wortliste zu Abstraktheit, Vorstellbarkeit, Arousal und Valenz.

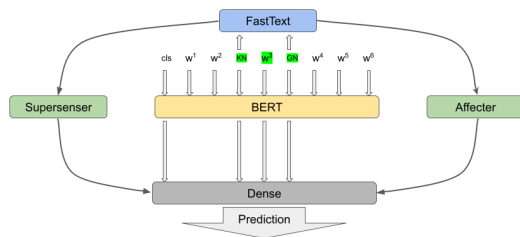


Abb. 1: Schema des Systems zur Erkennung von Metaphern in Genitivkonstruktionen. KN: Kopfnomen; GN: Genitivnomen.

Die Erkennung der Metaphern geschieht in zwei Schritten. Im ersten Schritt werden regelbasiert Genitivkonstruktionen erkannt (siehe Ressourcen); im zweiten Schritt werden diese als 'nicht-metaphorisch' oder 'metaphorisch' klassifiziert. Das System zur Klassifikation von Genitivmetaphern setzt sich aus drei Komponenten zusammen: Supersenser, Affecter und Bert (siehe Abb. 1). Damit folgt der Aufbau dem in Tsvetkov et al. (2014) vorgestellten Ansatz Metaphern unter Berücksichtigung von Abstraktheit, Vorstellbarkeit und Wortklasse ihrer Komponenten zu klassifizieren.

Das Training des Supersenser Moduls wird auf den FastText Vektoren der Wörter aus den 20 größten Supersense-Klassen für Substantive aus GermaNet durchgeführt. Diese werden durch überwachter Dimensionsreduktion (Szubert et al., 2019) in einen kleineren Raum mit 10 Dimensionen projiziert. Im Gegensatz zur direkten Verwendung von GermaNet als Eingabe in das System können so out-of-vocabulary Probleme vermieden werden. Außerdem wird durch die Projektion eine reichhaltigere Repräsentation erzeugt in der Supersense-Klassen in Beziehung gesetzt werden können. Eine Evaluation mittels kNN Klassifikation von Substantiven im projizierten Raum in ihre Supersense Klasse ergibt einen F-Score von 0.65.

Das Affecter Modul erhält ebenfalls FastText Vektoren, sowie die zugehörigen Werte aus der Wortliste von Köper und Schulte im Walde (2016). Eine 4-fach Regression durch ein MLP erreicht R^2 : 0.86.

Für die Klassifikation von Metaphern werden die Ausgaben aus Supersenser und Affecter an ein nach (Gao et al., 2019) modifiziertes BERT Modell übergeben. Dieses reicht nicht nur das CLS-Token, sondern auch die Embeddings der Genitivkonstruktion weiter. Bei der Unterscheidung zwischen Metaphern und sonstigen Genitivkonstruktionen erreicht das System einen F1 Score von 0.75 (Details siehe Tabelle 3). Die Klassifikation von Metaphern aus dem Realismus wird zwar leicht besser evaluiert als die aus den modernen Anthologien, der Effekt ist für die weitere Analyse aber vernachlässigbar.

		Precision	Recall	f1-score	Support
Realismus	Metapher	0.78	0.81	0.79	176
	Keine Metapher	0.75	0.72	0.73	141
Moderne	Metapher	0.65	0.82	0.73	132
	Keine Metapher	0.83	0.68	0.75	176
Insgesamt	Metapher	0.72	0.81	0.76	308
	Keine Metapher	0.79	0.69	0.74	317
Gesamt F1 Macro				0.75	

Tab. 3: Evaluation der Metaphern Klassifikation

Ergebnisse

Es gibt keinen Unterschied zwischen den Epochen in Hinsicht auf die Menge der Genitivkonstruktionen und den Anteil der Metaphern daran. Die These von der größeren Heterogenität und Individualität der modernen Metaphern haben wir mit zwei Operationalisierungen untersucht: Nimmt der Anteil an seltenen Wörtern zu und steigt der Type-Token-Ratio bei Metaphern der Moderne? Den Anteil der seltenen Wörter haben wir mit einem einfachen Verfahren überprüft, nämlich ob die Token im Wordembedding enthalten sind; wenn nicht, wurden sie als 'seltenen Wörter' identifiziert. Solche Wörter sind in der Lyrik häufig, zumeist handelt es sich um Komposita oder um Schreibvarianten aufgrund der Anpassung ans Metrum. Die Verwendung solcher seltenen Wörter ist auch schon im Realismus häufig: 26%, findet sich aber noch einmal häufiger in der Moderne 33%.

Wir haben den Type-Token-Ratio beider Bestandteile der Genitivmetaphern jeder Epoche untersucht und können auch hier einen signifikanten Unterschied feststellen: Bei den Metaphern der Moderne ist die Variabilität des Wortmaterials deutlich größer (siehe Abb. 2 und 3).

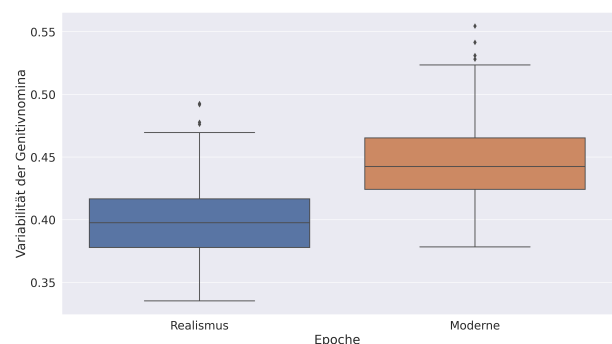


Abb. 2: Variabilität (str) der Genitivnomina nach Epoche.

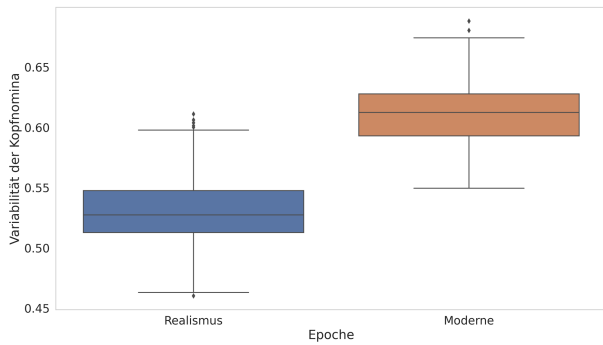


Abb. 3: Variabilität (str) der Kopfnomina nach Epoche.

Insgesamt können wir also den Eindruck bestätigen, dass die Metaphern der Moderne -- und das noch vor dem Expressionismus -- heterogener und individueller sind.

Die These zur Vergrößerung bzw. Verkleinerung der Bildspanne in den Metaphern haben wir überprüft, indem wir den Kosinusabstand der Substantive im FastText-Wordembedding gemessen haben. Wie Abb. 4 zeigt, hat der Abstand weder zu- noch abgenommen.

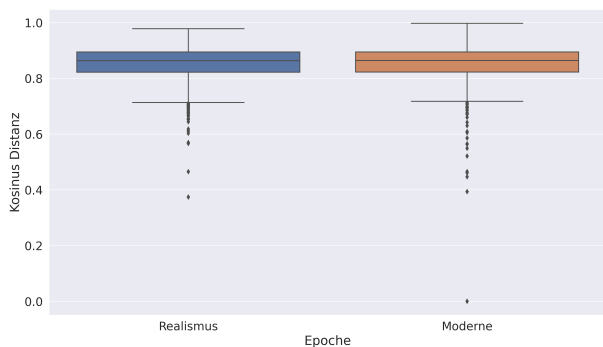


Abb. 4: Kosinusdistanz der Wordvektoren (fastText) der Metapherkomponenten nach Epoche.

Die These, dass in der Moderne der Abstand zwischen den Nomina einer Genitivmetapher in Hinsicht auf die Abstraktheit zugenommen hat, haben wir mit den Abstraktheits-Werten überprüft, die das oben dargestellte Affecter-Modul vorhergesagt hat (siehe Abb. 5), ebenso wie die These, dass die Moderne insgesamt zu abstrakteren (oder gerade zu weniger abstrakten) Metaphern neigt.

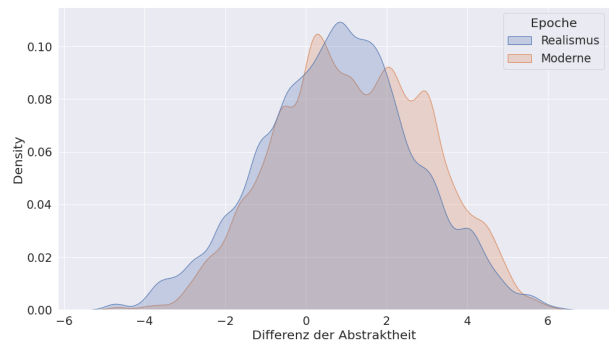


Abb. 5: Differenz der Abstraktheit in Metaphern. Formel: Abstraktheit der Kopfnomina - Abstraktheit des Genitivnomina. Hoher Wert: Demut der Finger, Reiz der Schale; Niedriger Wert: Krone des Strebens, Baum der Weisheit. Datengrundlage: 100 Zufallsstichproben mit je 200 Metaphern.

Die Forschungsthese lassen sich durch unsere Daten nicht bestätigen: Weder verändert sich das Level der durchschnittlichen Abstraktheit oder Konkretheit noch die *Differenz* der Abstraktheit-Werte wesentlich. Allerdings kann man einen andersartigen (schwachen, aber signifikanten⁷) Trend beobachten: Abb. 5 zeigt, dass in der Lyrik des Realismus ein leichtes Übergewicht von Metaphern des Typs konkret-abstrakt ("Wasser der Vergessenheit" usw.), in der Moderne hingegen ein leichtes Übergewicht des Typs abstrakt-konkret ("Demut der Finger" usw.) zu beobachten ist. Wie sich diese Metaphern auf Autorinnen und Autoren verteilen, wird in Zukunft genauer zu untersuchen sein. Die folgende Liste zeigt die Metaphern mit der größten Abstraktheit-Differenz: konkret-abstrakt: "Zweig der Pflicht", "Felsen der Gerechtigkeit", "Knospe der Hoffnung", "Blüten des Geistes", "Becher der Lust". Abstrakt-konkret: "Forderung des Meeres", "Lustgetön der Wälder", "Geist der Taube", "Zeit der Rose", "Glorie der Flammen".

Ein Blick auf die beliebtesten Wörter der Metaphern, getrennt nach der Verwendung in der Kopfposition oder im Genitiv, zeigt einige interessante Verschiebungen. Bei den Kopfnomina hat 'Meer' eine steile Karriere von Rang 10 auf Rang 1 gemacht, während 'Hauch', 'Geist' und 'Traum' deutlich weniger verwendet werden (siehe Abb. 6). Insgesamt sind die Wiederholungen bei den Kopfnomina weniger häufig und sie nehmen in der Moderne noch ab.

1 Hauch Geist _{22x}	1 Meer _{16x}
2 Lied _{17x}	2 Tag _{13x}
3 Traum Strom _{16x}	3 Lied _{12x}
4 Tag Licht _{15x}	4 Nacht Strom Land _{10x}
5 Bild Sturm Wort _{13x}	5 Duft Blut Reich _{8x}
6 Herz Nacht _{12x}	6 Zeit Hauch Kranz Baum _{8x}
7 Kelch Strahl _{10x}	7 Herz Auge Gold _{7x}
8 Schooß Zauber Reich _{8x}	8 Licht Stunde Schatten Strahl Traum Ge
9 Kind Glanz Tage _{8x}	9 Wolken Tal Sohn Blick Taumel Sturms
10 Nebel Schatten Meer Stern _{7x}	10 Mond Pforten Blume Sonne Quell Geister Zaub
11 Kuß Kampf Blume _{6x}	11 Schale Wogen Glanz Augen Himmel Fülle Dunkel TI
12 Schauer Baum Zeit Auge Blick Kreis Blumen Lauf	12 Gespenst Seele Harfe Augenblick Streit Werk Geheimnis F
13 Boten Glut Wunder Brand Zug Worte Blüte Fülle Reiz	
14 Gesang Kelche Thau Groß König Schmuck Blitz Glocken Land S	
15 Schwingen Paradies Grunde Wind Luft First Grund Saat Pforten Blüten	

Abb. 6: Kopfnomina Realismus und Moderne.

Die Genitivnomina (Abb. 7) dagegen zeichnen sich durch zahlreiche Wiederholungen aus, allerdings nimmt auch hier die Frequenz in der Moderne ab. Auffällig ist, wie stabil die Wortlisten in den oberen Rängen sind, mit der Ausnahme von "Glücks" und "Todes", die in der Moderne deutlich häufiger werden.

1 Liebe _{72x}	1 Lebens _{59x}
2 Nacht _{46x}	2 Nacht _{44x}
3 Lebens _{36x}	3 Liebe _{32x}
4 Natur _{31x}	4 Zeit _{24x}
5 Zeit _{30x}	5 Seele _{23x}
6 Himmels _{22x}	6 Glücks Natur _{15x}
7 Seele _{19x}	7 Todes Sonne _{14x}
8 Jugend Welt _{12x}	8 Himmels Erde _{12x}
9 Licht _{5x}	9 Ewigkeit Lust _{12x}
10 Wellen Zeiten _{12x}	10 Sterne _{10x}
11 Sonne Schmerzen _{12x}	11 Stunde Schönheit Frühlings Welt _{10x}
12 Tages Sterne Rosen _{10x}	12 Herzens Sehnsucht Seelen Menschheit _{10x}
13 Tage Lust Schönheit Schönen Erde _{10x}	13 Zukunft Tages Ferne _{10x}
14 Ewigkeit Wälder Tage _{10x}	14 Lichte Berge Einsamkeit Alles _{10x}
15 Glücks Nächte Einsamkeit Frieden _{10x}	15 Mondes Gedanken Wahrheit Freiheit Verweilung _{10x}
16 Zukunft Stunde Vögel Stern Flügeln Freiheit als Sehnsucht _{10x}	16 Güte Schmerzen Leidenschaft Wärme Kraft Hande Wandel Seiner Dasein Vögel Vorden Zeiten Tiden _{10x}

Abb. 7: Genitivnomina Realismus und Moderne.

Fazit

Insgesamt konnten wir einige der von der Literaturwissenschaft aufgestellten Vermutungen bestätigen, z. B. dass die Komplexität der Metaphern in der Moderne ansteigt. Da viele der Forschungsthemen für die Moderne insgesamt formuliert wurden, müssen wir allerdings dort, wo wir diese nicht bestätigen konnten, es zumindest offen halten, ob sie nicht auf spätere Phasen zutreffen. Allerdings deuten wir unsere Befunde doch so, dass die bisherige Forschung den Bruch zwischen den Epochen deutlich überbetont hat. Unsere Ergebnisse weisen vielmehr darauf hin, dass es sich um eine semantische Evolution handelt. Die hier dargestellten Ergebnisse sollen außerdem in Zukunft noch verbessert werden, indem die regelbasierte Erkennung der Genitivkonstruktionen auf eine bessere Grundlage gestellt wird: Da wir zur Zeit kein Korpus an Gedichten haben, in dem alle Genitivmetaphern annotiert sind, können wir den Recall nicht einschätzen. Außerdem soll ein umfassendes Korpus aus Texten des 19. Jahrhundert das Training eines domänenangepassten Fasttext-Modells ermöglichen, das historisch angemessenere semantische Distanzen zurückgibt. Nicht zuletzt soll auch die Erkennungsgenauigkeit des Systems zur Erkennung der Metaphern durch eine Vergrößerung des Trainingskorpus, durch eine aufwendigere Hyperparameter-Optimierung und durch Arbeit an der Architektur verbessert werden. Ein etwas anspruchsvolleres Ziel besteht außerdem in dem Versuch, die Typisierung der Metaphern auf eine andere Grundlage zu stellen, etwa durch Anschluss an die Kategorien linguistischer Metapherntheorien oder an große Ontologien. Nicht zuletzt werden wir untersuchen, wo die kanonisierten Lyriker dieser Zeit, George, Hofmannsthal, Holz und Rilke, stehen.

Fußnoten

- Wir werden im Folgenden abkürzend immer von ‘Moderne’ sprechen, meinen aber die Anfänge der Moderne um 1900, denn aus dieser Zeit stammen die Anthologien, deren Gedichte wir untersuchen.
- Unter diesen 4300 Konstruktionen finden sich 3784 echte Genitivkonstruktionen (true positives) und 516 sonstige Konstruktionen (false positives), was einer Precision von 0.88 entspricht. Über den Recall kann keine Auskunft gegeben werden, da keine Genitivkonstruktionen direkt im Text annotiert worden sind.
- Cohens Kappa
- <https://huggingface.co/deepset/gbert-large>
- Code und Material: https://github.com/LeKonArD/DH-d2022_Metaphern

6. Version 16

7. t-Test für unabhängige Stichproben: $p < .001$

Bibliographie

- Anderle, Martin** (1979): *Deutsche Lyrik des 19. Jahrhunderts: Ihre Bildlichkeit: Metapher – Symbol – Evokation*. volume Bd 287. Bouvier, Bonn.
- Andreotti, Mario** (2014): *Die Struktur der modernen Literatur: Neue Wege in die Textanalyse. Einführung Epik und Lyrik*. volume 1127. Wien/Köln/Weimar, 5. Auflage.
- Bojanowski, Piotr / Grave, Edouard / Joulin, Armand / Mikolov, Tomas** (2016): Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*, July. *arXiv: 1607.04606*.
- Chan, Braden / Schweter, Stefan / Möller, Timo** (2020): German’s Next Language Model. *arXiv:2010.10906 [cs]*, December. *arXiv: 2010.10906*.
- Friedrich, Hugo** (1992): *Die Struktur der modernen Lyrik. Von der Mitte des neunzehnten bis zur Mitte des zwanzigsten Jahrhunderts*. Rohwolt, Hamburg.
- Gao, Zhengjie / Feng, Ao / Song, Xinyu / Wu, Xi** (2019): Target-Dependent Sentiment Classification With BERT. *IEEE Access*, 7:154290–154299.
- Heyne, Moriz** (1895): *Deutsches Wörterbuch*. Hirzel, Leipzig.
- Köper, Maximilian / Schulte im Walde, Sabine** (2016): Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 {G}erman Lemmas. In S. 2595–2598, Portorož, Slovenia. ERLA.
- Lamping, Dieter** (2008): *Moderne Lyrik. Eine Einführung*. Göttingen, 2. Aufl. edition.
- Lamping, Dieter** (2010): *Das lyrische Gedicht. Definitionen zu Theorie und Geschichte der Gattung*. Vandenhoeck & Ruprecht, Göttingen, 3rd edition.
- Montani, Ines / Honnibal, Matthew** (2021): *explosion/spaCy: v3.1.0: New pipelines for Catalan & Danish, SpanCategorizer for arbitrary overlapping spans, use predicted annotations during training, bug fixes & more*. Zenodo, July.
- Neumann, Gerhard** (1970) Die ‘Absolute’ Metapher. Ein Abgrenzungsversuch am Beispiel Stéphane Mallarmés und Paul Celans. *Poetica*, 3(1):188–249.
- Ortiz Suárez, Pedro Javier / Romary, Laurent / Sagot, Benoît** (2020): A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, S. 1703–1714, Online. Association for Computational Linguistics.
- Sanders, Daniel** (1865): *Wörterbuch der Deutschen Sprache. Mit Belegen von Luther bis auf die Gegenwart*. Wigand, Leipzig, 3rd edition.
- Skirl, Helge / Schwarz-Friesel, Monika** (2013): *Metapher*. Winter, Heidelberg, 2nd edition.
- Specht, Benjamin** (2014): Epoche und Metapher Systematik und Geschichte kultureller Bildlichkeit. Einleitung. In Benjamin Specht, editor, *Epoche und Metapher*, S. 1–22. DE GRUYTER, Berlin, Boston.
- Specht, Benjamin** (2017): *Wurzel allen Denkens und Redens. Die Metapher in Wissenschaft, Weltanschauung, Poetik und Lyrik um 1900*. Winter, Heidelberg.
- Stockinger, Claudia** (2010): *Das 19. Jahrhundert. Zeitalter des Realismus*. Berlin.
- Zsubert, Benjamin / Cole, Jennifer / Monaco, Claudia / Drozdov, Ignat** (2019): Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific Reports*, 9(1):8914.

Tsvetkov, Yulia / Boytsov, Leonid / Gershman, Anatole / Nyberg, Eric / Dyer, Chris (2014): Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Weinrich, Harald (1963): Semantik der kühnen Metapher. *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte*, 37:325–344.

Hackathons als kollektiv-kreative Bildungsereignisse Ein Konzept zur Gestaltung offener Lehrveranstaltungen in den Digital Humanities

Mischke, Dennis

dennis.mischke@fu-berlin.de
Freie Universität Berlin, Germany

Trilcke, Peer

trilcke@uni-potsdam.de
Universität Potsdam, Germany

Sluyter-Gäthje, Henny

sluytergaeth@uni-potsdam.de
Universität Potsdam, Germany

Ausgangslage: Die DH als Motor der digitalen Transformation der Hochschullehre

Die digitale Transformation der Geisteswissenschaften vollzieht sich bereits seit einiger Zeit maßgeblich unter dem Dach der “Digital Humanities” (DH). Dabei profiliert sich dieses interdisziplinäre Forschungsfeld als eine eigenständige Disziplin im Grenzbereich der Informatik (Jannidis et. al. 2017). Die “radical Interdisciplinarity” (Ramsay 2011: 83) dieses Wissens- und Technologietransfers benötigt indes völlig neue Methodenkenntnisse, Fähigkeiten und Fertigkeiten, die allgemein als “Digital” (Hinrichsen et al. 2013) und “Data” (Heidrich et al. 2018) oder auch “Coding Literacies” (Vee 2017) verstanden werden. Das Konzept der “Digital Literacy” umfasst sowohl das kritische Lesen, Bewerten, Verarbeiten und Kommunizieren in digitalen Umgebungen (Hinrichsen et al. 2013) als auch die Fähigkeit komplexe rechner-gestützte Tools und Architekturen – auch ohne nutzerfreundliche GUIs – verstehen und angemessen verwenden zu können (Vee 2017). Die “weite Interdisziplinarität” (Schneidewind 2013) der DH zwischen computationalen Disziplinen und geisteswissenschaftlichen Fächern erfordert darüber hinaus ebenfalls reflexive Kompetenzen im Sinne eines “Knowledge Brokers” (Nowotny 2011). Die digitale Transformation der Metho-

den, Gegenstände und Forschungspraktiken zwingt Forschende und Studierende gleichermaßen dazu, neu erworbenes digitales und computationales Know-How in traditionelle Arbeitsweisen und Wissenschaftsformate der Humanities zu integrieren.

Insbesondere die kreativen, kollektiven und – im Sinne einer “maker culture” – bisweilen stark praxis- und produktorientierten Arbeitsweisen der DH sind hochschuldidaktisch bislang gar nicht oder nicht ausreichend reflektiert und aufgearbeitet worden. Im Rahmen des BMBF-Projektes “FoLD” (“Forschen | Lernen Digital”) entwickeln und erproben wir Konzepte zur fachübergreifenden Vermittlung digitaler Kompetenzen anhand von Workflows der digitalen Literaturwissenschaft und in Gestalt von kollektiv-kreativen Lehrveranstaltungen des Formats ‘Hackathons’. Die Kombination von einerseits stark durchstrukturierten und didaktisch geskripteten Trainings und Seminaren mit kreativen, offenen und kollektiven Arbeitsprozessen im Rahmen von Hackathons verbindet die wissenschaftliche DH-Ausbildung mit den agilen und interdisziplinären Arbeitsweisen und Projektformen, wie sie sowohl in der Digitalwirtschaft wie auch in digitalkulturellen NGOs oder Gedächtniseinrichtungen praktiziert werden.

Der Beitrag stellt unsere Idee von Hackathons als ein ereignisbasiertes Konzept des forschenden und kreativen Lernens vor und flankiert damit unsere hochschuldidaktische Konzeptualisierung der Workflow-basierten und stärker geskripteten Formate (Mischke, Trilcke & Sluyter-Gäthje 2021). Gleichzeitig berichten wir in zwei Fallvignetten von einem Hackathon zum Thema „Kulturdaten|Datenkulturen“, durchgeführt im Sommersemester 2021 an der Universität Potsdam.

Stand der Forschung: DH-Didaktik

Die Varianz und Dynamik DH-affiner Methoden und Werkzeuge in hochschuldidaktische Ansätze zu übersetzen, ist in Anbetracht der schnellen und breiten Entwicklung des Feldes eine Herausforderung. In der bestehenden Forschung zu DH-Didaktik wurden daher vor allem die digitalen Dimensionen verwandter Fachdidaktiken analysiert und versucht, wissenschaftliche Ansätze und Erkenntnisse insbesondere aus Medien- und Informatik-nahen Didaktiken für die DH fruchtbar zu machen (Bender 2016). Darüber hinaus finden sich Reflexionen und Anwendungen des Forschenden Lernens, um vor allem die Verbindung von Fachwissenschaften und digitalen Methoden zu stützen (Flemming 2017). Forschendes Lernen, so Flemming (2017), eignet sich insbesondere dafür, digitale Kompetenzen und die Ausbildung einer fachlichen Expertise zu verbinden; vor allem wenn das Lehr-Lernszenario kooperativ organisiert und digital unterstützt wird. Für Clement (2012) muss DH-Pädagogik in Anlehnung an die kollaborativen Arbeitsweisen aus der Forschungspraxis ebenso projekt-basiert und an der Förderung neuer Formen der Literacy ausgerichtet sein. Der Ansatz von Kirschenbaum (2010) geht sogar davon aus, dass eine DH-Didaktik gezielt ein sog. „living lab“ – ein interdisziplinäres Netzwerk von Forschenden und Studierenden – herstellen und befördern muss. Crane (2012) betont, dass diese “laboratory culture” der DH auch eine neue Kultur des Lehrens und Lernens hervorbringen kann, in der Studierende essentielle Beiträge zur Forschungsarbeit liefern können, wenn sie angemessen unterstützt, angeleitet und in die Forschung integriert werden. Eine entsprechende Labor-Didaktik, die Arbeitsweisen aus der Softwareentwicklung oder der Start-Up Szene nutzt, könnte Studierende und Lehrende im Umgang mit neuen Technologien schulen und geisteswissenschaftliche Lehr- und Arbeitsweisen um die “maker-culture” digitaler “makerspaces” (Sayers 2017) erweitern.

Zugleich bietet sich – in dezidiert abgegrenzter Weise – der an marktwirtschaftlich Effektivitätsmaßgaben orientierten Digitalwirtschaft – für die DH die Möglichkeit, in diese Arbeitsweisen die geisteswissenschaftlichen Praktiken sowohl der Reflexion und der Kritik einzubringen, als auch etwa ergebnisoffene, dabei kreativitätsaffine Denk- und Praxisräume im Sinne der Artistic Research oder des Design Thinking aufzugreifen (Schmidberger, Wippermann 2018). In diesem Sinne könnte sich eine DH-Didaktik auch aus der produktiven Zusammenführung von digitalwirtschaftlichen Arbeitsweisen der kollaborativen Produktion mit den offenen, reflexiv-kritischen Praxisformen der Geisteswissenschaften speisen.

Hackathons als Ereignisse kollektiver Kreativität

Der folgende, exemplarisch vorgehende und programmatisch-reflexiv ausgerichtete Praxisbericht aus dem lehrorientierten Forschungsprojekt “FoLD” setzt hier an. Hackathons werden dabei als praxeologische Intervention (Trilcke & Fischer 2018) begriffen, die die technisch-strikte Ausbildung von Coding Literacy in der DH-Lehre flankieren und dabei Prozesse kollektiver Kreativität anregen und einüben. Das Format des Hackathons, das insbesondere von IT-Unternehmen, der Open Source-Community, im GLAM-Sektor sowie in der DH-Community für die sprint-artige (Fort-)Entwicklung von (meist digitalen) Produkten sowie zur agilen Projektarbeit schon seit längerer Zeit genutzt wird, zeichnet sich vor allem dadurch aus, dass es die kollaborativen Arbeitsbedingungen eines Labors temporär nachbildet (Trilcke & Fischer 2018). Ganz im Sinne von Kirschenbaums’ ‘living lab’ (2010) soll ein Hackathon dabei als ein lebendiges Experiment Studierenden die Gelegenheit geben, eigene – d.h. allgemein digitale und DH-spezifische – Kompetenzen (KMK 2017) anzuwenden, neue Kompetenzen zu entdecken wie zu erwerben und diese im Kontext eines offenen, aber didaktisch orchestrierten Gruppenprozesses intensiv zu erleben. Die zeitliche, räumliche und soziale Intensität des Hackathons soll dabei – temporär verdichtet – ein kreatives Flow-Erleben erzeugen, welches Studierende und Lehrende entsprechend ihrer Interessen, Neigungen und Kompetenzen in einem Prozess des gemeinsamen Produzierens und “Machens” zusammenführt. Das übergreifende didaktische Ziel der im FoLD-Projekt konzipierten und durchgeführten Hackathons lässt sich dabei wie folgt fassen:

Hackathons in der DH-Lehre initiieren und begleiten digitale und DH-spezifische Kompetenzentwicklungsprozesse, die auf den reflektierten Einsatz einer fachwissenschaftlich fundierten, d.h. kritisch und analytisch geschulten, offenen kollektiven Kreativität zielen.

Diese didaktische Zielsetzung soll im Folgenden anhand von zwei Fallvignetten aus Hackathons umrissen werden, die wir an der Universität Potsdam im Rahmen des FoLD-Projektes im SoSe 2021 durchgeführt haben. Mit den beiden Fallvignetten wollen wir unsere Ansätze zur Lösung von zentralen didaktischen Herausforderungen während eines Hackathons präsentieren: erstens der Einübung offener Prozesse (Kreativität) im prinzipiell geschlossenen (weil zielbasierten) Kontext einer universitären Lehrveranstaltung; und zweitens des Zusammenspiels von kollaborativen und kompetitiven Dimensionen (Kollektivität). Anhand einer ersten Evaluation der Ergebnisse, basierend auf anonymisierten Interviews mit den Teilnehmer:innen skizzieren wir vorläufige Erkenntnisse und „lessons learned“.

Fallvignetten

Fallvignette 1: Offene Prozesse schulen (Kreativität)

Akademische Hackathons in der Hochschullehre ermöglichen eine offene Auseinandersetzung mit gegenwärtigen Problemen und Themen der Kultur im Allgemeinen wie der Geisteswissenschaften im Besonderen (Datafizierung der Kultur, Kulturdaten, Digitale Identität, Datenkommerzialisierung, Data Tracking, Umgang mit Hate Speech, gesellschaftliche Partizipation in der digitalen Welt, Digitale Bildung, Information Overload, Climate Change etc.). Dabei werden die Studierenden in die Lage versetzt, die eigenen fachwissenschaftlichen Kompetenzen in der Anwendung auf ihre eigene Gegenwartskultur wie ihre eigenen wissenschaftlichen Handlungskontexte aktiv zu entdecken und sowohl reflektiert wie kreativ für die strukturierte und kollektive Beschäftigung mit gegenwärtigen Phänomenen einzusetzen. Eine solche Einführung in die produktionsorientierte Forschungs- und Arbeitsweise vermittelt eine gegenwartsbezogene Wissenschaftskultur in “Echtzeit”.

Problem: Wie können komplexe Fragestellung der unmittelbaren Gegenwart (z.B. Digitalisierung oder Klimawandel) im Kontext des forschenden Lernens und Lehrens entdeckt und gemeinsam mit Studierenden bearbeitet werden? Wie können Studierende in geistes- und kulturwissenschaftlichen Master-Studiengängen (Germanistik und Anglistik/Amerikanistik) didaktisch an diese interdisziplinäre und thematische Offenheit herangeführt werden?

Lösung: Bereits vor der Vorbereitung und Durchführung eines thematisch spezifischen Hackathons als Blockseminar üben Studierende die offene, problem- und produktionsorientierte Lehr- und Arbeitsweise, indem sie in einem Miniaturdurchlauf den Prozess des Hackathons angeleitet durch eine Individual-Challenge durchspielen. Zentral bei diesen Individual-Challenges ist, dass die Studierenden das Thema ihrer Challenge vollständig selbst entwickeln. Die Dozierenden unterstützen und rahmen lediglich die kreativen Prozesse der Studierenden, ein Thema zu finden und auszuwählen, das ihren Interessen, aber auch ihren individuellen und heterogenen Vorwissensbeständen und Kompetenzen entspricht.

Fallvignette 2: Challenge-Auswahl und Teambildung (Kollektivität)

Der Ereignischarakter von Hackathons verdichtet die Lern-, Forschungs-, und Arbeitsaktivitäten in einer aktivierenden Hauptarbeitsphase; eines in unserem Fall zweitägigen Hackathons. Zuvor müssen die Studierenden jedoch zunächst geeignete Projektideen entwickeln und potentielle Mitarbeiter:innen für ihr Projekt gewinnen. Dabei kommt bereits ein zentraler aber produktiver Zielkonflikt von Hackathons zum Tragen: die Spannung zwischen dem Zusammenarbeiten auf der einen Seite (kollaborative Dimension) und dem Wettbewerb der Ideen auf der anderen Seite (kompetitive Dimension).

Problem: In einer Sitzung nach einer Vorbereitungsphase haben alle Studierende ihre Ideen vorgestellt und ‘gepitched’. Es können jedoch nicht alle Projekte realisiert werden. Einige Studierende müssen sich von ihren Projektideen verabschieden (Abstand nehmen) und sich anderen Gruppen anschließen. Wie kann diese soziale und kompetitive Dimension des Hackathons organisiert und

orchestriert werden, ohne dass die kollaborative Dimension dadurch gefährdet wird?

Lösung: Um das Problem sozial zu lösen, hat sich eine Umsetzung dieses Verteilungs- und Aushandlungsprozesses in einer räumlichen Umgebung als günstig erwiesen. Im Sommersemester 2021 haben wir die Umsetzung dieses Prozesses daher in Gather.town realisiert. Gather.town bietet als gamifizierter und räumlich 2D organisierter Videochat mittels Avataren (Embodiment) die Möglichkeit, Gespräche, Teilnahme und Kooperationsbereitschaft durch räumliche Nähe zu strukturieren und zu signalisieren. Studierende mit Projektpitches konnten sich so im Raum positionieren und auf diese Weise ein Angebot an interessierte Studierende unterbreiten, sich zu versammeln. Zur weiteren Teambildung sowie als Infrastrukturmaßnahme haben wir den Arbeitsgruppen einen DSGVO-konformen Live-Chat-Dienst (Mattermost) zur Verfügung gestellt. Auf diese Weise konnten die Studierenden notwendige arbeitsorganisatorische Rollen aushandeln, eine Team-Etikette ausbilden und eine niedrigschwellige und effiziente Kommunikationspraxis aufbauen. Aus Perspektive der Lehrenden erwies sich die technische Realisierung der Lehr-Lernkommunikation via Mattermost als besonders vorteilhaft, weil die entstandenen Gruppenprozesse und individuellen Arbeitsphasen optimal beobachtet, moderiert und motivierend unterstützt werden konnten. In zusätzlichen privaten Channels konnten sich Studierende jedoch jederzeit dem Dozierendenblick entziehen (auch auf Mattermost). Sowohl die Lösung über die räumliche Situierung (Gather.town) als auch die chatbasierte Organisation (Mattermost) der Projektkommunikation hat den Zielkonflikt zwischen Kollaboration und Kompetition spielerisch auflösen und produktiv umwenden können.

Vorläufige Auswertung

Die Teilnehmer:innen unseres Hackathons waren Studierende aus unterschiedlichen Masterstudiengängen der Fächer Anglistik | Amerikanistik und Germanistik sowie als Gäste der Fächer Archivwissenschaften (1 Studierender) und Wirtschaftsinformatik (1 Studierender). Zur Dokumentation der Lernfortschritte der Studierenden wurden systematisch Vorwissensbestände, Haltungen und Selbsteinschätzungen bezüglich digitaler Kompetenzen erhoben. Zusätzlich wurden die Teilnehmer:innen während und nach der Veranstaltung mit Online-Fragebögen zu ihren persönlichen Eindrücken und Wahrnehmungen befragt. Dazu haben wir eine quantitative und qualitative Befragung erarbeitet, die sich gegenwärtig noch in der abschließenden Auswertung befindet. Eine vorläufige Auswertung der qualitativen Erhebung lässt sich auf folgende Punkte verdichten:

- Insbesondere der Modus des intensiven kollektiven Arbeitens in Gemeinschaft „nach der langen Zeit der Isolation“ der Pandemie wurde trotz digitaler Umsetzung als sehr angenehm und produktiv empfunden. Des weiteren führte die Interdisziplinarität der Gruppen zu einer Erweiterung der Sichtweise auf den eigenen Fachbereich und zur Bewusstwerdung der eigenen Fähigkeiten.
- Mit Blick auf die Herausforderung der offenen Arbeitssituation (Themenfindung und Selbstorganisation) des Hackathons sahen sich viele Studierende zunächst überfordert, konnten sich im Verlauf der Veranstaltung durch die intensive Beratung jedoch gut zurechtfinden. Teilnehmer:innen berichteten positiv davon, dass gerade diese Offenheit Kreativität gefordert hat, was zur tiefergehenden Beschäftigung mit dem ausgewählten Projekt geführt hat.
- Als besonders erwähnenswert gaben alle Teilnehmer:innen den passend orchestrierten Einsatz von Mattermost an, dessen Multikanal-Kommunikation einer ständigen Einbindung in das

Geschehen sowohl in die einzelnen Arbeitsgruppen als auch in Gesamtgeschehen des Hackathons zuträglich war

Ausblick

Der Beitrag widmet sich programmatisch und konzeptionell einer didaktischen Aufbereitung, Erforschung und Erprobung des Formates “Hackathon” als hochschuldidaktisches Format im Kontext der Digital Humanities. Wir berichten exemplarisch von Überlegungen und Erfahrungen mit lehr-orientierten DH-Hackathons und deren bildungswissenschaftlicher Rahmung und Verzahnung mit stärker strukturfokussierten Lehrveranstaltungen zur DH-Methodenbildung. Erste Konzeptions- und Erprobungsphasen des Projekts sind abgeschlossen und werden zurzeit weiter umfassend bildungswissenschaftlich ausgewertet. Die finalen Ergebnisse dieser Begleitstudie sollen auf der DHd2022 vorgestellt werden.

Bibliographie

- Bender, Michael** (2016): “Digitalität in den Fachdidaktiken (DFd) – neues Projekt an der TU Darmstadt”, in: *DHdBlog. Digital Humanities im deutschsprachigen Raum*. <https://dhd-blog.org/?p=6812> [letzter Zugriff 14. Juli 2021].
- Clement, Tanya** (2012). “Multiliteracies in the Undergraduate Digital Humanities Curriculum”, in: Brett D. Hirsch (eds.): *Digital Humanities Pedagogy. Practices, Principles and Politics*. Cambridge: Open Book Publishers 365-388. <https://books.openedition.org/obp/1656> [letzter Zugriff 14. Juli 2021].
- Crane, Gregory** (2020): “Greek, Latin and a Global Dialogue among Civilizations”, in: *The Center for Hellenic Studies* [Website]. <https://chs.harvard.edu/gregory-crane-greek-latin-and-a-global-dialogue-among-civilizations/> [letzter Zugriff 14. Juli 2021].
- Flemming, Tobias** (2017): “Lernen an Handschriften. Studierende als Experten gewinnen”, in: *Forum Exegese und Hochschuldidaktik* 2.2: 69-79
- Hinrichsen, Juliet / Coombs, Antony** (2013): “The five resources of critical digital literacy: a framework for curriculum integration”, in: *Research in Learning Technology*. <https://doi.org/10.3402/rlt.v21.21334> [letzter Zugriff 14. Juli 2021].
- Heidrich, Jens / Bauer, Pascal / Krupka, Daniel** (2018): *Strukturen und Kollaborationsformen zur Vermittlung von Data-Literacy-Kompetenz* (= Hochschulforum Digitalisierung, Arbeitspapier Nr. 32). Berlin: Hochschulforum Digitalisierung. <https://doi.org/10.5281/zenodo.1408600> [letzter Zugriff 14. Juli 2021].
- Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte** (2017): *Digital Humanities. Eine Einführung*. Stuttgart: Metzler.
- KMK. Kultusministerkonferenz** (2017): *Bildung in der Digitalen Welt*. https://www.kmk.org/fileadmin/Dateien/pdf/Presse-UndAktuelles/2017/KMK_Kompetenzen_-_Bildung_in_der_digitalen_Welt_Web.html [letzter Zugriff 14. Juli 2021].
- Kirschenbaum, Matthew** (2010): “What is Digital Humanities and What’s it Doing in the English Department?”, in: *ADE Bulletin* 150: 55–61. <https://www.doi.org/10.1632/ade.150.55> [letzter Zugriff 14. Juli 2021].
- Mischke, Dennis / Trilcke, Peer / Sluyter-Gäthje, Henny** (2021): “Workflow-basiertes Lernen in den Geisteswissenschaften: digitale Kompetenzen forschungsnah vermitteln”, in: *Bildung in der Digitalen Transformation - GMW2021 Proceedings* (zur Publikation angenommen).

Nowotny, Helga / Scott, Peter / Gibbons, Michael (2001): *Rethinking Science-Knowledge and the Public in an Age of Uncertainty*. Cambridge, UK: Polity.

Sayers, Jentery (2018): *Making Things and Drawing Boundaries. Experiments in the Digital Humanities*. Minneapolis: Minnesota UP. <https://doi.org/10.5749/j.ctt1pwt6wq> [letzter Zugriff 14. Juli 2021].

Schneidewind, Uwe / Singer-Brodowski, Mandy (2013): *Transformative Wissenschaft. Klimawandel im Deutschen Wissenschafts- und Hochschulsystem*. Marburg: Metropolis.

Schmidberger, Iris / Wippermann, Sven (2018): *Design Thinking – ein Innovationsansatz für den Bildungsbereich?*. https://www.ph-ludwigsburg.de/fileadmin/phlb/hochschule/fakultaet1/bildungsmanagement/Bildungsmanagement/06_Design_Thinking/Design-Thinking-Innovationsansatz_Bildungsbereich.pdf [letzter Zugriff 14. Juli 2021].

Trilcke, Peer / Fischer, Frank (2018): "Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen." In: *Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden*. Hg. von Martin Huber / Sybille Krämer. Sonderband der Zeitschrift für digitale Geisteswissenschaften, 3. text/html Format. DOI: 10.17175/sb003_003 [letzter Zugriff 14. Juli 2021].

Ramsay, Stephen (2011): *Reading Machines. Toward an Algorithmic Criticism*. Champaign: Illinois UP.

Vee, Anette (2017): *Coding Literacy: How Computer Literacy is Changing Writing*. Boston: MIT.

Handwritten Text Recognition und Word Mover's Distance als Grundlagen der digitalen Edition "Die Kindheit Jesu Konrads von Fußesbrunnen"

Tomasek, Stefan

stefan.tomasek@germanistik.uni-wuerzburg.de
Universität Würzburg, Germany

Reul, Christian

christian.reul@uni-wuerzburg.de
Universität Würzburg, Germany

Wehner, Maximilian

maximilian.wehner@uni-wuerzburg.de
Universität Würzburg, Germany

Gegenstand und Fragestellung des Vortrags

OCR4all und HTR

Zur Zeit entsteht an der JMU Würzburg in einem Kooperationsprojekt zwischen dem Lehrstuhl für ältere deutsche Philologie und dem Zentrum für Philologie und Digitalität ein HTR-Projekt (Handwritten Text Recognition) zur Erfassung mittelhochdeutscher (mhd.) und frühneuhochdeutscher (frnhd.) Handschriften (Hss.) des 11.-15. Jh.s. Ausgangspunkt waren die zunächst in Transkribus¹ erstellten Transkriptionsdaten des Würzburger Editionsprojektes zur 'Kindheit Jesu' Konrads von Fußesbrunnen. Dieses Projekt baut neben den eigentlichen Editionstexten auf einer umfangreichen Datenmenge nicht normalisierten Mhd.s auf (s. u.). Im Open Source Bereich, sowohl mit Blick auf die automatische Texterkennung im Allgemeinen² als auch bei der Erkennung vormoderner volkssprachiger Hss., gab es in letzter Zeit erhebliche Fortschritte. Daher kommt mittlerweile für die Erstellung der Datengrundlage für das Editionsprojekt das an Frühdrucken³ erarbeitete, frei verfügbare Open Source Tool OCR4all⁴ zum Einsatz.

Die Grundidee von OCR4all ist es, insbesondere technisch weniger versierten NutzerInnen die Möglichkeit zu geben, anspruchsvolle historische Drucke und Handschriften selbstständig und in höchster Qualität zu erfassen. Dies wird v. a. dadurch ermöglicht, dass einzelne, auf unterschiedliche Schritte des OCR Workflows (Optical Character Recognition) spezialisierte (Kommandozeilen-)Werkzeuge in einem leicht zu installierenden Tool gekapselt und über eine einheitliche Benutzeroberfläche zugänglich gemacht werden. Die Konzeption als Client/Server-Anwendung und die Auslieferung mittels einer Containerlösung erlaubt dabei einen flexiblen Einsatz sowohl lokal beim Einzelnutzer als auch das kollaborative Arbeiten über eine zentralisierte Serverinstanz. Die Bearbeitung eines Werkes kann ebenfalls sehr flexibel erfolgen und an das vorliegende Material und die eigenen Ansprüche angepasst werden. Generell ist ein vollautomatischer Durchlauf möglich, dieser kann allerdings nach jedem Teilschritt unterbrochen und die Ergebnisse kontrolliert und bei Bedarf manuell nachkorrigiert werden, um Folgefehler zu vermeiden.

Im Vergleich zu herkömmlichen OCR-Verfahren ist bereits das Erfassen von Frühdrucken besonders anspruchsvoll, da hier z.T. komplexe Layoutstrukturen vorliegen und der Druck- bzw. Erhaltungszustand erheblich variiert. Zudem unterscheiden sich die verwendeten Drucktypen und Schriftarten innerhalb eines Werkes und zwischen unterschiedlichen Werken. Beide Kriterien gelten für die Erfassung historischer Handschriften in verstärktem Maße, da sich das Layout z.T. innerhalb der gleichen Hs. deutlich ausdifferenziert, die mittelalterlichen Schreiber vielfältige Schreibvarianten verwenden und sich das Schriftbild zwischen den einzelnen Schreibern erheblich unterscheidet bzw. im historischen Längsschnitt weiterentwickelt wurde (siehe hierzu unten).

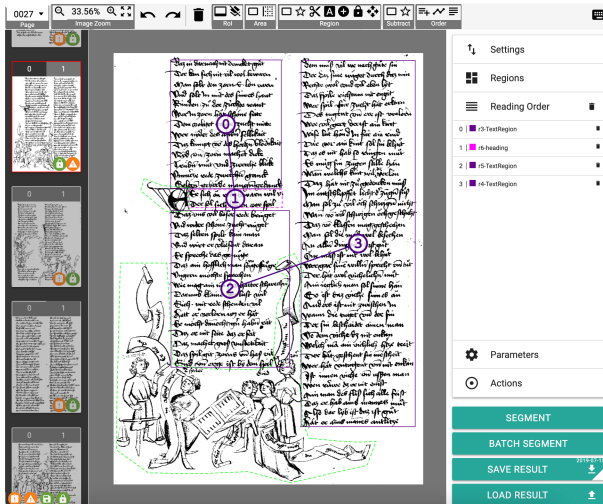


Abb. 1: OCR4all ermöglicht die Segmentierung und Typisierung der Layoutelemente einer Textseite ebenso wie die Festlegung deren Lesereihenfolge.



Abb. 2.: Im Kontext der 'Post Correction' kann automatisch generierter Text mithilfe einer virtuellen Tastatur (rechts) zeichengetreu nachkorrigiert werden.

Digitales Editionsprojekt "Die Kindheit Jesu Konrads von Fußesbrunnen" und Levenshtein-Distanzen bzw. Word Mover's Distance

Das Würzburger HTR-Projekt ist verzahnt mit dem digitalen Editionsprojekt der „Kindheit Jesu Konrads von Fußesbrunnen“ (KJ). Dieses in mhd. Reimpaarversen verfasste apokryphe Kindheitsevangelium (entstanden um 1200) weist erhebliche Varianten zwischen allen Textzeugen auf. Im Editionsprojekt wird daher die vollständige Überlieferungssituation des Textes (vier Haupthss., sieben Fragmente, vier Sekundärzeugnisse) synoptisch abgebildet. Um diese Synopse dennoch lesbar zu halten, soll mit Hilfe von Levenshtein-Distanzen (LevD) und einem Word-Embedding-Verfahren⁵ ein Filtersystem ermöglicht werden, mit dem die Genauigkeit des Anzeigemodus⁷ von den BenutzerInnen der Edition selbst festgelegt werden kann. Zum einen kann über die LevD zeichengenau gefiltert werden. Durch Teilnormalisierungen des Textes sollen zusätzlich verschiedene Parameter der Textvarianz (z.B. orthographische Varianz, dialektale Varianz, Graphemvarianten etc.) herauszufiltern sein.⁶ Mit diesen LevD-Werten kombiniert werden semantische Distanzwerte: Mithilfe von fastText⁷ werden die mhd. Begriffe in ein n-dimensionales Vektorensystem übertragen, in dem durch Word Mover's Distance⁸ (WMD) die semantischen Distanzen zwischen den mhd. Begriffen bestimmt werden. Damit sind jedem Verspaar des mhd.

Textes in jeder Hss.-Kombination mehrere Distanzwerte zugewiesen, die wiederum miteinander kombinierbar sind. So ist es für die BenutzerInnen der Edition möglich, die Anzeigegenauigkeit jeweils dem eigenen Leseinteresse entsprechend zu modellieren: Von kleinteiligen Zeichenvarianten über phonetische Unterschiede bis zu abgestuften Bedeutungsvarianten mit niedriger oder hoher semantischer Differenz bzw. Zusatz- oder Fehlversen kann voreingestellt werden, wie genau die Parallelüberlieferung zu der gewählten Lesehandschrift des Textes angezeigt werden soll. Dieser Genauigkeitsfaktor kann jederzeit dynamisch angepasst werden. Gleichzeitig sind die Distanzwerte in die Suchfunktion der Edition integriert, wodurch die historische Überlieferungssituation der KJ nachvollziehbar und abbildbar wird (siehe Abb. 3).

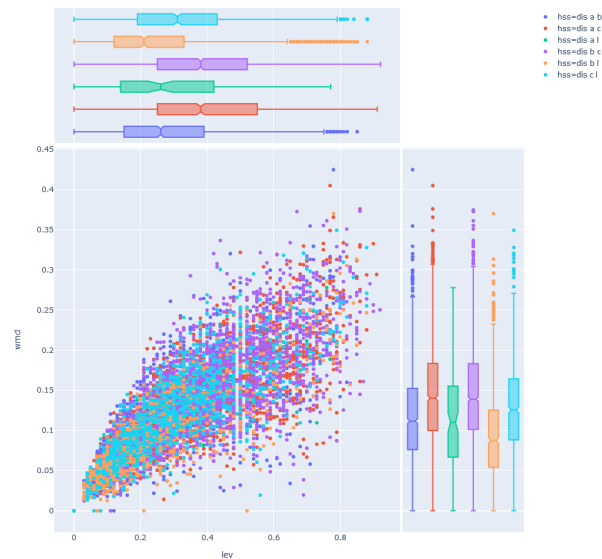


Abb. 3: Die Überlieferungssituation der vier Haupthss. der KJ: Abgebildet sind die Nähe-Distanz-Verhältnisse aller Verspaare in allen Hss.-Kombinationen nach LevD und WMD.

Die LevD und WMD-Werte haben sich für mhd. Texte als prinzipiell anwendbar gezeigt. Da die überwiegende Mehrheit der bisherigen Texteditionen, auf deren Grundlage das WMD-Modell bisher trainiert wurde,⁹ einen normalisierten mhd. Sprachstand aufweisen, zeigen sich bei zeichengenauen (nicht normalisierten) Texten dagegen noch deutliche Schwächen: Bestimmte Schreibvarianten, die im konstruierten "Normalmittelhochdeutsch" der traditionellen Editionen nicht vorkommen,¹⁰ werden durch die WMDs nicht als Synonyme erkannt und dementsprechend mit hohen semantischen Distanzen ausgezeichnet.

Da die WMDs diesen Anwendungsfall aber prinzipiell abdecken und gleichzeitig auf die (unnormalisierten) Transkriptionsdateien des genannten HTR-Projektes anwendbar sein sollen (s.u.), muss das Trainingscorpus erheblich ausdifferenziert werden. Daher sind beide Teilprojekte doppelt miteinander verzahnt: Die vollständig vorliegenden Transkriptionsdaten aller Hss. der KJ bilden das Grundmodell für OCR4all, auf dem das HTR-Modell trainiert wird. Alle Ground Truth Daten (GT), die aus den werkspezifischen Modellen generiert werden, bilden umgekehrt das Trainingscorpus für die WMDs der KJ. Die Filterstruktur der KJ wird damit auf einer wesentlich umfangreicheren "Datenbank nicht normalisiertes Mittelhochdeutsch" aufbauen, die ihrerseits aus HTR-Daten besteht.

Trotz insgesamt noch relativ geringer Mengen GT wurden durch das HTR-Projekt innerhalb kurzer Zeit bereits gute Texterkennungsergebnisse erzielt: So konnte in ersten Testdurchläufen auf einer anspruchsvollen Bastarden-Handschrift des 15. Jh.s¹¹ mit etwa 1.500 Zeilen GT eine Zeichenfehlerrate von etwas mehr als 3% erreicht werden. Auf einer Gotischen Buchschrift des 13. Jh.s¹² wurde mit nur knapp 600 Zeilen sogar eine Fehlerrate von ca. 2% erreicht. Durch die im nächsten Arbeitsschritt angestrebte Erweiterung der Trainingsmenge und weitere Ausdifferenzierung des Trainingsmodells sind werkspezifische Transkriptionsmodelle in Reichweite gelangt, mit denen relativ schnell auch umfängliche Handschriften bzw. Handschriftencorpora erschließbar sind. Perspektivisch sollen auch die bisher erstellten Trainingsdaten¹³ frei zur Verfügung gestellt werden (sofern dies die jeweiligen Bildrechte zulassen). Veröffentlicht werden darüber hinaus die derzeit entstehenden gemischten HTR Modelle, die sowohl externen NutzerInnen für eine out-of-the-box Anwendung zur Verfügung stehen als auch die Grundlage weiterer Trainingsprozesse darstellen können. Aus dem Würzburger HTR-Projekt gehen damit drei mögliche Anwendungsfälle hervor, die im Folgenden skizziert werden.

Anwendungsfälle

HTR-generierte Texte als Vorstufe für Editionsprojekte

Den gewissermaßen klassischen Anwendungsfall von (HTR-)Transkriptionen in der älteren deutschen Literaturwissenschaft stellt die zeichengenaue Umschrift als Vorstufe für (digitale) Editionsprojekte dar. Diplomatische Volltranskriptionen werden hier z.B. als Grundlage für stemmatologische Analysen verwendet,¹⁴ auf denen die Auswahl des Editionstextes fußt. Sie bilden den Ausgangspunkt für normalisierte Textstufen und stehen v.a. in den neueren digitalen Editionen gleichberechtigt neben dem normalisierten Lesetext.¹⁵ Die diplomatischen Texte erweitern in Hybrideditionen den printbasierten Editionstext¹⁶ oder ergänzen durch die Darstellung von Text-Bild-Beziehungen die traditionellen Editionen¹⁷ etc. Damit kommt den Transkriptionen der mhd. Hss. in der Editionsphilologie ein erheblicher Stellenwert zu.

Das Würzburger Projekt OCR4all soll hierfür die Datengrundlage für weitere Korrektur- und Bearbeitungsschritte liefern. Um hierbei die Effizienz zu maximieren, wird bei der GT Erstellung iterativ vorgegangen: Nach der initialen Erkennung einiger weniger Seiten mit einem gemischten Grundmodell erfolgt die manuelle Nachkorrektur, die zunächst noch vergleichsweise zeitaufwendig ist (je höher die Fehlerrate, desto größer der Korrekturaufwand). Die so gewonnene werkspezifische GT wird im Anschluss für das Training eines ersten werkspezifischen Modells eingesetzt. Dieses wiederum wird anschließend auf weitere Seiten angewendet und das, im Normalfall bereits deutlich bessere, Ergebnis erneut korrigiert. Dieser Vorgang wird iterativ wiederholt, bis entweder die gesamte Handschrift manuell nachkorrigiert wurde oder ein ausreichend gutes Modell vorliegt, mit dem die übrigen Seiten erkannt werden können. Die Anzahl der beschriebenen Trainings- und Korrekturiterationen sowie der mit ihnen verbundene zeitliche Aufwand hängen stark vom zugrundeliegenden Material und den eigenen/projektspezifischen Qualitätsansprüchen ab. Der oben erwähnte Anwendungsfall einer Gotischen Buchhandschrift deutet einen für OCR/HTR-Modelle

typischen Verlauf an: Die initiale Zeichenfehlerrate des gemischten Grundmodells (11%), in dem die zu erfassende Hs. nicht enthalten war, ließ sich in einem werkspezifischen Modell durch die Korrektur und das Training von lediglich drei Seiten (72 Zeilen) bereits auf 3,6% zu reduzieren. In weiteren Iterationen folgten unter Verwendung von sechs, zwölf und 24 Seiten Verbesserungen auf 3,1%, 2,6% und schließlich 2,1%.

Mit diesem Verfahren lässt sich der Zeitaufwand für die Herstellung der Volltranskription einer Hs. erheblich reduzieren. Hierdurch sind nun Editionsprojekte möglich, die auch bei umfänglicher Überlieferungssituation alle Textzeugen transkribieren und in einer Datenbank zur Verfügung stellen können. Daraus resultiert aber das Folgeproblem, dass große Datenmengen entstehen, die ihrerseits von den HerausgeberInnen systematisiert werden müssen. Für diesen Normalfall mhd./frnhd. Texte (mehrere Hss. mit divergenter Überlieferungssituation) soll daher bereits nach der HTR-Transkription mit dem oben beschriebenen kombinierten Filtersystem (LevD und WMD) eine Darstellung der Überlieferungsstruktur geboten werden. Alle Varianten innerhalb der Überlieferung können so systematisch identifiziert und klassifiziert werden. Diese mathematisch nachvollziehbaren Klassifizierungen können wiederum bei der Beschreibung der Editionsrichtlinien als (für die NutzerInnen der Edition überprüfbare) Kriterien angegeben werden. Hierdurch lässt sich, je nach gewünschter Editionsform, beispielsweise die gewählte Leithandschrift begründen und der Anmerkungsapparat erstellen etc. Damit ist das Filtersystem also bereits beim Vorgang der Texterstellung nutzbar. Natürlich kann auch das für die KJ erstellte Filtersystem selbst für eine digitale Edition übernommen werden.

HTR-Texte als „neuer Texttyp“

Alle in die „Datenbank nicht normalisiertes Mittelhochdeutsch“ aufgenommenen HTR-Transkriptionen sollen frei zur Verfügung gestellt werden. Die Transkriptionen sind hierbei zeilengenau mit den Digitalfaksimile der jeweiligen Hss. verzahnt. Zudem können über OCR4all ständig neue, von spezifischen Fragestellungen abhängige Corpora generiert werden. Damit gelangt ein neuer Texttyp in den altgermanistischen wissenschaftlichen Diskurs. Dieser weist auf der einen Seite eine höhere Fehlerquote auf als umfänglich (händisch) korrigierte Editionstexte. Auf der anderen Seite bietet er aber (anders als die meisten herkömmlichen Editionen) einen unmittelbaren Zugriff auf die historischen Handschriften und kann so die Grundlage der Erschließung und der Durchsuchbarkeit mhd./frnhd. Hss. darstellen. HTR-Transkriptionen können daher als Schlüssel zur mittelalterlichen Hs. genutzt werden. Das ist in den Fällen besonders relevant, in denen keine vollständig adäquate Editionssituation vorliegt bzw. nicht alle Textzeugen in bereits bestehende Editionen eingegangen sind. Bei Fragestellungen, die durch normalisierte Editionen erschwert werden, können HTR-Transkriptionen zudem als Ergänzung der bestehenden Texteditionen herangezogen werden. Sie nehmen damit generell eine Mittelstellung zwischen der Edition und dem (Digital-)Faksimile ein. Im Vergleich mit der durch Normalisierungen und Konjekturen geprägten Editionspraxis der älteren deutschen Literaturwissenschaft kann beispielsweise auch die Frage aufgeworfen werden, welche Rolle die HerausgeberInnen mhd. Texte eigentlich für unsere moderne Wahrnehmung der Texte und des historischen Sprachstands spielen etc. HTR-Transkriptionen gewähren so einen Blick in die historische Situierung mhd. Texte, der weit über den traditionellen Zugang des kritischen Anmerkungsapparats hinausgeht.

Weitere Anwendungsgebiete von HTR-Transkriptionen und Levenshtein- bzw. WMD-Filtern

Das mit dem Editionsprojekt der KJ verzahnte Würzburger HTR-Projekt soll in drei Anwendungsbereichen Ergebnisse generieren, die für potentielle Folgeprojekte über den Standort Würzburg hinaus frei zur Verfügung gestellt werden: 1. Das gemischte HTR-Grundmodell kann als Grundlage für weitere werkspezifische Erkennungsmodelle verwendet werden, wodurch sich der Transkriptionsaufwand in entsprechenden (externen) Folgeprojekten erheblich reduziert. Hierdurch werden jenseits von Editonsprojekten Fragestellungen ermöglicht, die auf Grundlage der momentan zur Verfügung stehenden Textdaten gar nicht oder nur mit erheblichem Aufwand beantwortet werden könnten (s. u.). Gleichzeitig lässt sich das Grundmodell mit jedem Folgeprojekt (und einer entsprechenden Erweiterung der GT) weiter ausdifferenzieren. 2. Als Folge der GT-Erstellung wächst auch die "Datenbank nicht normalisiertes Mittelhochdeutsch" kontinuierlich an. Die entstehenden Daten können einerseits einschlägigen Datenbanken wie der "Mittelhochdeutsche Begriffsdatenbank" zur Verfügung gestellt werden. Andererseits besteht beispielsweise für corpusanalytische Fragestellungen freier Zugriff auf alle erfassten Texte, die dementsprechend zur Nachnutzung zur Verfügung stehen. 3. Das auf der "Datenbank nicht normalisiertes Mittelhochdeutsch" basierende WMD- und LevD-Filterssystem kann für diverse weitere Fragestellungen angewendet werden (z. B. für diverse Fassungsvergleiche; automatisch erstellbare, auf WMD-Distanzen basierende Textsynopsen o. ä.). Daher werden das HTR-Grundmodell, die "Datenbank nicht normalisiertes Mittelhochdeutsch" und die WMD-/LevD-Daten als Open Source Datenbank zur Verfügung gestellt. Aus diesen drei Anwendungsbereichen folgen weitere mögliche Fragestellungen, die auf dem HTR-Projekt bzw. WMD-/LevD-Projekt fußen. Diese können im Folgenden nur knapp skizziert werden:

1. Die meisten überlieferungsgeschichtlichen Fragestellungen benötigen mehr Datenmaterial, als ein herkömmlicher Lesetext mit Anmerkungsapparat zur Verfügung stellt. Für alle corpusanalytischen Zugänge, die nicht auf die Edition der Corpustexte zielen, ist es zentral, mit möglichst wenig Arbeitsaufwand spezifische Untersuchungscorpora aufbauen zu können. Das ist mit dem HTR-Grundmodell möglich. 2. Durch die Erschließung der mittelalterlichen Hss. sind neue sprachgeschichtliche Erkenntnisse zu erwarten, da mit der "Datenbank nicht normalisiertes Mittelhochdeutsch" deutlich mehr nicht normalisiertes Datenmaterial zur Verfügung gestellt werden kann. Die WMDs lassen hierbei beispielsweise neue Perspektiven auf die Semantik historischer Sprachstufen zu. 3. Stilometrische Analysen können durch diese Datenbank quantitativ ausgeweitet und mit den WMDs kombiniert werden.¹⁸ 4. Phraseologische Querschnitte innerhalb eines Untersuchungscorpus erscheinen durch die WMDs möglich.¹⁹ 5. Überkommene stemmatologische Setzungen sind durch breit angelegte, von HTR-Modellen gestützte Levenshtein- und WMD-Analysen überprüfbar.²⁰ 6. LevD und WMDs sind für neue Fassungsdefinitionen anwendbar etc.²¹

Bereits diese kursorischen Überlegungen machen deutlich, wie gewinnbringend digitale Methoden, Textkorpora und Editionen besonders für vormoderne Texte nutzbar gemacht werden können. Die interdisziplinäre Zusammenarbeit zwischen den philologischen Disziplinen und den Digital Humanities dürfte hierbei das Potential haben, neue Fragen hervorzubringen und gleichzeitig traditionelle Fragen der Mediävistik neu zu beantworten.

Fußnoten

1. Vgl. Kahle et al. 2017.
2. Vgl. z. B. die DFG-Förderinitiative OCR-D. URL: <https://ocr-d.de>.
3. Vgl. das digitale Editionsprojekt Narragonien digital. URL: <http://www.narragonien-digital.de>.
4. Vgl. Reul et al. 2019 und <http://ocr4all.de>.
5. Vgl. Kusner et al. 2015.
6. Vgl. Dimpel 2017.
7. Vgl. Mikolov et al. 2017; Bojanowski et al. 2017.
8. Vgl. Hung et al. 2016.
9. Verwendet wurde der Datensatz der Mittelhochdeutschen Begriffsdatenbank, vgl. <http://www.mhdbdb.sbg.ac.at/>.
10. Vgl. Kragl 2015.
11. Vgl. Thomasin von Zerclaere, Der Welsche Gast, München, Bayerische Staatsbibliothek, Cgm 571 (3. Viertel 15. Jh.), vgl. https://digi.ub.uni-heidelberg.de/diglit/bsb_cgm571.
12. Vgl. Priester Wernher, Driu liet von der maget, Krakau, Bibl. Jagiellonska, Berol. mgo 109 (1. Viertel 13. Jh.), vgl. <https://jb-c.bj.uj.edu.pl/dlibra/doccontent?id=159362>.
13. Entspricht den mit OCR4all erstellten Rohdaten für jede Seite, bestehend aus dem Scan und der zugehörigen XML-Datei, die umfassende Informationen über die Seite enthalten kann, mindestens aber die Koordinaten und die korrekte Transkription einer jeden Zeile. Durch die Verwendung des etablierten Standard Formats PAGE wird eine problemlose und umfangreiche Nachnutzung durch eine Vielzahl von OCR/HTR Programmen sichergestellt.
14. Vgl. Stolz 2006.
15. Vgl. das Editionsprojekt Lyrik des deutschen Mittelalters. Digitale Edition. URL: <http://www.ldm-digital.de/>.
16. Vgl. das digitale Parzival-Projekt der Universität Bern. URL: <http://www.parzival.unibe.ch/>.
17. Vgl. das Projekt Welscher Gast digital. URL: <https://digi.u-b.uni-heidelberg.de/wgd/>.
18. Vgl. Krautter 2018.
19. Vgl. grundlegend Friedrich 2006.
20. Vgl. exemplarisch zur KJ Fromm 1971.
21. Vgl. Schiewer 2005.

Bibliographie

Bojanowski, Piotr / Grave, Edouard / Joulin, Armand / Mikolov, Tomas (2017): "Enriching word vectors with subword information", in: *TACL* 5: 135–146.

Dimpel, Friedrich Michael (2017): "Ein Delta-Rätsel. Nicht-normalisierte mittelhochdeutsche Texte, Z-Wert-Begrenzung und ein Normalisierungswörterbuch. Oder: Auf welche Wörter kommt es bei Delta an", in: *DARIAH-DE Working Papers* 25. URL: <https://cris.fau.de/converis/portal/publication/120046124>

Friedrich, Jesko (2006): *Phraseologisches Wörterbuch des Mittelhochdeutschen. Redensarten, Sprichwörter und andere feste Wortverbindungen in Texten von 1050-1350*. Tübingen: Niemeyer 2006.

Fromm, Hans (1971): "Stemma und Schreibnorm. Bemerkungen anlässlich der "Kindheit Jesu" des Konrad von Fußesbrunnen", in: Hennig, Ursula / Kolb, Herbert (eds.): *Mediaevalia litteraria. FS für Helmut de Boor zum 80. Geburtstag*. München: C.H. Beck 193–210.

Huang, Gao / Guo, Chuan / Kusner, Matt / Sun, Yu / Sha, Fei / Weinberger, Kilian

(2016): “Supervised Word Mover's Distance”, in: *NIPS* 29. URL: <https://proceedings.neurips.cc/paper/2016/hash/10c66082c124f8afe3df4886f5e516e0-Abstract.html>

Kahle, Philip / Colutto, Sebastian / Hackl, Günter / Mühlberger, Günter (2017): “Transkribus—a service platform for transcription, recognition and retrieval of historical documents”, in: *IAPR* 4: 19-24.

Kragl, Florian (2015): “Normalmittelhochdeutsch. Theorieentwurf einer gelebten Praxis”, in: *ZfdA* 144: 1-27.

Krautter, Benjamin (2018): “Über die Attribution hinaus. Forschungsperspektiven der Stilometrie als Anwendungsfeld in der Literaturwissenschaft”, in: Bernhart, Toni / Willand, Marcus / Richter, Sandra / Albrecht, Andrea (eds.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin/Boston: De Gruyter 289-314.

Kusner, Matt J. / Sun, Yu / Kolkun, Nicholas I. / Weinberger, Kilian Q. (2015): “From Word Embedding To Document Distances”, in: *ICML* 37: 957-966. URL: <https://dl.acm.org/doi/10.5555/3045118.3045221>

Mikolov, Tomas / Grave, Edouard / Bojanowski, Piotr / Puhrsch, Christian / Joulin, Armand (2018): “Advances in Pre-Training Distributed Word Representations”, in: *LREC* 2018. <https://aclanthology.org/L18-1008/>

Neudecker, Clemens / Baierer, Konstantin / Federbusch, Maria / Boenig, Matthias / Würzner, Kay-Michael / Hartmann, Volker / Herrmann, Elisa (2019): “OCR-D: An end-to-end open source OCR framework for historical printed documents”, in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*: 53-58.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): “OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings”, in: *Applied Sciences* 9,22. URL: <https://www.mdpi.com/2076-3417/9/22/4853>

Schiewer, Jans-Jochen (2005): “Fassung, Bearbeitung, Version und Edition”, in: Schubert, Martin J. (ed.): *Deutsche Texte des Mittelalters zwischen Handschriftennähe und Rekonstruktion. Berliner Fachtagung 1.-3. April 2004*. Tübingen: Niemeyer 35-50.

Stolz, Michael (2006): “Vernetzte Varianz. Mittelalterliche Schriftlichkeit im digitalen Medium”, in: Giuriato, Davide / Stingerlin, Martin / Zanetti, Sandro (eds.): *System ohne General. Schreibszenen im digitalen Zeitalter*. München: Wilhelm Fink Verlag: 217-244.

Internetadressen

<http://www.narragonien-digital.de>

<http://ocr4all.de>

<http://www.mhdbdb.sbg.ac.at/>

https://digi.ub.uni-heidelberg.de/diglit/bsb_cgm571

<https://jbc.bj.uj.edu.pl/dlibra/doccontent?id=159362>

<http://www.ldm-digital.de/>

<http://www.parzival.unibe.ch/>

<https://digi.ub.uni-heidelberg.de/wgd/>

Hemisphären des digitalen Gedächtnisses

Analyse von TEI-kodierten Bibelreferenzen mit XQuery im Rahmen der »Bibliothek der Neologie«

Stallmann, Marco

marco.stallmann@uni-muenster.de

Westfälische Wilhelms-Universität Münster, Germany

Sikora, Uwe

sikora@sub.uni-goettingen.de

Niedersächsische Staats- und Universitätsbibliothek Göttingen

Kreß, Hannah

hkress@uni-muenster.de

Westfälische Wilhelms-Universität Münster, Germany

Lemitz, Bastian

lemitz@uni-muenster.de

Westfälische Wilhelms-Universität Münster, Germany

Pietsch, Andreas

a.n.pietsch@uni-muenster.de

Westfälische Wilhelms-Universität Münster, Germany

Wünsch, Lukas

lukas.wuensch@uni-muenster.de

Westfälische Wilhelms-Universität Münster, Germany

Kontext

Editionen leisten einen wichtigen Beitrag zur geistes- und kulturwissenschaftlichen „Erinnerungsarbeit“ (vgl. Albrecht et al. 2006) und sind als erschließende Wiedergabe historischer Dokumente unter einem digitalen Paradigma zunehmend auch für die „Kulturen des digitalen Gedächtnisses“ von entscheidender Bedeutung. Doch obwohl digitale Ansätze in den historisch arbeitenden Fächern mittlerweile durchaus verbreitet sind, sprechen neuere Bestandsaufnahmen von längst noch nicht ausgeschöpften Potenzialen: Zum einen seien digitale Editionen oft weiterhin am klassischen Medium Buch orientiert und bestrebt, traditionelle Ansätze zu übertragen ohne neue Fragerichtungen zu entwickeln, sodass nicht selten eher von „digitalisierten Editionen“ gesprochen und die unzureichende Kommunikationsstruktur zwischen Informations- und Kulturwissenschaften kritisiert wird (vgl. Sahle 2017a). Zum anderen fehlen bisher etablierte Standards für die fachspezifische Identifikation und Klassifikation von Datenbeständen, auf deren Basis ein projektübergreifender Datenpool ent-

stehen könnte. Eine solche gemeinsame Semantik, so wird vermutet, lasse sich nur durch tatsächliche Nutzung, Überprüfung und Verfeinerung der Daten in konkreten Projekten etablieren, die einen klaren *open data*-Ansatz verfolgen und damit einen automatisierten Datenaustausch allererst ermöglichen (vgl. Zahnd 2020).

Die *Bibliothek der Neologie* (BdN) ist ein von der *Deutschen Forschungsgemeinschaft* (DFG) gefördertes Langzeitprojekt, welches das Ziel verfolgt, ausgewählte Texte der Neologie als zentraler Richtung der deutschen Theologie des 18. Jahrhunderts in kritischer Hybridedition für die interdisziplinäre Forschung bereitzustellen. Es handelt sich um eine Kooperation des Seminars für Kirchengeschichte II (Leitung: Prof. Dr. Albrecht Beutel) an der Evangelisch-Theologischen Fakultät der *Westfälischen Wilhelms-Universität Münster* sowie der Abteilung Forschung und Entwicklung der *Niedersächsischen Staats- und Universitätsbibliothek Göttingen* (Leitung: Dr. Jan Brase).¹ Der Beitrag befasst sich mit der Erschließung theologischer Editionsdaten im Rahmen des Projekts und zeigt exemplarisch auf, wie mithilfe von X-Technologien neue Blickwinkel auf kulturwissenschaftliche und speziell theologiehistorische Fragen eröffnet werden können.

Analyse

Das 18. Jahrhundert markiert in der Geschichte der Bibelwissenschaften eine folgenreiche Umbruchphase: Indem sich die neologische Schriftauslegung von kirchlich-dogmatischen Voraussetzungen emanzipierte und den Durchbruch zur historisch-kritischen Exegese vollzog, schuf sie die Grundlagen für eine aufgeklärte, neuzeitfähige Religionswissenschaft. In den zahlreichen Bibelreferenzen der im Rahmen der *Bibliothek der Neologie* edierten Quellschriften manifestiert sich diese Entwicklung in anschaulicher Weise. Die Referenzen werden bei der Erstellung der digitalen Edition mithilfe des Datenmodells der *Text Encoding Initiative* (TEI) vollständig erschlossen und in ein maschinenlesbares Format (mittels Elementen `tei:bibl` und `tei:citedRange` und entsprechenden Attributwerten, z.B. `from="Joh:1:1"` und `to="Joh:1:18"`) überführt.² Die Printedition transformiert diese Kodierung nicht zuletzt in ein ausführliches Register der einzelnen, in den Quellschriften angeführten Bibelverse. Damit ist allerdings das Potenzial der Auszeichnung von teilweise mehreren tausend Bibelstellen bei weitem noch nicht ausgeschöpft. Der hier verfolgte Ansatz schafft erweiterte Analysemöglichkeiten, indem die Bibelstellenauszeichnung systematisch mithilfe verschiedener XQuery-Module ausgewertet wird. Für die Verwendung von XQuery (etwa anstelle der Konversionssprache XSLT) spricht dabei neben der Zugänglichkeit (vgl. Anderson / Wicentowski 2020) auch die interne Projektinfrastruktur, auf die bei der Entwicklung zurückgegriffen wurde.

Im Projekt werden die TEI-kodierten Editionsdaten zunächst über eine Typeswitch-Transformation in ein einheitliches Zwischenformat konvertiert, um als solches in verschiedenen Anwendungen (s. unten 2.1 bis 2.3) weiterverarbeitet werden zu können. Über ein GitHub Repository arbeiten Fach- und InformationswissenschaftlerInnen kollaborativ an diesen Anwendungen.³ Die digitale Edition der *Anleitung zum Studium der populären Dogmatik* (1779, 1789) aus der Feder des Jenaer Theologen Johann Jakob Griesbach (1745–1812) verzeichnet etwa 2.500 Bibelreferenzen (vgl. Griesbach 2019) und dient im Folgenden als Beispiel.

Verweishäufigkeiten und Verwendungskontexte von biblischen Sinneinheiten

Bei der texthermeneutischen Analyse von Quellschriften mit einer großen Zahl von Bibelstellenverweisen sind Theologiehistorikerinnen und -historiker mit schwierigen Selektions- und Priorisierungsfragen konfrontiert: Beispielsweise ist nicht unmittelbar ersichtlich, welche Bibelstellen für die Argumentation des Autors wichtig und daher vertiefend zu betrachten sind. Das Bibelstellenregister der Printausgabe bietet hier einen wichtigen Überblick, geht aber über die Kapitel- und Versstrukturen einschlägiger Bibelausgaben nicht hinaus und lässt die dynamischen Gliederungs- und Sortierungsmöglichkeiten der XML/TEI-Daten weitgehend unberücksichtigt.

Im Rahmen einer weiterführenden Analyse von Bibelreferenzen werden daher in einer externen Bibel-XML-Repräsentation bestimmte exegetisch konsensfähige Sinneinheiten (z.B. „Johannesprolog“ = Joh 1,1–18) festgelegt.⁴ Über verschiedene XQuery-Funktionen werden diese Sinneinheiten in einem zweiten Schritt auf Entsprechungshäufigkeiten im Editionsdocument überprüft (Treffer wären im genannten Beispiel etwa `@n="Joh:1:1"`, `@from="Joh:1:1" @to="Joh:1:3"` oder `@n="Joh:1:17"`). Wendet man diese Abfrage auf sämtliche, im Rahmen der alt- und neutestamentlichen Exegese eingeteilten Sinneinheiten an, so entsteht eine Häufigkeitenstatistik, die es erlaubt, bestimmte biblische Schlüsseltexte (man könnte auch sagen: quantitative „Lieblingstexte“ der Autoren) zu identifizieren, deren Funktion für die Quellschrift nun textanalytisch näher untersucht werden kann.

Sinneinheiten	BdN III: Griesbach	BdN VIII: Steinbart	Gesamt
Röm 2,1–16: Gericht Gottes	47	10	57
Röm 4,1–25: Schriftbeweis (Abrahams Verheißung)	51	4	55
Joh 1,1–18: Prolog (Logosymnus)	32	12	44
1Kor 15,1–58: Die Auferweckung von den Toten	35	9	44
1Joh 3,1–24: Die Herrlichkeit der Gotteskindschaft	30	14	44
...

Abb. 1: Häufigkeiten zentraler Sinneinheiten bei verschiedenen Autoren (Ausschnitt)

Damit bewegt sich die Anwendung im Bereich derjenigen quantitativen Analysemethoden, die in qualitative Anschlussuntersuchungen übergehen (vgl. Jannidis 2010) und daher auch die kritisch-hermeneutische Interpretation keineswegs ausblenden, sondern allererst anregen (vgl. Wettlaufer 2016). So ist in methodischer Hinsicht der Kurzschluss zu vermeiden, den statistisch ermittelten Schlüsseltexten unmittelbar auch eine qualitative Schlüsselrolle zuzuschreiben – vielmehr ermöglichen sie aufgrund ihrer sichergestellten Häufigkeit und Anwendungsbreite im Rahmen des analysierten Textes neue, sinnvolle Erschließungszugänge, die sich bei der Beschränkung auf Einzelverse nicht zwangsläufig ergeben würden: Im Falle der theologiegeschichtlichen Untersuchung und Einordnung etwa der o.g. *Anleitung* von Johann Jakob Griesbach lässt sich beispielsweise das argumentative Begründungsgewicht des Johannesevangeliums und insbesondere des tiefgründigen, für die Logoschristologie zentralen Prologs belastbar einholen. Auf der Grundlage dieser Erkenntnisse könnten zudem die auslegungsgeschichtlichen und textkritischen Voraussetzungen des theologischen Systems näher un-

tersucht werden: Sie sind für die Gesamtheit mehrerer tausend Belegstellen (sog. *dicta probantia*) kaum angemessen und umfassend zu berücksichtigen – sehr wohl aber auf der Basis einer methodisch kontrollierten Identifikation argumentationsrelevanter Bibeltexte, auf die der hier dargestellte Ansatz zielt.

Neben den Häufigkeiten besteht eine sinnvolle Modifikation in der Abfrage, in wievielen verschiedenen Kapiteln (`tei:div` mit `@type = "section"` oder `"chapter"`) oder Paragraphen (`tei:p`) bestimmte Sinneinheiten vorkommen und in welchen. Auf diese Weise lassen sich weiterführende Forschungsfragen formulieren: Wo werden die Schlüsseltexte in ihrem Wortlaut, ihrer Intention, ihrer Sprache oder ihrer Entstehungssituation aufgegriffen und welche theologische Funktion haben sie an den jeweiligen Stellen? Darüber hinaus lässt sich für das Gesamtkorpus der im Rahmen der *Bibliothek der Neologie* edierten Quellschriften eine Häufigkeitsmatrix (vgl. Schöch 2017) der am meisten referenzierten biblischen Sinneinheiten erstellen, mit der sich textübergreifende Erkenntnisse gewinnen lassen. Auf diese Weise rückt der Ansatz in die Nähe bestehender Umgangsformen mit größeren digitalen Beständen, die bereits unter Schlagworten wie „distant reading“ (vgl. Moretti 2007) oder „cultural analytics“ (vgl. Piper 2016) diskutiert werden.

Bibelstellendichte und relative Häufigkeiten

Für einige theologiegeschichtliche Fragestellungen sind auch relative Verweishäufigkeiten im Sinne einer Bibelstellendichte und ihre Entwicklung im Argumentationsverlauf der Quellschrift von Interesse: In welchem Kapitel oder Abschnitt bedient sich der Autor einer vergleichsweise hohen Anzahl von Belegstellen – in welchem bleiben sie möglicherweise nahezu aus? Während sich eine solche Bibelstellendichte anhand der Printedition nur sehr mühsam und ungenau ermitteln lässt, ermöglicht die XML-Auszeichnung eine präzise Abfrage der Bibelstellenanzahl in Abhängigkeit von den entsprechenden Druckseiten-, Abschnitts- oder Wortanzahlen. Innerhalb der kirchenhistorischen Forschung gibt es bereits analoge Ansätze in dieser Richtung (vgl. etwa Beutel 2007), die im Rahmen von digitalen Editionen weiterzuentwickeln sind. Stellt man – beispielsweise über die Open-Source-Datenvisualisierung *Chart.js* – die Referenzhäufigkeiten im Verlauf der Textstruktur dar, so ergibt sich das kirchen- und theologiegeschichtlich aufschlussreiche Bild eines „Bibelstellenspannungsbogens“.

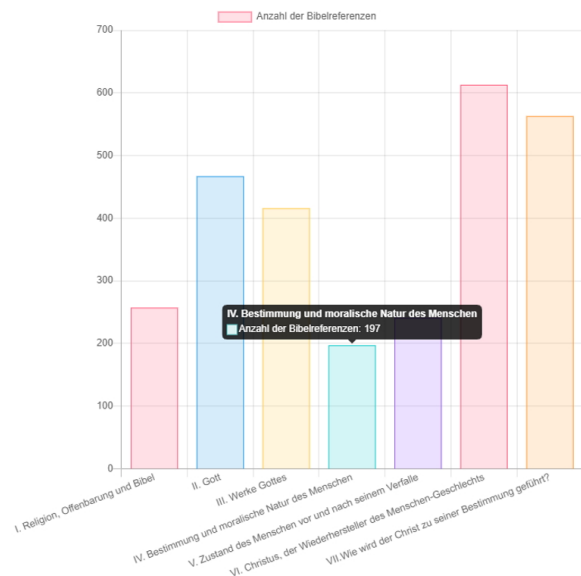


Abb. 2: Häufigkeiten von Bibelreferenzen im Verlauf der Hauptkapitel bei Griesbach

Obwohl die Frequenz der *dicta probantia* in der *Anleitung* für ihre Entstehungszeit angesichts der mittlerweile etablierten Bibelkritik noch relativ hoch ist, lassen sich hier signifikante Tiefpunkte ausfindig machen, d.h. bestimmte Paragraphen, in denen Griesbach zu erfahrungsbezogener, philosophischer oder religionspädagogischer Argumentation übergeht: Insbesondere seine Abschnitte zur Glückseligkeit, zur Trinität und zur moralischen Natur und Freiheit des Menschen kommen über weite Strecken ohne biblische Begründung aus. Sofern die Ergebnisse nicht für sich stehengelassen, sondern kritisch-hermeneutisch interpretiert werden, lässt sich beispielsweise die nicht ganz leichte Einordnung der Theologie Griesbachs zwischen supranaturalistischer Bibeldogmatik und anthropologisch fundierter Populartheologie erheblich präzisieren.

Ein charakteristischer Vorzug der digitalen Abfrage besteht darin, dass die Analyse nicht bei Einzelautoren stehen bleiben muss, sondern gezielte Vergleichsmöglichkeiten bietet: So lässt sich die Hypothese formulieren und überprüfen, dass etwa das mit der *Anleitung* verglichene *System der reinen Philosophie* (1778, ⁴1794; erscheint als BdN VIII) des progressiveren Neologen Gottlieb Samuel Steinbart (1738–1809) einen ganz anderen Bibelstellenspannungsbogen aufweist. Aber auch das im Editionsprozess befindliche *Wörterbuch des Neuen Testaments* (1772, ⁶1805; erscheint als BdN IX) von Wilhelm Abraham Teller (1734–1804) lässt sich aufgrund seines deutlich größeren Umfangs und der biblisch-theologischen Ausrichtung mit Gewinn auf die Bibelstellendichte einzelner Lemmata befragen, die sich im Laufe der mehr als dreißigjährigen Textentwicklung erheblich verändert haben dürfte.⁵

Bibelstellen und Textvarianz

Ein Alleinstellungsmerkmal der *Bibliothek der Neologie* als digitaler Edition ist schließlich der umfangreiche Variantenapparat. So wird im Projekt danach gefragt, wie die Bibelstellenauszeichnung mit dem ebenfalls in TEI kodierten *Critical Apparatus* korrespondiert, der in diesem Fall aufgrund der internen Varianz

der ausgewählten Quellenschriften eine hohe Komplexität aufweist.⁶ Dabei sind orthographische Textvarianten und Korrekturen von echten Ersetzungen, Hinzufügungen oder Löschungen zu unterscheiden, um festzustellen, wie sich die Bibelargummentation im Laufe der edierten Auflagen entwickelt und ob dabei gegebenenfalls textkritische oder exegetische Weichenstellungen im Hintergrund stehen. Das entsprechende XQuery-Modul und die Darstellung seines Outputs in einer dynamischen Gesamtübersicht beschleunigt deutlich den Erkenntnisprozess im Vergleich zur Komplettsichtung des teilweise sehr umfangreichen kritischen Apparats und erlaubt Rückschlüsse auf bestimmte Modifikationsentscheidungen, die der Autor im Verlauf der Textgenese vorgenommen hat.

Bibelref.	Abschnitt	1. Aufl.	2. Aufl.	3. Aufl.	4. Aufl.
Joh 1,14	§ 135	✓	✓	✓	✗
Joh 1,17	§ 132	✗	✗	✗	✓
Joh 3,16	§ 13	✓	✓	✓	✗
...

Abb. 3: Modifikationen von Bibelreferenzen bei Griesbach (ausschnittshaftre Nachbildung des mittels XQuery erstellten „textkritischen Bibelstellenregisters“)

So wird beispielsweise in der vierten Auflage von Griesbachs *Anleitung* in § 135 der mit Joh 1,14 belegte Zusatz, dass Jesus es zu Lebzeiten an Merkmalen seiner göttlichen Majestät nicht habe fehlen lassen, entfernt und damit das theologische Verständnis von dessen Menschlichkeit als ethischer Lehrer der Glückseligkeit an dieser Stelle geringfügig modifiziert (vgl. Griesbach 2019, 139). In diesen und zahlreichen weiteren mikroskopischen Details äußern sich theologiegeschichtliche Zentralfragestellungen, die in den Möglichkeitsraum der Digital Humanities sinnvoll einzuholen sind (vgl. Anderson 2019).

Ausblick

Über die herkömmliche Registeranalyse im Rahmen kritischer Editionen hinausgehend lassen sich mindestens drei Vorteile einer digitalen Analyse von TEI-kodierten Bibelreferenzen plausibilisieren: So ermöglicht die eindeutige, maschinenlesbare Bibelstellenauszeichnung eine umfassende Erschließung auf Datenebene und damit die Identifikation spezifisch kirchenhistorischer Daten, die zu den explizit markierten Desideraten der Digital Humanities gehört und die sich nur in konkreten Projekten mit einem klaren *open data*-Ansatz realisieren lässt (vgl. Zahnd 2020). Zweitens bietet eine entsprechende Portaldarstellung im Rahmen der digitalen Edition erkennbar übersichtlichere Präsentations- und präzisere Auswertungsmöglichkeiten, die im Vergleich zu herkömmlichen Publikationsformaten mit differenzierten Analyseergebnissen einhergehen. Der stärkste Vorzug eines digitalen Abfrageansatzes besteht jedoch drittens darin, dass er nicht auf die Einzeldition beschränkt ist, sondern darüber hinaus auf weitere historische Quellen mit entsprechender Bibelstellenerschließung angewendet werden kann (Nachnutzbarkeit). Geht man davon aus, dass die Anzahl digital erschlossener Quellentexte auch im Bereich der Kirchen- und Theologiegeschichte weiter steigt, dürfte sich langfristig auch eine erhöhte Nachfrage hinsichtlich entsprechender Analyseverfahren ergeben, mit denen ein weiterführender Beitrag zum vertieften Verständnis der neuzeitlichen Schriftaus-

legung und ihrer kulturgeschichtlichen Bedeutung geleistet werden kann.

Unabhängig von seinem fachwissenschaftlichen Output soll der vorgestellte Ansatz aber vor allem dem interdisziplinären Diskurs zur engeren Vernetzung und Verzahnung von „D“ und „H“ dienen und neue Forschungsfragen ermöglichen, die gerade durch die veränderten Informationsbedingungen und Publikationsmöglichkeiten provoziert werden. Denn obwohl die Digital Humanities mittlerweile ein hohes Maß an Professionalisierung erreicht haben, dokumentieren jüngere Bestandsaufnahmen noch immer eine unzureichende Kommunikationsstruktur zwischen Informationstechnologie und Geisteswissenschaften (vgl. Sahle 2017b). Die damit verbundenen, fachkulturellen und wissenschaftspolitischen Problemstellungen sind dabei ebenso in den Blick zu nehmen wie die spezifischen Potenziale langzeitgeförderter (Hybrid-)Editionen: Solange das „Digitale“ auf ein anempfohlenes Hilfsmittel oder die „Edition“ auf ein notwendiges Trägermedium reduziert wird, dürften sich weiterführende Fragestellungen in Grenzen halten. Wo aber Digital Humanities auf der Schnittstelle zwischen Geistes- und Informationswissenschaft gestärkt werden und etwas Eigenes hervorbringen sollen, da sind auf der Ebene der forschenden sowie der fördernden Einrichtungen integrative Denkweisen zu unterstützen.

Das in den letzten Jahren zu einem kulturwissenschaftlichen Leitbegriff avancierte Gedächtnisparadigma (vgl. Radonic / Uhl 2016) dürfte sich nur dann als tragfähig erweisen, wenn es dieses Integrationsmoment abzubilden und somit gewissermaßen die „Hemisphären“ des digitalen Gedächtnisses zu synchronisieren imstande ist. Kulturen zeichnen sich schließlich – neben ihren Techniken der Überlieferung, der Speicherung und des Erinnerns (vgl. das CfP der Konferenz) – auch durch ihre handlungsleitenden Sinnstrukturen aus (vgl. Assmann 2005). In diesem Zusammenhang kann die Theologie, wie gezeigt wurde, einen Beitrag leisten: Es wäre zu prüfen, ob nicht mit der Analyse TEI-kodierter Bibelreferenzen im Rahmen digitaler Editionen ein bisher vernachlässigtes, theologiehistoriographisches Bewährungsfeld der Digital Humanities vorliegt, das künftig mehr Aufmerksamkeit bekommen sollte.

Fußnoten

1. Die digitale Publikation erfolgt auf einem Onlineportal (<https://bdn-edition.de>) sowie im Langzeitarchiv *TextGrid Repository* (<https://textgridrep.org> [letzter Zugriff 25. November 2021]) unter der freien Lizenz CC BY-SA 3.0, während die aus dem gleichen Datenbestand generierte Printausgabe (vgl. Nösel 2019; Griesbach 2019; Bahrdt/Semler 2020) vom Tübinger Wissenschaftsverlag *Mohr Siebeck* begleitet wird.
2. Vgl. TEI P5: Guidelines for Electronic Text Encoding and Interchange by the TEI Consortium. Originally edited by C. M. Sperberg-McQueen and L. Burnard [...] Version 4.3.0. Last updated on 31st August 2021 (<https://guidelines.tei-c.de> [letzter Zugriff 25. November 2021]), „3.12.2.5 Scopes and Ranges in Bibliographic Citations“ sowie die im Projektportal (s. Anm. 1) dokumentierte Schemaanpassung.
3. Vgl. <https://github.com/Bibliothek-der-Neologie/bdnBible> sowie zur o.g. Print-Serialisierung: <https://gitlab.gwdg.de/bibliothek-der-neologie/print> [letzter Zugriff 25. November 2021].
4. Die externe XML-Repräsentation der Bibel dient als Referenz für die Analyse der in den Editionsdaten vorkommenden Bibelverweise. Sie ist angelehnt an die *Zefania XML Bible Markup Language* (<https://sourceforge.net/projects/zefania-sharp> [letzter Zugriff 25. November 2021]), wird allerdings projektspezi-

fisch angepasst. – Bei den Sinneinheiten handelt es sich weniger um nutzerabhängige Interpretationserzeugnisse als vielmehr um naheliegende Sinnabschnitte, die aus den Handlungsverläufen und Gliederungen der biblischen Texte recht eindeutig hervorgehen und auf diese Weise traditionell geworden sind. Dass das Modul selbstverständlich auch andere Sinneinheiten ermöglicht, die dann zu anderen Zwischenergebnissen führen, wird im Vortrag diskutiert. – Während die externe Bibel-XML vor allem in dem unter 2.1 erläuterten Modul („Sinneinheiten“) Verwendung findet, dient das zu Beginn des Abschnitts genannte Zwischenformat der Vereinheitlichung und Vereinfachung des unübersichtlichen Editionsdatenbestands für die weitere Auswertung und liegt als Vorstufe allen drei Modulen (2.1 bis 2.3) zugrunde. 5. Die in den Abschnitten 2.1 und 2.2 vorgestellten Module zielen insofern auf den Übergang von quantitativen Statistiken in kritisch-hermeneutische Textuntersuchungen, als damit in unübersichtlichen Editionsdaten zentrale biblische Sinneinheiten oder Abschnitte mit auffälliger Bibelstellendichte identifiziert werden können, die eine weitere inhaltliche Kontextualisierung und Analyse anregen. Dabei steht weniger ein Kategoriensystem-orientiertes Vorgehen im Mittelpunkt, wie es Verfahren der qualitativen Inhaltsanalyse oder Mixed Methods (vgl. Kuckartz 2014) nahelegen, als vielmehr eine datenbasierte Schaffung von Textzugängen, die durch traditionelle Verfahren so nicht erreicht würden. Für mögliche Weiterentwicklungen mit Blick auf qualitative Forschungsmethoden ist der Ansatz aber durchaus offen. 6. Vgl. das Kapitel „12 Critical Apparatus“ der TEI Guidelines (s. Anm. 2). Gemäß der projektspezifischen Schemaanpassung erfolgt die Unterscheidung der kritischen Apparate im tei:rdg über @wit mit Werten „#a“, „#b“, „#c“ usw. (Textzeugen) sowie @type mit Werten „v“ (variant), „pp“ (paraphrase), „pt“ (parenthesis) und „om“ (omission). Mit dieser Klassifikation lassen sich die textkritischen Phänomene der neologischen Schlüsseltexte im Rahmen der Hybridausgabe präzise beschreiben, aber auch die entsprechenden Interdependenzen zwischen den zahlreichen Bibelverweisen und der Textgenese automatisiert abfragen.

Bibliographie

- Albrecht, Christian / Chapman, Mark D. / Kaufmann, Thomas / Marksches, Christoph / Molendijk, Arie L. / Von Soosten, Joachim** (2006): "Erinnerungsarbeit durch Klassikeredition. Die Bedeutung akademischer Selbsthistorisierung für die Zukunft des Protestantismus" in: Graf, Friedrich Wilhelm (ed.): *Geschichte durch Geschichte überwinden. Ernst Troeltsch in Berlin* (Troeltsch-Studien. Neue Folge 1). Gütersloh: Gütersloher Verlagshaus 253–284.
- Anderson, Clifford** (2019): "Digital Humanities and the Future of Theology" in: *Cursor. Zeitschrift für explorative Theologie* 1. <https://doi.org/10.17885/heup.czeth.2019.1.24000> [letzter Zugriff 25. November 2021].
- Anderson, Clifford / Wicentowski, Joseph** (2020): *XQuery for Humanists* (Coding for Humanists). College Station: Texas A&M University Press.
- Assmann, Jan** (2005): "Der Begriff des kulturellen Gedächtnisses" in: Dreier, Thomas (ed.): *Kulturelles Gedächtnis im 21. Jahrhundert*. Tagungsband des internationalen Symposiums vom 23. April 2005 (Schriften des Zentrums für Angewandte Rechtswissenschaft 1). Karlsruhe: Universitätsbibliothek 21–29.
- Bahrdt, Carl Friedrich / Semler, Johann Salomo** (2020): *Glaubensbekenntnisse (1779–1792)*, eds. Pietsch, Andreas / Weidemann, Christian (BdN I). Tübingen: Mohr Siebeck.
- Beutel, Albrecht** (2007): "Biblischer Text und theologische Theoriebildung in Luthers Schrift ‚Von weltlicher Oberkeit, wie weit man ihr Gehorsam schuldig sei‘ (1523)" in: Beutel, Albrecht: *Reflektierte Religion*. Beiträge zur Geschichte des Protestantismus. Tübingen: Mohr Siebeck 21–46.
- Griesbach, Johann Jakob** (2019): *Anleitung zum Studium der populären Dogmatik* (1779, 4. Aufl. 1789), ed. Stallmann, Marco (BdN III). Tübingen: Mohr Siebeck.
- Jannidis, Fotis** (2010): "Methoden der computergestützten Textanalyse" in: Ansgar Nünning / Vera Nünning (eds.): *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*. Ansätze - Grundlagen - Modellanalysen. Stuttgart/Weimar: Metzler 109–132.
- Kuckartz, Udo** (2014): *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: Springer VS.
- Moretti, Franco** (2007): *Graphs, Maps, Trees. Abstract Models for Literary History*. <https://hdl.handle.net/2027/heb.08911> [letzter Zugriff 25. November 2021].
- Nösselt, Johann August** (2019): *Anweisung zur Bildung angehender Theologen* (1786/89, 3. Aufl. 1819/19), eds. Beutel, Albrecht / Lemitz, Bastian / Söntgerath, Olga (BdN VI). Tübingen: Mohr Siebeck.
- Piper, Andrew** (2016): *There will be numbers*. Cultural Analytics 1. <https://doi.org/10.22148/16.006> [letzter Zugriff 25. November 2021].
- Radonic, Ljiljana / Uhl, Heidemarie** (2016): *Gedächtnis im 21. Jahrhundert*. Zur Neuverhandlung eines kulturwissenschaftlichen Leitbegriffs. Bielefeld: transcript-Verlag.
- Sahle, Patrick** (2017a): "Digital Humanities und die Fächer. Eine schwierige Beziehung?" in: *Forum Exegese und Hochschuldidaktik*. Verstehen von Anfang an 2: 7–27.
- Sahle, Patrick** (2017b): "Digitale Editionen" in: Fotis Jannidis / Hubertus Kohle / Malte Rehbein (eds.): *Digital Humanities*. Eine Einführung. Stuttgart: J.B. Metzler 234–249.
- Schöch, Christof** (2017): "Quantitative Analyse", in: Jannidis / Kohle / Rehbein: *Digital Humanities* 279–298.
- Wettlaufer, Jörg** (2016): "Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern" in: *Zeitschrift für digitale Geisteswissenschaften* https://doi.org/10.17175/2016_011 [letzter Zugriff 25. November 2021].
- Zahnd, Ueli** (2020): "Netzwerke, historisch und digital. Digital Humanities und die Mittlere und Neue Kirchengeschichte" in: *Verkündigung und Forschung* 65: 114–123.

iART

Eine Suchmaschine zur Unterstützung von bildorientierten Forschungsprozessen

Schneider, Stefanie

stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Germany

Springstein, Matthias

matthias.springstein@tib.eu
Technische Informationsbibliothek (TIB), Germany

Rahnama, Javad

javad.rahnama@uni-paderborn.de
Universität Paderborn, Germany

Kohle, Hubertus

hubertus.kohle@lmu.de
Ludwig-Maximilians-Universität München, Germany

Ewerth, Ralph

ralph.ewerth@tib.eu
Technische Informationsbibliothek (TIB), Germany;
Forschungszentrum L3S, Leibniz Universität Hannover,
Germany

Hüllermeier, Eyke

eyke@ifi.lmu.de
Ludwig-Maximilians-Universität München, Germany

Kunsthistorische Erkenntnisprozesse werden über Ähnlichkeitsbeschreibungen angetrieben: Bei Heinrich Wölfflin finden diese in formalanalytischer Perspektive statt, bei dem gleichzeitig tätigen Aby Warburg aus der Sicht des kulturwissenschaftlich forschenden Ikonologen. Beide gelten als Väter der modernen Kunstgeschichte, die deren bis heute aktuellen methodischen Zugriffe im Bereich von Form und Semantik grundgelegt haben. Wölfflin (1915) visualisiert Ähnlichkeiten und Differenzen in seinen „Kunstgeschichtlichen Grundbegriffen“ auf einer Doppelseite mit gegenüberliegenden Vergleichsbeispielen, kategorisiert in fünf binären Gegensätzen. Warburg gestaltet seine zu Berühmtheit avancierten Bilderatlastafeln in Gruppen mit loserer, ähnlichkeitsnaher Verbindung, deren Qualität nicht immer leicht zu erkennen ist (Warnke und Brink 2000; Ohrt et al. 2020). Ein digitales Werkzeug, das ein umfangreiches, nicht über Kanonisierungsprozesse vorselektiertes Bildmaterial nach unterschiedlich gewichteten Ähnlichkeitskriterien zu filtern in der Lage ist, scheint daher in hohem Maße erwünscht – gerade, weil elektronische Bilddatenbanken inzwischen über große Mengen von Reproduktionen verfügen.

Bisherigen Ansätzen fehlt es jedoch entweder an einer Feinabstimmung auf die kunsthistorische Domäne (Rossetto et al. 2016), an flexiblen Suchabfragestrukturen, die sich den Bedürfnissen der Nutzer:innen anpassen (Lang und Ommer 2018), oder an der Möglichkeit, eigene Datensätze hochzuladen und zu verwalten (Offert, Bell und Harlamov 2021). Mit iART¹ wird der Versuch unternommen, dieses Desiderat mithilfe einer offenen Web-Plattform zu schließen. Das Retrieval von Objekten erfolgt, wie im Weiteren gezeigt wird, nicht nur mit durch Deep Learning generierte Schlagwörter, sondern auch unter Verwendung multimodaler Embeddings, die eine Suche bspw. auf Grundlage detaillierter Szenenbeschreibungen ermöglichen. Eine intuitive Benutzeroberfläche unterstützt die Nutzer:innen bei der Definition von Abfragen und der Untersuchung der Ergebnisse.

Infrastruktur

Das DFG-geförderte Projekt wurde von 2019 bis 2021 umgesetzt vom Lehrstuhl für Mittlere und Neuere Kunstgeschichte der Ludwig-Maximilians-Universität München, der Forschungsgruppe „Visual Analytics“ der TIB Hannover und der Fachgruppe

„Intelligente Systeme und Maschinelles Lernen“ des Heinz Nixdorf Instituts der Universität Paderborn. Es ist geplant, dass die TIB Hannover die entwickelte Plattform auch über die Projektlaufzeit hinaus als Infrastrukturdienst zur Verfügung gestellt und somit dessen Nachhaltigkeit sichert. Weiterhin ist der Quellcode für alle Komponenten frei verfügbar, so dass andere Forscher:innen die Software nachnutzen und erweitern können.²

Die beschriebene Software läuft auf zwei Rechnern: Das erste System ist mit einer Grafikkarte (GeForce GTX 1080 Ti) ausgestattet und dient der Indizierung der Daten und der Ähnlichkeitssuche neu hochgeladener Bilder. Weiterhin läuft auf diesem Rechner eine Elasticsearch-Instanz. Das zweite System stellt die Webseite bereit und speichert zu importierende Bilder.

Backend

Die Aufgabe des Backend besteht sowohl darin, Informationen über Datensätze und Nutzer:innen zu speichern, als auch mithilfe einer API (Application Programming Interface) verschiedene Möglichkeiten des Retrieval zur Verfügung zu stellen. Um die Anpassung an unterschiedliche Forschungsinteressen zu erleichtern, ist die Software modular aufgebaut. Daher sind die einzelnen Indizierungsschritte in Plug-ins ausgelagert und die Benutzerverwaltung von der Suchinfrastruktur getrennt (Abb. 1). Alle Modelle werden mit einem RedisAI-Inferenzserver³ beschleunigt, um die für die Berechnung benötigten Ressourcen optimal zu verwalten. Dieser Schritt erleichtert es, verschiedene Deep-Learning-Modelle auf einer einzigen Grafikkarte laufen zu lassen und ermöglicht den Einsatz von Backend-Systemen wie PyTorch oder TensorFlow. Die Kommunikation zwischen Indexserver und Frontend wird mithilfe eines auf Python basierenden Django-Webservice umgesetzt.⁴ Dieser Service kümmert sich auch um die Verwaltung der Nutzer:innen, von ihnen angelegte Lesezeichen und hochgeladene Bildbestände.

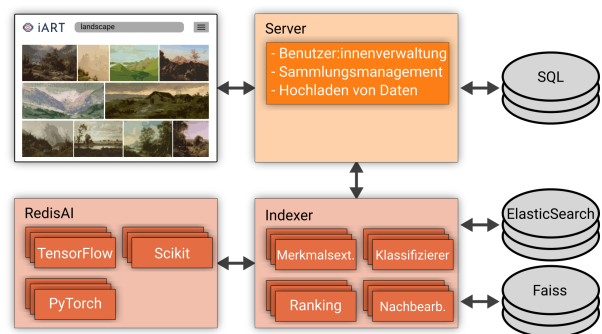


Abb. 1: Architektur mit zugehöriger Datenbankstruktur und RedisAI-Inferenzserver.

Plug-ins

Zumeist extrahieren Suchmaschinen eine einzige Repräsentation eines Bildes, die für alle Suchanfragen verwendet wird. Kunsthistoriker:innen müssen Objekte jedoch unter verschiedensten Gesichtspunkten finden, etwa in Hinblick auf Komposition oder Farbe. Daher generiert iART eine Vielzahl von Merkmalen pro Bild, die je nach Interesse der Forscher:innen gewichtet werden können. Das zugrundeliegende System wurde flexibel entwickelt, sodass es durch Plug-ins leicht erweitert werden kann:

Die einzelnen Plug-ins wurden nach einer Schnittstellendefinition für den jeweiligen Erweiterungstyp in einer eigenen Klasse implementiert, die zur Laufzeit geladen wird. So kann das Framework durch Anlegen einer einzigen Datei und Aktualisieren einer Konfigurationsdatei erweitert werden. Dies gilt insbesondere für die Merkmalsextraktion, die Klassifikation von Bildinhalten, das Ranking der Ergebnisse und verschiedene Nachbearbeitungsschritte, die der Visualisierung und dem Clustering dienen. Die Pipeline ist in Abb. 2 dargestellt. Folgende Plug-ins sind u. a. implementiert:

1. *iMet*. Das Metropolitan Museum of Art stellt im seit 2019 auf Kaggle stattfindenden iMet-Wettbewerb einen Teil seines Bestands für das Training von Bildklassifikatoren zur Verfügung.⁵ Dabei sind die 142.119 Bilder des Datensatzes mit 3.474 Tags, etwa aus den Bereichen Region, Kultur und Medium, annotiert. Ein ResNet-Ansatz (He et al. 2016) wird verwendet, um die Schlagwörter vorherzusagen. Dabei erreicht das Modell einen F_2 -Score von 73,6 Prozent auf einem 20-Prozent-Subsample des Datensatzes. Die Metrik wurde für den iMet-Wettbewerb festgelegt und berechnet sich aus der Anzahl der gefundenen Bilder in Relation zu deren richtiger Klassenzuordnung.
2. *Painter by Numbers*. Der ebenso im Rahmen eines Kaggle-Wettbewerbs erstellte Painter-by-Numbers-Datensatz umfasst 80.000 kunstwissenschaftlich relevante Trainingsbilder aus WikiArt, die u. a. mit Genre- und Stilzuschreibungen versehen sind.⁶ Das Modell für dieses Plug-in besteht aus einer ResNet-Architektur, die 43 Genre- und 142 Stilskategorien vorhersagt und dabei eine Genauigkeit von 63,8 bzw. 40,7 Prozent auf einem 20-Prozent-Subsample des Datensatzes erreicht. Ein weiteres Plug-in mit derselben Architektur wird genutzt, um Genre- und Stilmerkmale für die bildgesteuerte Ähnlichkeitssuche zu extrahieren.
3. *BYOL*. BYOL (Bootstrap Your Own Latent; Grill et al. 2020) ist eine selbstüberwachte Lernmethode für neuronale Netze, bei deren Verwendung keine Annotationen für Trainingsbilder benötigt werden. Ziel während des Trainings war es, dass sich zwei Ausschnitte desselben Bildes im Merkmalsraum ähnlicher sein sollten als Ausschnitte anderer Bilder. Die in iART verwendete Version des Modells wurde auf Bildern eines kunsthistorisch spezifischen Wikidata-Samples trainiert.
4. *CLIP*. CLIP (Contrastive Language-Image Pre-training; Radford et al. 2021) beinhaltet Modelle für visuelle und textuelle Informationen, die mithilfe von mehreren Millionen Bild-Text-Paaren trainiert wurden. Ziel des Trainings war es, dass visuelle und textuelle Darstellung für ein Bild-Text-Paar im Merkmalsraum möglichst nah beieinander liegen. Durch die multimodale Ausrichtung wird das Modell in iART auf zwei Weisen eingesetzt: Zum einen werden Suchanfragen mithilfe des Textmodells in den Bildbereich überführt. Dadurch können Nutzer:innen nicht nur nach extrahierten Tags oder Metadaten suchen, sondern Bilder auch direkt beschreiben. Zum anderen kann das Modell mit einem Zero-Shot-Ansatz textuelle Schlagwörter vorhersagen, ohne mit diesen trainiert worden zu sein. Wir verwenden diesen Ansatz, um 21.484 Notationen des etablierten Dezimalklassifikationssystems Iconclass (van de Waal 1973–85) zu generieren. Hierzu werden textuelle Merkmale für alle den Notationen zugehörigen Beschreibungen extrahiert und mit den visuellen Embeddings der entsprechenden Bilder verglichen. Im System gespeichert

werden alle Labels, die über einem bestimmten Schwellenwert liegen.

Der Einfluss jedes Plug-ins kann mithilfe von Schiebereglern präzise modifiziert werden, um Schwächen einzelner Modelle auszugleichen: Ein generischer ImageNet-Merkmalsextraktor wurde bspw. nur auf wenigen visuellen Konzepten trainiert, sodass eher im Bildhintergrund situierte Phänomene unberücksichtigt bleiben. Konträr dazu berücksichtigt das CLIP-Modell aufgrund der anderen Trainingsmethode wesentlich mehr Informationen des gesamten Bilds.

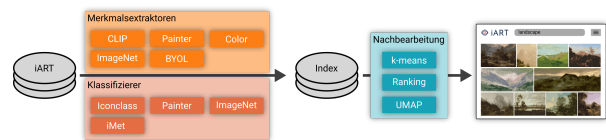


Abb. 2: Indizierungs- und Nachbearbeitungsschritte mit entsprechenden Plug-ins.

Frontend

Die webbasierte Benutzeroberfläche von iART wurde mit Vue.js⁷ erstellt und in JavaScript geschrieben. Sie integriert durch das UI-Framework Vuetify⁸ bewusst Googles Material Design. Diese Entscheidung hat zwei Gründe. Zum einen ist Google als Suchstandard etabliert: Plattformen, die sich stark von Google unterscheiden und nicht dessen Usability-Standards entsprechen, sind im Nachteil; wie zuletzt wieder Kröber, Münster und Messmer (2020) gezeigt haben. Zweitens soll der Zugang für Lai:innen nicht unnötig erschwert werden; allein die nicht Metadaten-getriebene Suche kann schließlich als zunächst gewöhnungsbedürftig empfunden werden. Dementsprechend klassisch ist die Positionierung der Einzelkomponenten in iART: Der altbekannte Suchschlitz befindet sich oben in der Mitte, während erweiterte Einstellungsmöglichkeiten in einem Banner darunter erscheinen (Abb. 3, oben).

Verschiedene Objektansichten vereinfachen die Exploration der Suchergebnisse. Standardmäßig wird ein Bildraster angezeigt, über das bei Bedarf weitere Details, wie z. B. Metadaten des jeweiligen Objekts, bereitgestellt werden. Die Ergebnisse können jeweils auf- und absteigend nach Relevanz, Titel oder Entstehungszeitpunkt sortiert werden. Eine Clustering visualisiert die Objekte als Bilderkarussells vertikal nach Gruppen getrennt (Abb. 3, Mitte). Der zugrundeliegende Algorithmus ist entsprechend der eigenen Forschungsinteressen konfigurierbar: Unterstützt wird aktuell das partitionierende Verfahren k-means; in Zukunft implementiert werden sowohl hierarchische als auch dichtebasierte Ansätze, bspw. DBSCAN. Für fortgeschrittene Anwendungsfälle ist es möglich, die Bilder mit der Dimensionsreduktionstechnik UMAP (Uniform Manifold Approximation and Projection; McInnes, Healy und Melville 2018) auf einer zweidimensionalen Leinwand anzuordnen, in der farbliche Markierungen die Gruppenzugehörigkeiten der jeweiligen Objekte indizieren (Abb. 3, unten). Durch den in SciPy implementierten Jonker-Volgenant-Algorithmus können die Bilder zudem überlappungsfrei in einem Raster positioniert werden (Virtanen et al. 2020, Crouse 2016). Zoom- und Filteroperationen, etwa ein interaktives Drag-Select zur Gegenüberstellung mehrerer Objekte, unterstützt iART mithilfe der Bibliothek vis.js.⁹

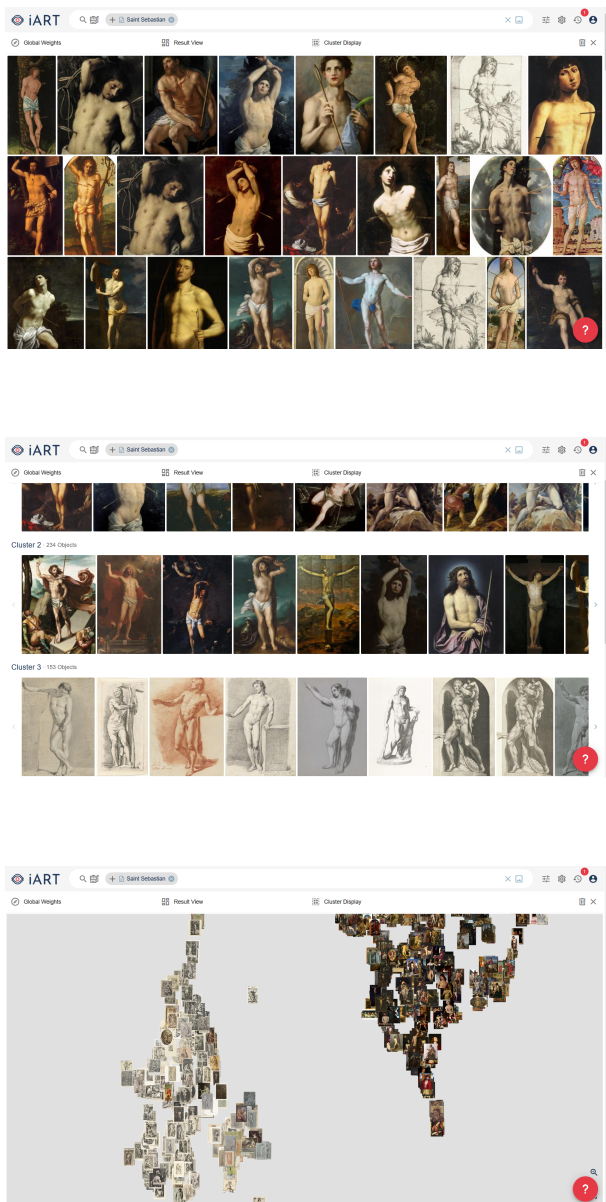


Abb. 3: Verschiedene Objektsichten zur Exploration der Suchergebnisse. Standardbildraster (oben), nach Gruppen getrennte Bilderkarussells (Mitte), Anordnung auf einer zweidimensionalen Leinwand (unten).

Datensätze

iART integriert ein breites Spektrum offen lizenzierter Bildinventare, das fortlaufend erweitert wird. Momentan bereitgestellt werden Daten aus fünf kunsthistorisch relevanten Quellen: des niederländischen Rijksmuseums (392.624 Objekte), der Wikidata (347.448 Objekte), des virtuellen Münzkabinetts KENOM (119.580 Objekte), der Social-Tagging-Plattform AR-Tigo (54.411 Objekte; Becker et al. 2018) und des Museumsportals Kulturerbe Niedersachsen (12.085 Objekte).¹⁰ Die Objekte wurden entweder mittels Web Scraping oder über offiziell verfügbare APIs extrahiert. Demnächst folgen u. a. Bestände des Musée du Louvre und Victoria and Albert Museums.¹¹ Zum Metadaten-gestützten Retrieval offeriert iART im Frontend eine

facettierte Suche, die die Objekte nach von den jeweiligen Institutionen vorgehaltenen Kategorien, z. B. Genre oder Medium, unscharf filtert (Abb. 4). Die Kategorien wurden manuell auf ein gemeinsames Schema überführt.

Um zu gewährleisten, dass selbst spezifische kunsthistorische Forschungsanliegen flexibel adressiert werden können, ist der Import von eigenen Datenbeständen für registrierte Nutzer:innen möglich. Zum einen werden bspw. als CSV-Datei bereitgestellte Metadaten für die facettierte Suche nutzbar gemacht, zum anderen in einer ZIP-Datei gebündelte Bildinhalte mit den zuvor beschriebenen Plug-ins analysiert. Anschließend können Nutzer:innen ihre hochgeladenen Sammlungen einzeln oder im Kontext mit frei verfügbaren Inhalten untersuchen. Damit ist eine Deep-Learning-gestützte Suche auch für Lai:innen praktikabel, die nicht nur auf übliche Schnittstellen, wie Googles Cloud Vision API,¹² zurückgreifen möchten.

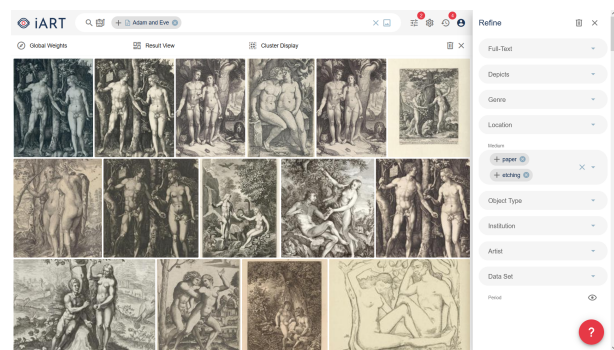


Abb. 4: Beispiel für eine nach Medium facettierte Suche mit der Abfrage „Adam and Eve“.

Use Cases

Die Vorteile von iART erschließen sich insbesondere bei Suchanfragen, die aufgrund ihrer semantischen Komplexität bislang nahezu unmöglich waren – oder nur in äußerst feingranular verschlagworteten Systemen sinnvolle Ergebnisse bringen. Die textbasierte Suche „Death of Marat“ gibt bspw. vier relevante Ergebnisgruppen zurück: erstens Jacques-Louis Davids „Der Tod des Marat“ (1793) in verschiedenen Reproduktionen, dazu weitere Beispiele für diese Ikonographie; zweitens Beispiele für Darstellungen des toten Christus, auf den sich auch die Marat-Darstellungen beziehen; drittens andere Figuren, denen der Arm in ähnlicher Weise herabhängt wie dem David’schen „Marat“; und viertens völlig andere Ikonographien, die formal – etwa in der Anordnung der gebogenen Linie vom Oberkörper des toten Marats und seinem Arm – auf den David’schen „Marat“ bezogen werden könnten (Abb. 5, oben).¹³ Wir sehen vor allem drei Anwendungsszenarien, die im Folgenden exemplarisch beschrieben werden:

1. *Suche nach ähnlichen Ikonographien.* Die Abfrage mit einem Bild von Leonardo da Vincis „Salvator Mundi“ (um 1500) führt zu einer Reihe von Salvator-Darstellungen, unter die sich nur wenige andere Ikonographien mischen.¹⁴ Dabei werden vor allem auch Objekte gefunden, die anders – z. B. in einer anderen Sprache – bezeichnet werden. Die vom System ermöglichte Clusterung mit k-means trägt zu einer weiteren Präzisierung der Ergebnismenge bei.

2. *Suche zur Rekontextualisierung.* Eine Recherche auf Basis von Vincent van Goghs „Die Kartoffelesser“ (1885) als Referenzbild weist auf spezifisch holländische bzw. nordeuropäische Darstellungen bäuerlicher Mahlzeiten und liefert eine empirische Begründung für bekannte Vermutungen zur Geschichte der Genreverteilung in der europäischen Malerei.¹⁵

3. *Suche im historischen Umkreis.* Eine Query-by-Image-Suche mit Tizians „Jacopo de Strada“ (1567/68) resultiert in einer Reihe von Halbfigurenporträts, die häufig aus dem venezianisch-norditalienischen Raum stammen (Abb. 5, unten).

¹⁶ Warum das so ist, lässt sich kaum eindeutig bestimmen: Es könnte mit Gestaltungseigenheiten zusammenhängen, dürfte letztendlich aber ein Beispiel für die Blackbox des Deep Learning sein. Gerade hier erweist sich die facettierte Suche mit ihrer Möglichkeit, die Objekte nach Medium einzuschränken, als sinnvoller Filtermechanismus, bei dem die Konzentration im norditalienischen Bereich am deutlichsten zur Geltung kommt.

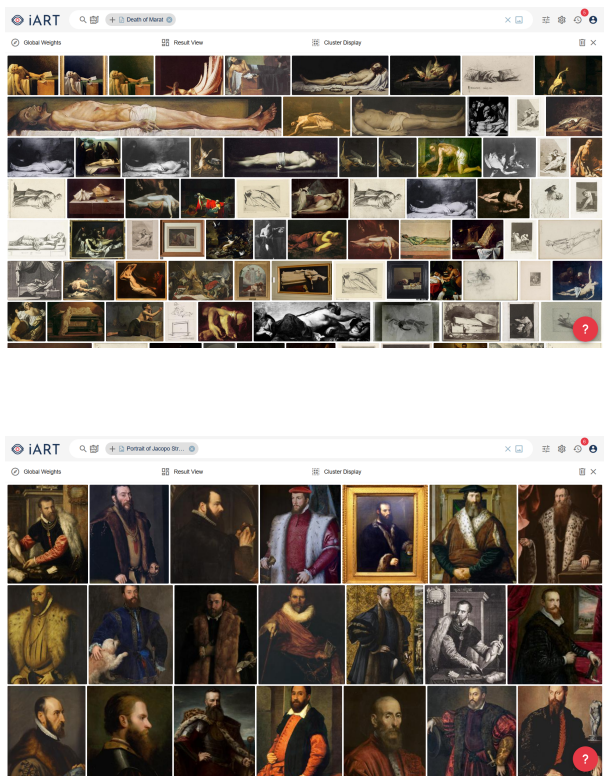


Abb. 5: Ergebnisse für eine Query-by-Text-Suche nach „Death of Marat“ (oben) und eine Query-by-Image-Suche nach Tizians „Jacopo de Strada“ (unten).

Fazit

Die Anwendungsszenarien zeigen, dass iART als unterstützendes Werkzeug für die kunst- und kulturwissenschaftliche Forschung dienen kann, indem es für eine Forschungsfrage interessante Bildobjekte identifiziert, extrahiert und analysiert. Da das System verschiedene Klassifizierungs-Plug-ins und Feature-Extraktoren unterstützt, können Nutzer:innen es an ihre Bedürfnisse anpassen. Auch die maschinell gesteuerte Suche sollte dabei prinzipiell als Anreiz zur weiteren Exploration verstanden werden und nicht als Instrument, das per se perfekte Ergebnisse liefert. Ge-

rade durch zunächst „unsinnig“ oder offensichtlich „falsch“ erscheinende Resultate können sich durchaus näher zu begutachtende Forschungsperspektiven ergeben.

Danksagung

Das Projekt „iART: Ein interaktives Analyse- und Retrieval-Tool zur Unterstützung von bildorientierten Forschungsprozessen“ wurde von der Deutschen Forschungsgemeinschaft (DFG) gefördert (Projektnummer 415796915).

Fußnoten

1. <https://iart.vision>, letzter Zugriff 30. November 2021.
2. <https://github.com/TIBHannover/iart>, letzter Zugriff 30. November 2021.
3. <https://oss.redis.com/redisai/>, letzter Zugriff 30. November 2021.
4. <https://www.djangoproject.com/>, letzter Zugriff 30. November 2021.
5. <https://www.kaggle.com/c/imet-2021-fgvc8>, letzter Zugriff 30. November 2021.
6. <https://www.wikiart.org/de> und <https://www.kaggle.com/c/painter-by-numbers/data>, jeweils letzter Zugriff 30. November 2021.
7. <https://vuejs.org/>, letzter Zugriff 30. November 2021.
8. <https://vuetifyjs.com/en/>, letzter Zugriff 30. November 2021.
9. <https://visjs.org/>, letzter Zugriff 30. November 2021.
10. <https://www.rijksmuseum.nl/en/>, <https://www.wikidata.org/>, <https://www.kenom.de/>, <https://www.artigo.org/> und <https://kulturerbe.niedersachsen.de/>, jeweils letzter Zugriff 30. November 2021.
11. <https://www.louvre.fr/en> und <https://www.vam.ac.uk/>, jeweils letzter Zugriff 30. November 2021.
12. <https://cloud.google.com/vision>, letzter Zugriff 30. November 2021.
13. <https://iart.vision/search?query=%2Btxt%3ADeath+of+Marat>, letzter Zugriff 30. November 2021.
14. <https://iart.vision/search?query=%2Bidx%3A3fa6b53c7e163ebd9663a01ab3efd24a>, letzter Zugriff 30. November 2021.
15. <https://iart.vision/search?query=%2Bidx%3A2e0d7a-d7a1733fdbbad8134b871dd8b0>, letzter Zugriff 30. November 2021.
16. <https://iart.vision/search?query=%2Bidx%3A673681d267163d84900c41c77ce951e2>, letzter Zugriff 30. November 2021.

Bibliographie

Becker, Matthias / Bogner, Martin / Bross, Fabian / Bry, François / Campanella, Caterina / Commare, Laura / Cramerotti, Silvia / Jakob, Katharina / Josko, Martin / Kneißl, Fabian / Kohle, Hubertus / Krefeld, Thomas / Levushkina, Elena / Lücke, Stephan / Puglisi, Alessandra / Regner, Anke / Riepl, Christian / Schefels, Clemens / Schemainda, Corina / Schmidt, Eva / Schneider, Stefanie / Schön, Gerhard / Schulz, Klaus / Sigmüller, Franz / Steinmayr, Bartholomäus / Störkle, Florian / Teske, Iris / Wieser, Christoph (2018): *ARTigo. Social Image Tagging [Dataset and Images]* 10.5282/ubm/data.136 .

Crouse, David F. (2016): „On Implementing 2D Rectangular Assignment Algorithms“, in: *IEEE Transactions on Aerospace and Electronic Systems* 52(4) 10.1109/TAES.2016.140952.

Grill, Jean-Bastien / Strub, Florian / Alché, Florent / Tallec, Corentin / Richemond, Pierre H. / Buchatskaya, Elena / Dörsch, Carl / Pires, Bernardo Ávila / Guo, Zhaohan Daniel / Azar, Mohammad Gheshlaghi / Piot, Bilal / Kavukcuoglu, Koray / Munos, Rémi / Valko, Michal (2020): „Bootstrap Your Own Latent. A New Approach to Self-Supervised Learning“, in: *Proceedings of Advances in Neural Information Processing Systems* 21271–21284.

Ohrt, Roberto / Heil, Axel / The Warburg Institute / Haus der Kulturen der Welt (Eds., 2020): *Aby Warburg. Bilderatlas Mnemosyne. The Original*. Berlin: Hatje Cantz.

He, Kaiming / Zhang, Xiangyu / Ren, Shaoqing / Sun, Jian (2016): „Deep Residual Learning for Image Recognition“, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

Kröber, Cindy / Münster, Sander / Messemer, Heike (2020): „Bildrepositorien und Forschung mit digitalen Bildern im Bereich der Kunstgeschichte“, in: Schöch, Christof (ed.): *DHd 2020. Spielräume. Digital Humanities zwischen Modellierung und Interpretation. 7. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum* 87–90 10.5281/zenodo.3666690.

Lang, Sabine / Ommer, Björn (2018): „Attesting Similarity. Supporting the Organization and Study of Art Image Collections with Computer Vision“, in: *Digital Scholarship in the Humanities* 845–856.

McInnes, Leland / Healy, John / Melville, James (2018): *UMAP. Uniform Manifold Approximation and Projection for Dimension Reduction* <https://arxiv.org/pdf/1802.03426.pdf> [letzter Zugriff 14. Juli 2021].

Offert, Fabian / Bell, Peter / Harlamov, Oleg (2020): *imgs.ai* <https://imgs.ai/> [letzter Zugriff 14. Juli 2021].

Radford, Alec / Wook Kim, Jong / Hallacy, Chris / Ramesh, Aditya / Goh, Gabriel / Agarwal, Sandhini / Sastry, Girish / Askell, Amanda / Mishkin, Pamela / Clark, Jack / Krueger, Gretchen / Sutskever, Ilya (2021): *Learning Transferable Visual Models From Natural Language Supervision* <https://arxiv.org/pdf/2103.00020.pdf> [letzter Zugriff 14. Juli 2021].

Rossetto, Luca / Giangreco, Ivan / Tanase, Claudiu / Schuldt, Heiko (2016): „vitriV. A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections“, in: *Proceedings of the 24th ACM International Conference on Multimedia* 1183–1186.

van de Waal, Henri (1973–85): *Iconclass. An Iconographic Classification System. Completed and Edited by L. D. Couprie with R. H. Fuchs*. Amsterdam / Oxford / New York: North-Holland Publishing Company.

Virtanen, Pauli / Gommers, Ralf / Oliphant, Travis E. / Haberland, Matt / Reddy, Tyler / Cournapeau, David / Burovski, Evgeni / Peterson, Pearu / Weckesser, Warren / Bright, Jonathan / van der Walt, Stéfan J. / Brett, Matthew / Wilson, Joshua / Millman, K. Jarrod / Mayorov, Nikolay / Nelson, Andrew R. J. / Jones, Eric / Kern, Robert / Larson, Eric / Carey, CJ / Polat, İlhan / Feng, Yu / Moore, Eric W. / VanderPlas, Jake / Laxalde, Denis / Perktold, Josef / Cimrman, Robert / Henriksen, Ian / Quintero, E.A. / Harris, Charles R. / Archibald, Anne M. / Ribeiro, Antônio H. / Pedregosa, Fabian / van Mulbregt, Paul / SciPy 1.0 Contributors (2020): „SciPy 1.0. Fundamental Algorithms for Scientific Computing in Python“, in: *Nature Methods* 17(3): 261–272.

Warnke, Martin / Brink, Claudia (Eds., 2000): *Aby Warburg. Gesammelte Schriften. Der Bilderatlas Mnemosyne*. II(1), Berlin: Akademie Verlag.

Wölfflin, Heinrich (1915): *Kunstgeschichtliche Grundbegriffe. Das Problem der Stilentwicklung in der neueren Kunst*. München: Bruckmann.

Inter Annotator Agreement und Intersubjektivität

Ein Vorschlag zur Messbarkeit der Qualität literaturwissenschaftlicher Annotationen

Gius, Evelyn

evelyn.gius@tu-darmstadt.de
Technische Universität Darmstadt, Germany

Vauth, Michael

michael.vauth@tu-darmstadt.de
Technische Universität Darmstadt, Germany

Annotationen als etablierte Praxis der DH

Die in den Sozialwissenschaften und der Computerlinguistik schon lange etablierte Praxis der computergestützten und häufig kollaborativen manuellen Annotation ist mittlerweile auch im Zentrum der digitalen Geisteswissenschaften angekommen. Deshalb möchten wir unsere Beobachtungen zu einem zentralen Punkt teilen: dem Inter Annotator Agreement bzw. Inter Coder Agreement. Wir betrachten dieses aus der Sicht der Computational Literary Studies (CLS) anhand unseres Projekts „Evaluating Events in Narrative Theory (EvENT)“¹. In diesem annotieren wir Ereignisse als kleinste Handlungseinheiten in Prosatexten und nutzen die Annotationen, um die Erkennung der Ereignisse zu automatisieren.² Diese automatisierte Ereignisanalysen können anschließend für korpusbasierte Untersuchungen zur Ereignishaftigkeit oder generell zur Ereignisstruktur literarischer Texte genutzt werden. Wie wir zeigen möchten, spielt das Inter Annotator Agreement eine vielfältige Rolle in der Arbeit an und mit manuellen Annotationen und sollte möglichst im Einklang mit der Praxis des Erkenntnisgewinns in der Literaturwissenschaft genutzt und weiterentwickelt werden.

Zum Nutzen von Inter Annotator Agreement-Messungen

Es gibt eine Vielzahl von Inter Annotator Agreement-Metriken, die als Maß eingesetzt werden, um die Verlässlichkeit manuell erstellter Annotationen zu beurteilen, die zum Überprüfen einer These oder zur Entwicklung und zum Testen computatio-

eller Modelle genutzt werden (Artstein & Poesio 2008:556). Da bei der Betrachtung des Inter Annotator Agreements von Menschen annotierte Daten – und damit deren Analysen bestimmter Texte (oder anderer Artefakte) – miteinander verglichen werden, ist dies auch literaturwissenschaftlich interessant. Der literaturwissenschaftliche Erkenntnisgewinn basiert nämlich, in Ermangelung objektiver Fakten, ganz wesentlich auf intersubjektiver Übereinstimmung bzw. deren Abgleich.

Grundsätzlich lassen sich fünf Einsatzgebiete von Inter Annotator Agreement-Messungen unterscheiden:

1. Man kann mittels Inter Annotator Agreement-Messung feststellen, wie hoch die *Reliabilität einzelner Annotator*innen* ist. Dies wird etwa genutzt, wenn man aus Ressourcengründen größtenteils nur eine:n Annotator:in die Texte annotieren lässt und einschätzen möchte, ob diese:r die Annotationen in der gewünschten Qualität anfertigt.
2. Bei der *Entwicklung von Guidelines* können Inter Annotator Agreement-Messungen als Heuristik zum Aufdecken nicht übereinstimmender Annotationen genutzt werden, deren Erkenntnisse anschließend in die Überarbeitung des Annotationsworkflows und insbesondere der Guidelines einfließen.
3. Auch für eine Bewertung oder auch den Vergleich der *Qualität von Guidelines* werden Inter Annotator Agreement-Verfahren genutzt. Gute Guidelines sollten eine hohe Übereinstimmung zwischen Annotationen ermöglichen. Dies ist in Verfahren wichtig, in denen das annotierte Korpus zum Testen einer Hypothese konzipiert wird. Dort werden nämlich die Guidelines nicht im Annotationsprozess überarbeitet, sondern stehen mit Beginn der Annotation fest (vgl. dazu Artstein & Poesio 2008 bzw. Krippendorff 2004).
4. Desweiteren wird in Fällen, in denen Korpora – wie etwa Referenzkorpora – zur Nachnutzung für weitere Forschung aufgebaut werden, die *Qualität bzw. Validität der Daten* an den Inter Annotator Agreement-Angaben gemessen.
5. Schließlich können Inter Annotator Agreement-Werte auch genutzt werden, um etwas über die *annotierten Phänomene bzw. deren Operationalisierbarkeit* auszusagen. Bei einem hohen Inter Annotator Agreement-Wert kann man von einer geringen Komplexität der annotierten Phänomene bzw. einer guten Operationalisierbarkeit ihrer Bestimmung (in Form der Guidelines) ausgehen. Das Inter Annotator Agreement ist dann ein Maß, das zum einen die Schwierigkeit der Automatisierungsaufgabe beschreibt (je geringer das Agreement, desto anspruchsvoller die Aufgabe) und zum anderen die Qualität der Automatisierung evaluiert.

Für das EvENT-Projekt sind diese Einsatzbereiche unterschiedlich stark von Interesse. Wir nutzen Inter Annotator Agreement für (1) die Reliabilität von Annotator:innen und die (2) Entwicklung von Guidelines. Die (3) Bewertung der Qualität der Guideline spielt im EvENT-Projekt nur eine nachgeordnete Rolle. Im Gegensatz zur Computerlinguistik gibt es nämlich für die CLS bislang mangels Erfahrung keine Inter Annotator Agreement-Werte, an denen man sich orientieren kann. Dasselbe gilt für (4) Datenvalidität und (5) Operationalisierbarkeit.

Inter Annotator Agreement als literaturwissenschaftliches Qualitätskriterium

Literaturwissenschaftliche Befunde basieren meistens weder auf streng formalisierten Schlussfolgerungssystemen³ noch ist mit ihnen der Anspruch verbunden, eine empirische Wahrheit abzubilden. Die Wissenschaftlichkeit der Befunde wird vielmehr durch ihre “prinzipielle intersubjektive Vermittelbarkeit – einen ‘*sensus communis*’” garantiert.⁴ Literaturwissenschaftliche Analyse bedeutet in einem ersten Schritt, ohne Wertung “die Feststellung von allgemein beobachtbaren und intersubjektiv anerkannten Eigenheiten bestimmter Texte zu fixieren”⁵. Dieser Anspruch der intersubjektiven Vermittelbarkeit von beobachtbaren Texteeigenschaften legt nahe, dass Inter Annotator Agreement-Maße geeignete Kandidatinnen für das ‘Messen’ von Intersubjektivität sind.

Mit Blick auf Intersubjektivität kann man in den fünf genannten Bereichen, in denen Inter Annotator Agreement-Messungen zum Einsatz kommen, feststellen: Im Kontext der Reliabilität von Annotator*innen (Fall 1) geht es um den Abgleich einer an sich aber *intrasubjektiven* Qualität, nämlich die Frage, wie gut Annotator:innen annotieren bzw. welche besonders gut sind. Intersubjektivität spielt hier eine untergeordnete Rolle.

Bei der Entwicklung bzw. Qualität von Guidelines (Fall 2 bzw. 3) geht es hingegen um die Frage, inwiefern eine *Guideline* ein geteiltes Verständnis von Phänomenen unterstützt. Damit geht es um die intersubjektive Übereinstimmung bei der Interpretation der Guidelines, die sich in den Annotationen niederschlägt.

Im Kontext der Qualität bzw. Validität der Daten und der Operationalisierbarkeit von Phänomenen (Fall 4 und 5) steht schließlich die intersubjektive Übereinstimmung bei der Beurteilung der Phänomene im Text im Fokus.

Aus literaturwissenschaftlicher Sicht ist die Intersubjektivität insbesondere in den letzten beiden Fällen abgebildet. Bei der Frage nach Qualität bzw. Validität der Daten und der Operationalisierbarkeit von Phänomenen wird nämlich der Grad der Übereinstimmung zwischen Annotationen auf die oben erwähnten „Eigenheiten bestimmter Texte“ bezogen. Die beiden Aspekte sind auch aus computationeller Sicht wichtig, denn sie betreffen die analysierten Phänomene und damit das zentrale Forschungsinteresse vieler literaturwissenschaftlicher Ansätze in den Digital Humanities. Wie bereits angesprochen, fehlen allerdings gerade zu diesen beiden Fällen Erfahrungswerte, auf die zurückgegriffen werden kann. Da die Inter Annotator Agreement-Werte in literaturwissenschaftlichen Annotationsprojekten zudem meist deutlich unter den in anderen Disziplinen gängigen Grenzwerten liegen, können diese nicht sinnvoll genutzt werden. Stattdessen müssen Strategien entwickelt werden, die eine Beurteilung der Annotationsqualität in philologischen Forschungskontexten ermöglichen.

Wir stellen deshalb im Folgenden eine Anpassung des Verfahrens der Annotation und der Inter Annotator Agreement-Messung vor, mit der man diesem Manko in bestimmten Forschungszusammenhängen begegnen kann.

Literaturwissenschaftlich adäquate Inter Annotator Agreement-Messung

Inter Annotator Agreement-Metriken basieren auf differenzierten Formeln, die typischerweise erwartete (Nicht-)Übereinstimmungswerte berücksichtigen und z.T. auch die Gewichtung bestimmter Aspekte der Annotationen zulassen (z.B. durch das Festlegen von Ähnlichkeiten zwischen Kategorien oder die Gewichtung der Segmentierungsentscheidungen). Die Wahl der eingesetzten Metrik sollte in Abhängigkeit von den Eigenschaften der Annotationen getroffen werden. Zu diesen Eigenschaften gehören die Anzahl und Verteilung der genutzten Annotationskategorien, die Häufigkeit, mit der Annotationskategorien auftreten, die Frage, ob die Bestimmung der zu annotierenden Textsegmente Teil der Annotationsaufgabe ist und viele mehr (vgl. dazu Artstein & Poesio 2008 sowie Mathet et al. 2015). Das Problem, vor dem wir zumindest bislang stehen, ist nicht nur, dass es eine ziemliche Herausforderung ist, diese Eigenschaften zu identifizieren, sondern noch mehr, dass uns etablierte Strategien fehlen, um diese zu beurteilen.

Ein wesentlicher Grund dafür ist, dass literaturwissenschaftliche Textanalysen oft Phänomene in den Blick nehmen, die bei näherer Betrachtung keine Merkmale der Textoberfläche sind. Da diese Phänomene nicht direkt an bestimmten Texteneigenschaften festgemacht werden können, muss man bei der Operationalisierung auf mit dem Phänomen mutmaßlich zusammenhängende Merkmale zurückgreifen, die sich textlich realisieren.⁶ So modellieren wir im EvENT-Projekt die Ereignishaftigkeit von Texten mit der vergleichsweise granularen Annotation von Verbalphrasen, da wir diese in unseren Untersuchungen als kleinste Textspannen identifiziert haben, die auf ein Ereignis referieren können.⁷ Wir annotieren also Mikrophänomene auf der Textoberfläche, um ein erzähltheoretisches Makrophänomen zu beschreiben, das sich nicht unmittelbar an der Textoberfläche manifestiert.

Eine Folge dieser indirekten Annäherung an die untersuchten Phänomene ist, dass eine Agreement-Messung mit den üblichen Metriken für bestimmte literaturwissenschaftliche Einsatzgebiete nicht sinnvoll ist, da diese für die Annotation von Textphänomenen wie etwa Wortarten oder semantische Klassen entwickelt wurden.

Nun könnte man versuchen neue, für literaturwissenschaftliche Fragestellungen passende Annotationsmetriken zu entwickeln. Ähnlich hilfreich und leichter umsetzbar ist allerdings eine Anpassung des Operationalisierungsverfahrens an das, was mit bestehenden Metriken gemessen wird.

Konkret sollte man versuchen, die genutzten Annotationskategorien so zu gestalten, dass sie:

1. einen möglichst klaren Textumfang haben sowie möglichst im ganzen Text vorkommen und
2. numerischen Werten belegt werden können.

Beim ersten Punkt ist es erstrebenswert, dass die genutzten Kategorien eine möglichst eindeutig festlegbare Texteinheit umfassen und im ganzen Text vorkommen.⁸ Der zweite Punkt bedeutet, dass die genutzten Kategorien möglichst in numerische Werte überführt werden. (Dies ist übrigens in allen Fällen ein hilfreicher Schritt, in welchem es darum geht, quantifizierend mit Annotationen umzugehen.) Dies bedeutet nicht nur, dass man Annotationskategorien in numerische Werte überführt, sondern auch, dass die Werte in Bezug auf ihren Intervall bedeutungshaft sind und es

zudem idealerweise auch einen absoluten Nullpunkt gibt, zu dem sie im Verhältnis stehen. Der Grad der Umsetzbarkeit dieser Vorschläge hängt natürlich von der Forschungsfrage und dem untersuchten Phänomen ab.

Eine mögliche Umsetzung dieser Punkte lässt sich an unserem Beispiel verdeutlichen.

Inter Annotator Agreement-Messung im EvENT-Projekt

Ausgehend von den erzähltheoretischen Ereigniskonzepten haben wir im EvENT-Projekt vier Annotationskategorien definiert:

- `change_of_state`: Die Verbalphrase referiert auf die Zustandsveränderung einer Entität in der erzählten Welt (Diegese)
- `process_event`: Die Verbalphrase referiert auf einen zeitlichen Vorgang in der erzählten Welt, der keine Zustandsveränderung enthält.
- `stative_event`: Die Verbalphrase referiert auf einen Sachverhalt in der erzählten Welt, der keine zeitliche Dimension hat.
- `non_event`: Die Verbalphrase oder ein elliptisches Textsegment referiert nicht auf einen Sachverhalt in der erzählten Welt.

Wir haben also eine syntaktisch weitgehend eindeutige Einheit – die Verbalphrase – identifiziert, die sich als Annotationseinheit eignet und deren Inhalt zur Bestimmung der Kategorisierung geeignet ist. Durch die Ausweitung der `non_event`-Kategorie auf nicht vollständige Verbalphrasen kann ein Text außerdem durchgängig mit unseren Kategorien annotiert werden kann.

Auch die Überführung der kategorialen Skalierung in eine numerische Skalierung basiert auf dem literaturwissenschaftlichen Verständnis der Kategorien. Entsprechend dem literaturwissenschaftlichen Ereignisverständnis nehmen wir an, dass diese vier Kategorien in unterschiedlichem Maß die Ereignishaftigkeit eines Textes konstituieren: Zustandsveränderungen, aber auch Bewegungs- und Kommunikationsvorgänge tun dies in stärkerem Maß als Landschafts-, Raum- oder Figurenbeschreibungen, die in vielen erzählenden Texten eher Expositionsfunktionen erfüllen.⁹ Aus diesem Grund haben wir die Narrativität der Annotationskategorien mit folgenden Werten festgelegt:

- `change_of_state`: 7
- `process_event`: 5
- `stative_event`: 2
- `non_event`: 0

Doch dies war noch nicht ausreichend, um ein Agreement zu erzielen, welches aus computerlinguistischer Sicht gut ist. Hinzu kommt, dass die Agreement-Werte unsere Intuition über die Qualität der Annotationen nicht widerspiegeln (vgl. Tabelle 1).

Tab. 1: Inter Annotator Agreement-Werte für annotierte Texte

	Erdbeben in Chili	Krambambuli	Effi Briest
Cohen's κ	0.73	0.63	0.58
Krippendorff's α	0.73	0.63	0.58

Deshalb haben wir unser Vorgehen entsprechend weiterentwickelt. Der Schlüssel zu einer aussagekräftigeren Inter Annotator Agreement-Perspektive lag in der Erkenntnis, dass uns die Entwicklung von Ereignishaftigkeit im Textverlauf und entsprechend Narrativitätsverläufe interessieren.

Wir haben deshalb nicht nur die Ergebnisse der Annotationen als Verlauf visualisiert, sondern auch entschieden, die Einschätzung des Inter Annotator Agreement – ebenso wie übrigens die Qualität der automatisierten Erkennung von Ereignissen – anhand von Verläufen vorzunehmen.

Für die Darstellung des Narrativitätsverlaufs wurden die Werte der Annotationen innerhalb eines Textabschnitts anhand der Narrativitätswerte der umliegenden 50 Verbalphrasen mit einer Kosinuskewichtung geglättet. Die Kosinuskewichtung sorgt dabei dafür, dass näher liegende Textsegmente einen stärkeren Einfluss auf den Narrativitätswert des untersuchten Textsegments haben.

Auf Grundlage dieser Zuweisungen konnten wir die Narrativitätsverläufe in Einzeltexten wie in Abbildung 1 untersuchen:

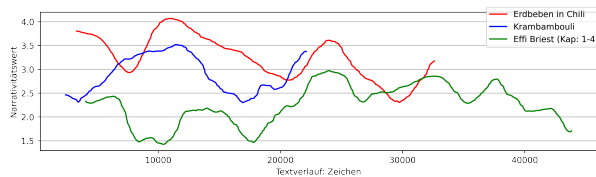


Abb. 1: Narrativitätsverläufe zu drei Prosatexten aus dem Event-Projekt.

Um die Stabilität des Verfahrens zu prüfen, haben wir mit der Zuweisung der Zahlen zu den Kategorien experimentiert, dabei aber ihre Anordnung gemäß ihrer Narrativität nicht verändert. Eine umfassende Evaluation steht noch aus, aber die bisherigen Versuche deuten darauf hin, dass die Narrativitätsverläufe dabei strukturell nicht stark variieren (vgl. Abbildung 2).

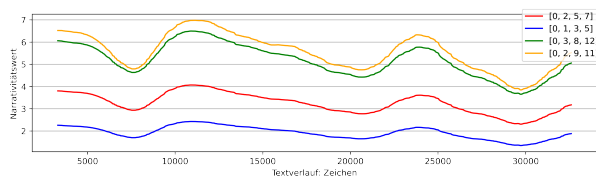


Abb. 2: Narrativitätsverläufe von Kleists *Das Erdbeben in Chili* bei variierenden Skalierungen der Ereignistypen. Die Zahlenlisten in der Legende geben die verwendeten Werte für *non_events*, *stative_events*, *process_events* und *change_of_states* in dieser Reihenfolge an.

Wir konnten also auf der Grundlage unserer Wertzuweisung für die Ereignistypen die Annotationen der unterschiedlichen Annotator:innen miteinander vergleichen (vgl. Abbildung 3).

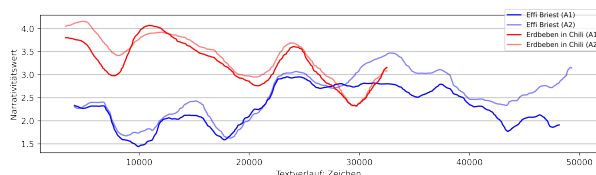


Abb. 3: Narrativitätsverläufe in Kleists *Das Erdbeben in Chili* und *Effi Briest* zum Vergleich der beiden Annotator:innen (A1 und A2).

Die Narrativitätsgraphen ähneln sich deutlich stärker als es angesichts des eher niedrigen Inter Annotator Agreement zu vermuten gewesen wäre. Das drückt sich auch in der mittleren bis starken Korrelation der Graphen aus (vgl. Tabelle 2).¹⁰

Tab. 2: Korrelation (Pearson) der Narrativitätsgraphen der annotierten Texte.

	Erdbeben in Chili	Krambambuli	Effi Briest
Korrelation (Pearson)	0.94	0.81	0.8

Unsere Annäherung an ein Inter Annotator Agreement, das auf die Modellierung eines literarischen Phänomens ausgerichtet ist, scheint also unsere literaturwissenschaftlich fundierte Intuition besser abzubilden als gängige Inter Annotator Agreement-Metriken.

Dafür sind zwei Aspekte entscheidend:

1. Wenn die Annotator:innen systematische Fehler machen, die sich auf ein unterschiedliches Verständnis der Annotationsrichtlinien zurückführen lassen, schlägt sich das durch die Parametrisierung nur insofern wieder, als der Narrativitätswert im Durchschnitt etwas niedriger oder höher ist. Die strukturellen Eigenschaften des Graphen berührt das kaum. Ein Beispiel: In den vier längeren Abschnitten aus *Effi Briest* zeigt sich in 3 von 4 Abschnitten eine Tendenz von Annotator:in 1 (A1) häufiger *non_events* und seltener *stative_events* zu annotieren. Annotator:in 2 (A2) identifiziert in drei Abschnitten seltener *process_events* (vgl. Tabelle 3).
2. Das Glättungsverfahren, das ein notwendiger Schritt ist, um die Parametrisierung der Mikrophenomene zur Modellierung von Makrophenomenen umzusetzen, nivelliert Flüchtigkeitsfehler bei der Annotation und das hinsichtlich der Segmentierung und der Klassifizierung.

	Teil 1		Teil 2		Teil 3		Teil 4	
	A1	A2	A1	A2	A1	A2	A1	A2
<i>non_event</i>	0.43	0.37	0.38	0.36	0.43	0.40	0.45	0.50
<i>stative_event</i>	0.28	0.38	0.31	0.38	0.31	0.40	0.28	0.28
<i>process_event</i>	0.37	0.25	0.30	0.30	0.25	0.20	0.25	0.22
<i>change_of_state</i>	0.02	0.01	0.01	0.00	0.00	0.00	0.01	0.00

Tab. 3: Bias der beiden Annotator:innen (A1 und A2) bei der Annotation von Ereignistypen in *Effi Briest* (vier Teile mit einem Umfang von je 4–7 Kapitel).

Durch dieses Vorgehen gelingt es uns, den Fokus auf das eigentlich untersuchte Phänomen – in unserem Fall die Ereignishaftekeit von erzählenden Texten – zu richten. Damit lässt sich die intersubjektivität der Analysen besser messen als anhand der Annotationen, die das Phänomen anhand von Oberflächenphänomenen (Verbalphrasen) operationalisieren und die im Kontext von gängigen Inter Annotator-Metriken entsprechend nur bedingt aussagekräftig sind. Hinzu kommt, dass es zwei wichtige Fehlerquellen bei literaturwissenschaftlichen Annotationen – nämlich einfache Fehler sowie divergierende Voranalysen (vgl. Gius & Jacke 2017) – ausgleicht.

Fußnoten

1. Das Projekt wird im DFG-Schwerpunktprogramm Computational Literary Studies (SPP 2207) gefördert und wird seit 09/2020 an der Technischen Universität Darmstadt und der Universität Hamburg durchgeführt.
2. vgl. zu den ersten Automatisierungsergebnissen Hans Ole Hatzel, Michael Vauth, Chris Biemann und Evelyn Gius (2021).
3. vgl. Danneberg und Albrecht (2016), insbesondere S. 6–8.
4. vgl. Stöckmann 2013, S. 475.
5. vgl. Fricke et al. 2000, S. 447.

6. Vgl. dazu die sogenannten 'instrumental variables' von Graham Sack, die eingesetzt werden, wenn keine die eigentlichen Phänomene nicht messbar sind (Moretti 2013:104).
7. Vgl. dazu unsere Guideline (Vauth & Gius 2021) unter <http://doi.org/10.5281/zenodo.5078175>.
8. Dadurch werden potentielle Probleme mit den Aspekten *Unitizing* und *Sporadicity* (Mathet et al. 2015) verringert. Die weiteren für Inter Annotator Agreement relevanten Aspekte *Categorization*,
9. Es gibt einige zusätzliche Ereignisseigenschaften, die die Ereignishaftigkeit eines Textes determinieren und in unserem Annotationsschema Berücksichtigung erfahren haben. Auf ihnen liegt in diesem Beitrag allerdings nicht unser Fokus.
10. Für den Vergleich filtern wir die Annotationen der beiden Annotator:innen so, dass sie einen etwa gleichen Startpunkt haben (+/- 3 Zeichen). Wie bei üblichen Verfahren der Inter Annotator Agreement-Messung wird so verhindert, dass zwei Annotationen miteinander verglichen werden, die sich nicht auf die gleiche Textspanne beziehen.

Bibliographie

- Artstein, Ron, und Massimo Poesio.** 2008. „Inter-Coder Agreement for Computational Linguistics“. *Computational Linguistics* 34 (4): 555–96. <https://doi.org/10.1162/coli.07-034-R2>.
- Danneberg, Lutz, und Andrea Albrecht.** 2016. „Beobachtungen zu den Voraussetzungen des hypothetisch-deduktiven und des hypothetisch-induktiven Argumentierens im Rahmen einer hermeneutischen Konzeption der Textinterpretation“. *Journal of Literary Theory* 10 (1): 1–37. <https://doi.org/10.1515/jlt-2016-0001>.
- Fricke, Harald, Klaus Grubmüller, Jan-Dirk Müller, und Klaus Weimar,** Hrsg. 2000. *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. 3., Neubearb. Aufl. Berlin: De Gruyter.
- Gius, Evelyn, und Janina Jacke.** 2017. „The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis“. *International Journal of Humanities and Arts Computing* 11 (2): 233–54. <https://doi.org/10.3366/ijhac.2017.0194>.
- Krippendorff, Klaus.** 2004. „Reliability in Content Analysis: Some Common Misconceptions and Recommendations“. *Human Communication Research* 30 (3): 411–33. <https://doi.org/10.1093/hcr/30.3.411>.
- Mathet, Yann, Antoine Widlöcher, und Jean-Philippe Métivier.** 2015. „The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment“. *Computational Linguistics* 41 (3): 437–79. https://doi.org/10.1162/COLI_a_00227.
- Moretti, Franco.** 2013. „‘Operationalizing’. Or, the Function of Measurement in Literary Theory“. *New Left Review*, Nr. 84 (Dezember): 103–19.
- Stöckmann, Ingo.** 2013. „Ästhetik“. *Handbuch Literaturwissenschaft*, herausgegeben von Thomas Anz, 1:465–91. Stuttgart: J.B. Metzler. <https://doi.org/10.1007/978-3-476-01271-5>.
- Vauth, Michael, und Gius, Evelyn.** 2021. *Richtlinien für die Annotation narratologischer Ereigniskonzepte*, Juli. <https://doi.org/10.5281/ZENODO.5078175>.
- Vauth, Michael, Hatzel, Hans Ole, Biemann, Chris, und Evelyn Gius.** 2021. „Automated Event Annotation in Literary Texts“ Workshop on Computational Humanities Research 2021. *CHR 2021: Computational Humanities Research Conference*, 333–45. Amsterdam, The Netherlands. http://ceur-ws.org/Vol-2989/short_paper18.pdf

Japanese Visual Media Graph Bündelung des Wissens von Fan-Gemeinschaften in einem domänenspezifischen Knowledge Graph

Pfeffer, Magnus

pfeffer@hdm-stuttgart.de
Hochschule der Medien Stuttgart, Deutschland

Kacsuk, Zoltan

kacsuk@hdm-stuttgart.de
Hochschule der Medien Stuttgart, Deutschland

Roth, Martin

roth1003@fc.ritsumei.ac.jp
Ritsumeikan University, Kyoto, Japan

Einleitung

Die Umstellung auf digitale Vertriebswege hat visuelle Medien in einer Fülle verfügbar gemacht, die zuvor nur schwer vorstellbar war. Medien für spezielle Zielgruppen und Nischen haben nun ein potenziell globales Publikum und Teile dieses Publikums formieren sich in Fan-Gemeinschaften, die sich auf eigens entwickelten Online-Plattformen über „ihre“ Medien austauschen, sie analysieren, katalogisieren und Informationen über sie sammeln (Price & Robinson 2017). Einige dieser Gemeinschaften fokussieren sich auf bestimmte Genres, andere auf ein Herkunftsland oder gar nur das Werk einzelner Autor:innen. Schon die Menge von Daten zu visuellen Medien, die von diesen Gemeinschaften produziert wird, ist beeindruckend, noch mehr aber ist es die Qualität der Datenmodellierung, die Liebe zum Detail und der hohe Grad an Koordination in den Gemeinschaften.

Für Wissenschaftler:innen, die zu visuellen Medien forschen, sind die Daten dieser Gemeinschaften sehr wertvoll, denn sie liefern die benötigte Kontextualisierung und dokumentieren Beziehungen zwischen Werken, Genres und Medienarten. Gleichzeitig geben sie Auskunft über die Rezeption und Bewertung von Medien durch die verschiedenen Zielgruppen und erlauben einen Einblick in lokale und globale Medienkulturen. Die Nutzung dieser Daten ist jedoch mit Herausforderungen verbunden: Zum einen ist es von außen nur schwer einzuschätzen, wie die unterschiedlichen Gemeinschaften in Bezug auf Arbeitsweise und Qualitätsstandards arbeiten, und zum anderen kann es je nach Fragestellung erforderlich sein, Daten aus mehreren Quellen zu kombinieren – was erweiterte Kompetenzen im Bereich der Data Science und beim Modellieren von Daten erfordert. Nicht zuletzt sind die kleineren Gemeinschaften inhärent fragil und haben keine Ressourcen außer der Bereitschaft ihrer Mitglieder, sich an Aufbau, Programmierung und Finanzierung der Online-Angebote zu beteiligen.

Das Japanese Visual Media Graph Projekt

In diesem Projekt sollen Methoden entwickelt werden, um die von den Fan-Gemeinschaften gesammelten Daten für die Wissenschaft zugänglich zu machen und dauerhaft zu sichern. Dies umfasst Software-Werkzeuge, Workflows und die Dokumentation von Best-Practice-Verfahren für das Entdecken, Extrahieren, Sammeln, Konsolidieren und Verknüpfen der Daten. Die aufbereiteten Daten sollen in einem zentralen Repositorium zur Verfügung gestellt und mit einem an den Bedürfnissen der Forschenden orientierten Benutzerinterface versehen werden.

Auch wenn diese Methoden und Vorgehensweisen weitestgehend universell anwendbar sind, beschränkt sich das Projekt zunächst auf die Domäne der japanischen visuellen Medien mit einem Fokus auf Manga, Anime und Computerspielen. Diese Medien haben in den vergangenen zwei Jahrzehnten einen wahren Boom erlebt und sind wesentlicher Bestandteil der japanischen Soft-Power-Strategie, die auch als "cool Japan" bekannt ist (McGray 2002, Oyama 2016, Valaskivi 2013). Dieser Teilbereich der visuellen Medien ist eine besondere Herausforderung für die Erprobung unserer Ansätze und ermöglicht zugleich einen signifikanten Betrag zur Forschung. Die japanische Kreativ-Industrie präferiert weitläufige Story-Universen und ein auf die visuellen Charaktere aufbauendes cross-media Franchising, das auch als "media-mix" bezeichnet wird (Steinberg 2012, Picard & Pelletier-Gagnon 2015, Nozawa 2013), und hat eine enge und besondere Beziehung zu Fan-Werken und Fan-Aktivitäten (Condry 2013). Aus dieser Praxis entsteht ein komplexes und weites Netzwerk aus Werken, das einen Ansatz erforderlich macht, der die Verbindungen zwischen den Inhalten, Genres und Charakteren über Mediengrenzen und Datenquellen hinweg detailliert beschreiben kann. Dazu kommt, dass ein signifikanter Teil der Inhalte und des Contents nur innerhalb Japans verfügbar ist, was das Sammeln und Validieren von Informationen darüber erschwert.

Im Rahmen der Vorarbeiten zu dem Projekt wurden zwei Untersuchungen durchgeführt: Im Rahmen eines Seminars haben Studierende über 40 Fan-Websites zu japanischen visuellen Medien untersucht und nach einheitlichen Kriterien beschrieben. Die Bandbreite reichte von vergleichsweise großen, internationalen Gemeinschaften bis zu kleineren lokalen Gruppen, deren Webseiten nicht in englischer Sprache gehalten sind. Die Untersuchung zeigte, dass nahezu alle Webseiten neben der Möglichkeit, sich online mit Gleichgesinnten auszutauschen, in irgendeiner Form Daten zu den Medien selbst, ihren Inhalten und Genres oder den Charakteren sammeln und systematisch aufbereiten. Diese Datensammlungen werden dem Umfang und der Komplexität dieser Medienwelt mehr als gerecht und zeichnen sich durch einen hohen Detailgrad in der Beschreibung der Medien, große Sorgfalt bei der Kuratierung der Daten und ein beeindruckendes technisches Niveau aus.

Weiterhin wurde in einer qualitativen Befragung der tatsächliche Bedarf in der Fachwissenschaft abgefragt. Zielgruppe waren dabei sowohl Forschende aus dem Bereich der Japanologie, die an modernen Medien interessiert sind, als auch Forschende aus den Medienwissenschaften, sofern Japan in deren Fokus steht. Die Interviews wurden vor Ort auf Konferenzen und Workshops sowie per Email und Videokonferenz geführt. In der Auswertung zeigten sich drei Aspekte, die einen wesentlichen Einfluss auf die Projektziele hatten: Zum einen wird das Feld als ein interessanter Forschungsgegenstand gesehen, was sich in der steigenden Zahl einschlägiger Veröffentlichungen, Workshops und Konfe-

renzen niederschlägt. Dazu kommt ein erkennbarer Mangel an Informationsressourcen, die einen Zugang zu dieser Medienwelt ermöglichen. Bestehende Lexikon-artige Einzelveröffentlichungen wurden als veraltet oder nicht übergreifend genug und zu sehr auf einzelne Aspekte fokussiert wahrgenommen. Die von Fan-Gemeinschaften gesammelten Informationen waren teilweise bekannt, aber der uneinheitliche Zugang, Unklarheit über die Qualität und Vorgehensweise und die Befürchtung, dass diese unvermittelt nicht mehr zur Verfügung stehen könnten, wurden als Hindernisse für deren Nutzung gesehen. Als dritter Aspekt wurde der Wunsch genannt, mehr Optionen für die Bearbeitung von Forschungsfragen zu haben. Die Forschenden möchten vor allem datengetriebene Methoden wie z.B. Netzwerkanalysen, die in anderen Bereichen der Medienwissenschaften bereits etabliert sind, auch auf japanische visuelle Medien anwenden können. Dies und die starke Vernetzung der Medien über Franchises, Charaktere, Genres und Themen sowie die daran beteiligten Personen erfordert Datenbanken, die über die reine Auflistung bibliografischer Angaben hinausgehen. Explizit wurde ein Zugang zu den Medien über die in ihnen auftretenden Charaktere und deren Rollen sowie über das Setting oder andere inhaltliche Aspekte erwähnt.

Projektziele und aktueller Stand

Im Rahmen des Projektantrags wurden vier Kernziele formuliert, die sich aus den Vorarbeiten herleiten. Sie sollen nun kurz vorgestellt und der aktuelle Stand des Erreichten zusammengefasst werden.

Vereinbarungen zum Datenaustausch mit Fan-Gemeinschaften

Für eine erfolgreiche und dauerhafte Partnerschaft ist es wichtig, die Motivation und Bedürfnisse der Fan-Gemeinschaften zu verstehen und zu wissen, vor welchen Herausforderungen und Problemen sie stehen. Direkt zu Projektbeginn wurden mehrere Gemeinschaften angeschrieben und Vertreter:innen zu einem gemeinsamen Workshop mit den Projektbeteiligten eingeladen. Schon vor dem Workshop war auffällig, wie unterschiedlich der Zugang zu den erstellten Daten der Gemeinschaften ist: die Bandbreite reicht hier vom aktiven Blockieren von Crawlern und anderen Harvestern über einen freien Zugang bis hin zu ausgefeilten Schnittstellen (APIs). Lizenzinformationen allerdings fehlten teilweise, waren unvollständig oder sogar widersprüchlich. In der gemeinsamen Diskussion zeigte sich, dass es von Seiten der Fan-Gemeinschaften eine große Bereitschaft zum Teilen der eigenen Daten sowohl mit anderen Gemeinschaften als auch Forschenden gibt. Problematisch hingegen wurde die kommerzielle Nutzung der Daten gesehen. Zugleich wurde von mehreren Seiten eine gewisse Unsicherheit in Bezug auf die rechtlichen und lizenztechnischen Rahmenbedingungen geäußert, was die beobachteten Inkonsistenzen in der Lizenzierung erklärt.

Um Daten mit unterschiedlichen Lizenzen in einem gemeinsamen Portal oder Knowledge Graphen anbieten zu können, müssen die Lizenzen zueinander kompatibel sein. Die häufig anzutreffenden Creative-Commons-Lizenzen sind dabei problematisch, sobald die "share-alike" Klausel genutzt wird. So sind Daten mit CC-BY-SA und CC-BY-SA-NC nicht kombinierbar, da die kommerzielle Nutzung nicht gleichzeitig erlaubt und verboten sein kann. Unproblematisch hingegen sind gänzlich freie Lizenzen wie CC-0 oder die alleinige Verwendung der "by" Klausel, die nur

die Nennung der Urheber verlangt. Im Projektkontext wird angestrebt, mit allen Partnern entweder eine Lizenzierung mit einer aktuellen CC-BY-Lizenz oder der Variante CC-BY-SA-NC zu vereinbaren. Alle Daten des Projekts können dann gemeinsam mit der CC-BY-SA-NC Lizenz Dritten zur Verfügung gestellt werden.

Erstellen eines Datenmodells für die Domäne der japanischen visuellen Medien

Ausgehend von den Daten, die von den Gemeinschaften zur Verfügung gestellt werden, werden formale Modelle erstellt, die die jeweiligen Entitäten, ihre Attribute und Beziehungen untereinander abbilden. Diese sind zunächst spezifisch für die jeweiligen Gemeinschaften, haben aber klar identifizierbare Schnittmengen untereinander. In einem zweiten Schritt wird ein gemeinsames Modell erstellt, das die Domäne der japanischen visuellen Medien in ihrer Gänze umfasst. Die Modellierung ist bereits weit fortgeschritten und die formalen Modelle sind als Ontologien in der Web Ontology Language (OWL) zusammen mit einer Beschreibung veröffentlicht worden (Kiryakos & Pfeffer 2021a,b,c).

Aufbau einer zentralen Datenbank

Die OWL-Ontologien werden genutzt, um die Daten der Fan-Gemeinschaften in einzelne Aussagen gemäß dem Resource Description Framework (RDF) zu konvertieren. RDF hat viele Vorteile: Die Aussagen bilden einen Graphen und kommen damit der vernetzten Struktur der Domäne nahe; auch erlaubt die Datenhaltung ein iteratives Vorgehen bei der Integration in das gemeinsame, übergreifende Modell (Kiryakos & Pfeffer 2021d). So können die Daten ohne weitere technische Trennung in einer gemeinsamen Triple-Store-Datenbank gespeichert werden und mittels SPARQL gezielt durchsucht werden. In einem Matching-Schritt werden alle Entitäten identifiziert, die von mehreren Datenquellen beschrieben werden, und geclustert. Die Informationen aus den Clustern können dann in einer "merged Entity" zusammengeführt werden, die im gemeinsamen Modell nur noch an einer Stelle beschrieben sein wird.

Um eine Vorstellung der Größe der Datenbank zu bekommen, sind in Abbildung 1 die Bezeichnungen und Anzahl der Kern-Entitäten für drei Fan-Gemeinschaften zusammengefasst. Werke und Medien werden mit unterschiedlicher Granularität beschrieben, was für den Matching-Schritt eine besondere Herausforderung darstellt. Da jede Entität mit anderen verknüpft und durch weitere Attribute beschrieben wird, ergeben sich eine große Zahl an einzelnen Aussagen. In Summe sind es für die drei Quellen in der Tabelle über 10 Millionen Aussagen.

Fan-Gemeinschaft	Werke und Medien				Firmen	Charaktere	Werk-Eigenschaften	Charakter-Eigenschaften	Beteiligte Personen
ACDB	Work					Character	Work Tag	Character Tag	People
	10207					107369	1088	4051	5557
AnimeClick	Animation Work	Comic Work				Character			Staff
	9491	11762				102143			39604
VNDB		Visual Novel	Release	Producer		Character	Tag	Trait	Staff
		28190	71349	10394		90077	2585	2777	21164

Abb. 1: Struktur und Anzahl der Kern-Entitäten

Der Clustering-Schritt wird zum Zeitpunkt der Einreichung dieser Veröffentlichung durchgeführt und sollte zum Jahresende ab-

geschlossen sein. Problematisch gestaltet sich neben der unterschiedlichen Granularität der Daten zu den Werken auch die Disambiguierung von Personen mit gleichem Namen. Die betroffenen Entitäten machen aber nur einen vergleichsweise kleinen Anteil an den Gesamtdaten aus und werden bei Bedarf manuell bearbeitet.

Zusätzlich zur Bereitstellung der Daten über die SPARQL-Schnittstelle wurde ein Web-Frontend neu entwickelt, das neben den Standardfunktionen für die Anzeige von RDF-Daten auf Basis der Entity-URLs einfach an die Bedürfnisse des Projekts angepasst werden kann. So können die Aussagen einzelner Datenquellen ein- und ausgeblendet werden und die Sprache für die Label der einzelnen Entitäten gewählt werden. Darüber hinaus steht eine schnelle Suchfunktion über einen Elasticsearch-Index zur Verfügung und es können Erweiterungen in Python realisiert werden, die ausgehend vom Frontend und dem aktuell angezeigten Entity-URI Funktionalitäten im Webinterface anbieten. Exemplarisch wurden bereits Korrelationsanalysen, Teilgraph-Exporte und Visualisierungen implementiert.

Die Entwicklung des Frontends ist weit fortgeschritten. Abbildung 2 zeigt die Ansicht eines Spiels vom Typ "Visual Novel". Die Bezeichnungen der Attribute (1) stammen aus der Ontologie, der Link führt zum Eintrag in selbiger. Verknüpfte Entitäten - hier: Charaktere (2), beteiligte Personen (3), Tags (4) - werden durch ihre Label repräsentiert und verlinkt. Für jede einzelne Aussage ist die Quelle (5) in hellgrau angegeben. Die Inhalte der Datenbank sind über das Frontend auf der Website mediagraph.link zugänglich. Die Software für das Frontend wird aktuell noch bearbeitet und wird vor Projektende als Open Source bereitgestellt werden.



Abb. 2: Ansicht einer Visual Novel im Frontend

Evaluation der Datenqualität und der Eignung der Daten für Fragestellungen aus Medienwissenschaften und Japanologie

Um eine klare Vorstellung davon zu bekommen, welche Qualität die Daten der Fan-Gemeinschaften haben, wurden für drei unterschiedliche Datenquellen randomisierte Stichproben gezogen und auf Korrektheit überprüft. Die Größe der Stichproben wurde so gewählt, dass das Ergebnis mit einer Konfidenz von 95% (+/- 5%) auch für die Grundgesamtheit gilt. Als Datenelemente wurden die Titel von Anime-Filmen und Computerspielen von Genre "Visual Novel" in der Originalsprache Japanisch und der englischen Übersetzung ausgewählt. Die Titel sind prominent auf den Covern der Vertriebsmedien ersichtlich und können daher mit

vertretbarem Aufwand auch ohne Inspektion des Mediums selbst geprüft werden. Konkret konnte dafür auf Abbildungen aus Online-Shops und Webseiten der Vertriebsfirmen zurückgegriffen werden. Abbildung 3 fasst die Ergebnisse der Untersuchung zusammen. Die beobachteten Fehler waren überwiegend typografischer Natur und betrafen primär nicht-sintragende Elemente der Titel (5,57%-28,16%): Leerzeichen, Interpunktion, Anführungszeichen und Sonderzeichen wie Sterne, Herzen oder auch Sonderformen von Bindestrichen. Es gab deutlich weniger echte semantische Fehler (0%-2,48%), die meisten davon Hinzufügungen zum Titel, die offenbar der Disambiguierung von gleichnamigen Medien dienen soll.

Datenfeld	Angabe	Anzahl	Anteil	KI untere Grenze	KI obere Grenze
ACDB English title Stichprobe: 424 Auswahlgrundlage: 2435	Korrektur Titel	312	73.585%	68.585%	78.585%
	Typografische Abweichungen	111	26.179%	21.179%	31.179%
	Falsche Angabe	1	0.236%	0.041%	5.236%
ACDB Japanese title Stichprobe: 424 Auswahlgrundlage: 2435	Korrektur Titel	345	81.368%	76.368%	86.368%
	Typografische Abweichungen	77	18.160%	13.160%	23.160%
	Falsche Angabe	2	0.472%	0.082%	5.472%
AnimeClick English title Stichprobe: 483 Auswahlgrundlage: 9167	Korrektur Titel	333	68.944%	63.944%	73.944%
	Typografische Abweichungen	136	28.157%	23.157%	33.157%
	Falsche Angabe	12	2.484%	0.131%	7.484%
	Fehlende Daten	2	0.414%	0.022%	5.414%
AnimeClick Japanese title Stichprobe: 483 Auswahlgrundlage: 9167	Korrektur Titel	367	75.983%	70.983%	80.983%
	Typografische Abweichungen	88	18.219%	13.219%	23.219%
	Falsche Angabe	8	1.656%	0.087%	6.656%
	Fehlende Daten	20	4.141%	0.218%	9.141%
VNDB English title Stichprobe: 503 Auswahlgrundlage: 28170	Korrektur Titel	475	94.433%	89.433%	99.433%
VNDB Original title Stichprobe: 503 Auswahlgrundlage: 28170	Typografische Abweichungen	28	5.567%	0.567%	10.567%
	Korrektur Titel	460	91.451%	86.451%	96.451%
	Typografische Abweichungen	40	7.952%	0.142%	12.952%
	Falsche Angabe	2	0.398%	0.007%	5.398%
	Nicht ermittelbar	1	0.199%	0.004%	5.199%

Abb. 3: Detailangaben zur Qualitätsuntersuchung

Der Aufbau einer Infrastruktur für die Forschung sollte nicht von informationstechnischen Notwendigkeiten, sondern von den Bedürfnissen der Forschenden geleitet werden. Im Rahmen des Projekts werden eine Reihe von kleineren wissenschaftlichen Fragestellungen bearbeitet, die wir als "Tiny Use Cases" bezeichnen (Freybe & Rämisch & Hoffmann 2019). Ausgehend von einer Forschungsfrage werden von Projektmitarbeitern mit einem medienwissenschaftlichen Hintergrund Anforderungen an Daten zur Beantwortung der Frage und Suchstrategien formuliert. Diese prototypischen Nutzungsszenarien leiten die Projektmitarbeiter mit Informatik- bzw. informationswissenschaftlichen Hintergrund bei der Aufbereitung der Daten und dienen als jederzeit testbare Fallstudien für entstehende Prototypen (für ein Beispiel s. Kacsuk 2021). So entsteht eine enge Feedback-Schleife auf den Ebenen der Datenmodellierung, Datenintegration und der Benutzer- und Suchinterfaces. Diese Phase ist im Projekt fast abgeschlossen und der aktuelle Prototyp wird in Kürze externen Forschenden zur Verfügung gestellt, die den Knowledge Graphen für ihre eigenen Forschungsfragen einsetzen und ebenfalls Feedback und Anregungen für die weitere Entwicklung geben können.

Ausblick

Das Projekt "Japanese Visual Media Graph" startete im Juni 2019 und wird für 36 Monate von der Deutschen Forschungsgemeinschaft in der Förderlinie "E-Research Technologies" gefördert. Die aktuellen Arbeiten und Ergebnisse werden im Projekt-Blog dokumentiert (Website: blog.mediagraph.link). Zum Zeitpunkt des Vortrags werden die Arbeiten weitestgehend abgeschlossen sein und der Knowledge Graph als Prototyp der Fachöffentlichkeit zur Verfügung stehen. Bereits jetzt ist deutlich, dass die Daten nicht nur nutzbar sind, sondern auf breites Interesse in der Forschungsgemeinschaft stoßen. Durch den Ausbau entsprechender Kollaborationen hoffen wir, die Datenbasis und ihre Anwendungsszenarien sukzessive erweitern und konsolidieren zu können.

Bibliographie

Condry, Ian (2013): *The soul of anime*. Collaborative creativity and Japan's media success story. Durham: Duke University Press.

Freybe, Konstantin / Rämisch, Florian / Hoffmann, Tracy (2019): "With small steps to the big picture: A method and tool negotiation workflow.", in: Steven Krauwer / Darja Fišer (eds.): *Proceedings of the Twin Talks Workshop at DHN 2019*, co-located with Digital Humanities in the Nordic Countries (DHN 2019). Aachen: CEUR-WS.org 13-24 http://ceur-ws.org/Vol-2365/03-TwinTalks-DHN2019_paper_3.pdf [letzter Zugriff 15. Juli 2021].

Kacsuk, Zoltan (2021): "Using fan compiled metadata for anime, manga and video game studies research: Revisiting Hiroki Azuma's 'Otaku: Japan's Database Animals' twenty years on", in: Roth, Martin / Picard, Martin / Yoshida, Hiroshi (eds.): *Japan's Media between Local and Global. Current Perspectives on Regionality, Representation, Culture and Technology*. Heidelberg: CrossAsia-eBooks, Universitätsbibliothek Heidelberg (erscheint in Kürze).

Kiryakos, Senan / Pfeffer, Magnus (2021a): *Japanese Visual Media Graph - Visual Novel Database Ontology*. Zenodo 10.5281/zenodo.5036040.

Kiryakos, Senan / Pfeffer, Magnus (2021b): *Japanese Visual Media Graph - Anime Characters Database Ontology*. Zenodo 10.5281/zenodo.5710959.

Kiryakos, Senan / Pfeffer, Magnus (2021c): *Japanese Visual Media Graph - AnimeClick Ontology*. Zenodo 10.5281/zenodo.5508683.

Kiryakos, Senan / Pfeffer, Magnus (2021d): "The Benefits of RDF and External Ontologies for Heterogeneous Data: A Case Study Using the Japanese Visual Media Graph", in: Schmidt, Thomas / Wolff, Cristian (eds): *Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)*. Regensburg, Germany, 8th-10th March 2021. Glückstadt: Verlag Werner Hülsbusch 308-320 10.5283/epub.44950.

McGray, Douglas (2002): "Japan's Gross National Cool", in: *Foreign Policy* 130: 44-54 10.2307/3183487.

Nozawa, Shunsuke (2013): "Characterization", in: *Semiotic Review* 3 <https://www.semioticreview.com/ojs/index.php/sr/article/view/16> [letzter Zugriff 15. Juli 2021].

Oyama, Shinji (2016): "Japanese creative industries in globalization", in: Hjorth, Larissa / Khoo, Olivia (eds): *Routledge Handbook of New Media in Asia*. Abingdon Oxon UK: Routledge 322-332 10.4324/9781315774626.

Picard, Martin / Pelletier-Gagnon, J  r  mie (2015): "Introduction: Geemu, media mix, and the state of Japanese video game studies", in: *Kinephanos. Journal of media studies and popular culture* 5: 1-19 <https://www.kinephanos.ca/2015/introduction-geemu-media-mix-en/> [letzter Zugriff 15. Juli 2021].

Price, Ludi / Robinson, Lyn (2017): "'Being in a knowledge space': Information behaviour of cult media fan communities", in: *Journal of Information Science* 43(5): 649-664 10.1177/0165551516658821.

Steinberg, Marc (2012): *Anime's media mix. Franchising toys and characters in Japan*. Minneapolis: University of Minnesota Press.

Valaskivi, Katja (2013): "A brand new future? Cool Japan and the social imaginary of the branded nation", in: *Japan Forum* 25(4): 485-504 10.1080/09555803.2012.756538.

Jung, wild, emotional? Rollen und Emotionen Jugendlicher in zeitgen  ssischer Fantasy- Literatur

Fl  h, Marie

marie.flueh@uni-hamburg.de
Universit  t Hamburg, Germany

Schumacher, Mareike

schumacher@linglit.tu-darmstadt.de
Technische Universit  t Darmstadt

Im literatur- und kulturwissenschaftlich ausgerichteten Projekt *m*w* untersuchen wir den Zusammenhang von Genderaspekten und Emotionen in deutschsprachigen literarischen Texten (Schumacher und Fl  h 2020). Um bestehende Einzelfallstudien und Erkenntnisse   ber Genderrollen (vgl. Beauvoir 2018, Bourdieu 2010, Butler 2016, Connell 1996, 2015 Heilman 2003) und Emotionen (vgl. Winko 2003, 2020) in einen gr   eren Kontext setzen zu k  nnen, fragen wir nach Gelingensbedingungen und konkreten Umsetzungsm  glichkeiten einer reflektierten, digitalen Textanalyse. Gleichzeitig geht es darum, die Interdependenz zwischen Gender und Emotionen auf der einen Seite und Textsorte, Genre und zeitlichem Kontext auf der anderen Seite herauszustellen.

Rollen und Emotionen jugendlicher Hauptfiguren in zeitgen  ssischer Fantasy-Literatur

Dieser Beitrag zielt darauf ab, das f  r phantastische Literatur als genrekonstitutiv geltende, aber recht allgemein beschriebene "Klima des Grauens" (Caillois 1974: 56), das Unheimliche (Todorov 2018), ausdifferenzieren. Dar  ber hinaus betrachten wir die f  r phantastische Literatur als spezifisch herausgestellten Emotionstypen in Abh  ngigkeit zu den Genderrollen, die den Emotionsender:innen zugeschrieben werden. Dabei r  cken wir die – ebenfalls f  r das betrachtete Genre typischen – jugendlichen Hauptfiguren in den Fokus. Der Beitrag widmet sich drei eng

miteinander verkn  pften Forschungsfragen: Welche Genderrollen werden den Protagonist:innen in Fantasyromanen zugeschrieben? Welche Emotionstypen bestimmen das Korpus und wie sehen die Rollen- und Emotionsprofile der Protagonist:innen aus? Fallen genderstereotype, statische Muster auf?

Abgeschlossene Vorarbeiten: Annotation von Genderrollen und Emotionstypen im Korpus

Unser Beitrag kn  pft direkt an eine allgemeinere Fallstudie zum Thema Genderrollen in zeitgen  ssischer Jugend-Fantasy-Literatur an (Fl  h, Horstmann und Schumacher, im Erscheinen) und vertieft die gewonnenen Einsichten. Grundlage der bereits abgeschlossenen Fallstudie stellen 28 deutschsprachige kontempor  re (im Zeitraum zwischen 2015 und 2020 publizierte) Fantasy-Romane f  r Jugendliche dar. Hierbei haben wir im Rahmen eines Mixed-Methods-Ansatzes Emotionen in Abh  ngigkeit zu Genderkategorien (m  nnlich, weiblich, neutral) betrachtet, indem ein eigens trainierter Gender-Classifer (Schumacher 2021), der auf Conditional-Random-Fields-Algorithmen (vgl. Sutton und McCallum 2010) basiert und mit dem Stanford Named Entity Recognizer (vgl. Finkel et al. 2005) kompatibel ist, zur automatischen Annotation von Genderrollen mit der digitalen manuellen Annotation von Emotionsinformationen kombiniert wurde.

Das Korpus wurde zun  chst mit dem Gender-Classifer annotiert, der in allen Texten m  nnliche, weibliche und neutrale Figurenrollen markiert. Die Erkennungsgenauigkeit der genutzten Version erreicht   ber alle Kategorien hinweg bei gattungsspezifischem Testmaterial einen F1-Score von rund 72% (vgl. Schumacher 2021). Auf die automatische Vorannotation aufbauend, wurde in 25 Romanen mit dem Textanalysetool CATMA (Gius et al. 2021) eine taxonomiebasierte Emotionsanalyse durchgef  hrt. Die digitale manuelle Annotation funktioniert auf Grundlage eines f  r die Emotionsanalyse in literarischen Texten entworfenen Tagsets und hierf  r entworfenen Guidelines (vgl. Fl  h 2020). Die Emotionsanalyse bezog sich auf die Textstellen, an denen vom Gender-Classifer besonders zahlreiche Genderannotationen gemacht wurden. An diesen Gender-Peaks wurde jeweils das semantische Umfeld der Gender-Annotationen nach Emotionsinformationen untersucht. Wir fokussieren also das unmittelbare Textumfeld der Figurenreferenzen und bestimmen, ob und welche emotionstragenden Textstrukturen zu finden sind.

Emotionen der Jugend – Ausgestaltung von Emotionsprofilen junger Protagonist:innen

Unabh  ngig von Genderrollen zeigt sich, dass im Korpus Basisemotionen mit negativer Qualit  t roman  bergreifend die Erz  hlwelten bestimmen. Im gesamten Korpus etabliert sich ein Emotionsprofil, das sich mit abnehmender H  ufigkeit zusammensetzt aus:

1. Angst (925 Annotationen)
2. Zorn (388 Annotationen),
3. Freude/Gl  ck (149 Annotationen)
4. Liebe (219) Annotationen)

5. Trauer (198 Annotationen)
6. Ekel (123 Annotationen)

Ein Klima des Grauens etabliert sich deutlich über die besonders häufig vorkommenden Angst-Emotionen. Ein genauer Blick auf die Vertreter dieser Kategorie zeigt ein facettenreiches Emotionsprofil, das unterschiedliche Angstzustände und Spielarten der Angst beinhaltet (s. Tabelle 1).

Tab. 1: Die unterschiedlichen Formen der Angst im Überblick

Emotionstyp	Vertreter der Kategorie	Anzahl der Annotationen
Angst	Besorgnis	383
	Schrecken	148
	Panik	123
	Nervosität	90
	Entsetzen	73
	Bestürzung	39
	Zaghafteigkeit	29
	Gruseln	20
	Grauen	20

Die Annotation der Romane zeigt darüber hinaus, dass gerade die beiden bewusst bedeutungs offen gestalteten Annotationskategorien “UNCATEGORIZABLE” (797 Annotationen) und “PROBLEMFÄLLE” (813 Annotationen) quantitativ ins Gewicht fallen. Während in die Kategorie “Problemfälle” Emotionstypen fallen, die u.a. mehrere polare Gegensätze vereinen, also nicht eindeutig als positiv oder negativ kategorisiert werden können, versammelt die Oberkategorie “Uncategorizable” Emotionen, die nicht einer der Basisemotionen zugeordnet werden können, die aber in ihrer Polarität durchaus eindeutig sind. Viele Textpassagen lassen sich auf Grundlage der strukturorientierten Typologisierung, die sich an der Einteilung von Basisemotionen orientiert (Ekman 1972, Schwarz-Friesel 2007), nicht adäquat beschreiben. In diesen Fällen wurden weitere Emotionstypen definiert; besonders ins Gewicht fallen dabei:

1. Erstaunen (414)
2. Bedauern (179)
3. Interesse (102)
4. Scham (77)
5. Aggression (35)

Die häufigsten Vertreter dieser Kategorien nuancieren das Emotionsprofil. Scham, Aggression und Bedauern weisen eine negative Qualität auf, während Interesse und Erstaunen eher in die Kategorie positiver Emotionstypen fallen. Auffällig ist, dass alle hier vertretenen Emotionstypen im Diskurs über Sprache und Emotionen als strittige Emotionstypen verhandelt werden. Scham, eine eher intrasubjektiv empfundene Emotion, wird häufig mit ähnlichen Emotionskategorien wie Schuld, Reue oder Bedauern in Verbindung gebracht. Unklar ist hierbei, ob Scham eine eigene Kategorie darstellt oder nicht. Deutlich explosiver und als Selbst- oder Fremdschädigung auftretende Aggressionen lassen sich als Trieb oder als Emotion beschreiben. Fraglich ist auch, ob Erstaunen und Interesse als Emotion klassifiziert werden sollen (Schwarz-Friesel 2007). Hier offenbart sich ein facettenreiches negatives Emotionsprofil, das neben recht eindeutig bestimmbar negativen Basisemotionen auch eher nach innen gerichtete und Verbundemotionen wie Scham beinhaltet.

Da im Annotationsprozess für jede Emotionsannotation die Values männlich, weiblich oder neutral festgelegt wurden, lässt sich nachvollziehen, welche Emotionen mit männlichen Figuren und welche mit weiblichen Figuren in Verbindung stehen. Innerhalb der untersuchten Textpassagen konnten weiblichen Figuren 2200

Mal eine emotionale Reaktion zugeordnet werden, männliche Figuren lediglich 1474 Mal. Es zeigt sich, dass Angst im untersuchten Korpus die zentrale Emotion für beide Gender darstellt. Charaktere beider Geschlechter bilden ein negativ geprägtes Emotionsprofil aus, das unterschiedliche Angstzustände beschreibt: Besorgnis, Erschrecken und Panik bestimmen das Korpus.

Nach diesem übergeordneten Blick auf das Korpus, stellt sich nun die Frage, welche Genderrollen für die jungen Protagonist:innen besonders häufig sind, ob diese spezifisch für einzelne Erzähltexte sind, oder ob sich im gesamten Korpus romanübergreifende Muster bilden.

Genderrollen der Jugend – Ausgestaltung der Rollenprofile junger Protagonist:innen

Um herauszufinden, welche Genderrollen für die jungen Protagonist:innen besonders häufig sind, ob diese spezifisch für einzelne Erzähltexte sind, oder ob sich im gesamten Korpus romanübergreifende Muster bilden, haben wir für alle Protagonist:innen in einer relationalen Graphdatenbank mithilfe der Webapplikation Graphcommons (vgl. Arian et al., o.J.) Rollenprofile angelegt.¹ Genderrollen definieren wir nach einem theoretischen Modell, in dem verschiedene in Gender und Masculinity Studies skizzierte Geschlechterrollen in übergeordnete Kategorien zusammengefasst wurden (vgl. Schumacher und Flüh 2020). Rollen wie “Knabe”, “Kind” oder “Mutter” können einer dieser Genderkategorien zugeordnet werden (Mann, Genderneutral, Frau), sind also Genderrollen.²

Die vom Classifier annotierten Genderrollen wurden durch eine Kollokationsanalyse mit den Protagonist:innen in Verbindung gebracht. Dabei haben wir nach den Gendertags gesucht, in deren Wortumfeld (fünf Wörter davor und danach) der Name der Hauptfigur steht. Anschließend wurde der Annotationskontext daraufhin überprüft, ob mit der annotierten Genderrolle die Hauptfigur bezeichnet wird. Wenn dies der Fall war, wurde die Rolle im Graph mit der Figur verknüpft. Da unser Korpus auch Romanreihen beinhaltet und einige der Hauptfiguren in mehr als einem Roman als Protagonist:in auftreten – was die Vermutung nahelegt, dass die Protagonist:innen von Reihen aufgrund der größeren Textmenge auch mit mehr Rollen bezeichnet werden könnten – wurden die Texte ebenfalls als Knoten im Graphen angelegt und die Hauptfiguren mit den jeweiligen Texten verknüpft. Auf diese Weise entstand im ersten Schritt ein komplexer Graph mit drei Ebenen von Knoten: Texte, Hauptfiguren und Rollen (vgl. Abb. 1). In einem weiteren Schritt wurden Emotionen als vierter Knotentyp hinzugefügt.



Abb. 1: Graph der Texte, Hauptfiguren und Rollen im Fantasy-Jugendroman-Korpus

Betrachtet man die Rollenstruktur des Fantasy-Korpus, fallen mehrere Eigenheiten auf. Zunächst einmal beinhalten die Romane mehr Protagonistinnen als Protagonisten. Elf weibliche Hauptfiguren der Romane und Romanzyklen stehen vier männlichen gegenüber. Zwei der vier männlichen Protagonisten fallen durch besonders vielfältige Rollenprofile auf, während die weiblichen Hauptfiguren insgesamt weniger vielfältige Rollenprofile ausbilden. Die einzelnen Rollen sind dabei häufig stereotyp angelegt, männlichen Protagonisten werden hauptsächlich männlich stereotype Genderrollen zugeschrieben und Protagonistinnen hauptsächlich weibliche. In seltenen Fällen weisen die Protagonist:innen zusätzlich zu ihren binär gegenderten Rollen auch genderneutrale Rollen auf wie „Kind“, „Findelkind“, „Mensch“ und „guter Mensch“.

Die meisten Hauptfiguren bilden im Rollennetzwerk Cluster³ aus, wobei die Rollen häufig text- oder reihenspezifisch sind. Ein Beispiel für eine sehr romanspezifische Rolle ist die „Traumhändlerin“ in *Die Stadt der gläsernen Träume* und auch die Magierinnen-Rollen von Robin aus *Burning Magic I-III* sind sehr an die erzählte Welt dieser Trilogie gebunden. Die Rolle der Tochter bildet dagegen ein eigenes Cluster (vgl. Abb. 2), was zeigt, dass die familiäre Struktur im gesamten Korpus von herausragender Bedeutung ist.

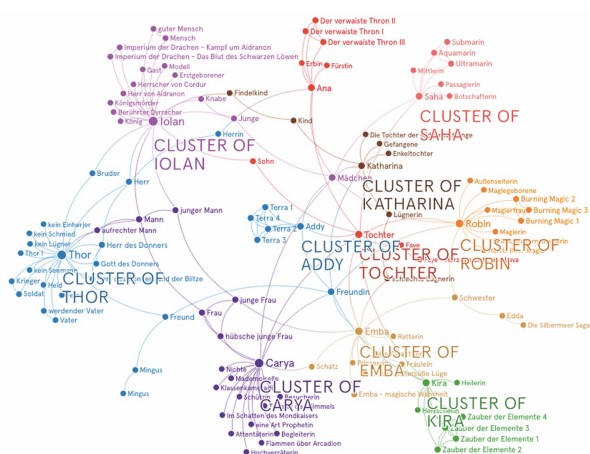


Abb. 2: Rollennetzwerk geclustert

Darüber hinaus sind die Genderrollen der „Freundin“ und des „Mädchens“ romanübergreifend bedeutsam; sie sind jeweils mit sieben Protagonistinnen verknüpft. Die Hauptfiguren in dieser Stichprobe deutschsprachiger Jugend-Fantasy-Romane sind also meist Mädchen, für die einerseits das familiäre und andererseits das freundschaftliche Umfeld eine wichtige Bedeutung haben.

Zwei der Hauptfiguren fallen dadurch auf, dass sie nur mit einer einzigen Rolle verbunden sind. Es handelt sich dabei um Edda aus der *Silbermeer Saga* und Mingus aus dem gleichnamigen Roman. Tatsächlich werden Edda und ihr Bruder als Findelkinder vorgestellt, deren familiäre Herkunft unbekannt ist und deren wichtigste Verbindung die zueinander ist. Mingus ist ein künstlich erzeugtes Wesen, eine Chimäre aus Mensch und Tier. Es verwundert darum kaum, dass hier die (stereotypen) Genderrollen, die der Classifier erkennt, selten sind. Die einzige Genderrolle, mit der Mingus referenziert wird, ist die des Freundes. Dieser Befund ist für Anschlussuntersuchungen besonders interessant. Da der Gender-Classifier stereotype Genderrollen am besten erkennt, könnte es sein, dass den Protagonist:innen dieser Texte zwar noch mehr Rollen zugeschrieben werden, dass diese aber die hier betrachteten Genderkategorien sprengen und nicht in das zugrundeliegende Kategoriensystem passen. Eine genauere Betrachtung liegt zwar außerhalb des Fokus dieses Beitrags, könnte aber im Rahmen einer Close-Reading-Analyse darauf aufbauen. Die Analyse des Rollennetzwerkes zeigt also insgesamt bereits erste Hinweise auf die Ausgestaltung der Hauptfiguren unseres Korpus. Ein genaueres Bild ergibt sich, wenn die Emotionen, die den Protagonist:innen zugeschrieben werden, mit einbezogen werden.

Zusammenführung von Genderrollen und Emotionstypen

Um zu erproben, wie das Zusammenspiel von Genderrollen und Emotionen gemeinsam betrachtet werden kann, haben wir zunächst Emotionen in einem kompletten Roman manuell annotiert. Dabei handelt es sich um einen Text mit einer Protagonistin – Ana aus der Trilogie *Der verwaiste Thron* von Claudia Kern (Abb. 3).

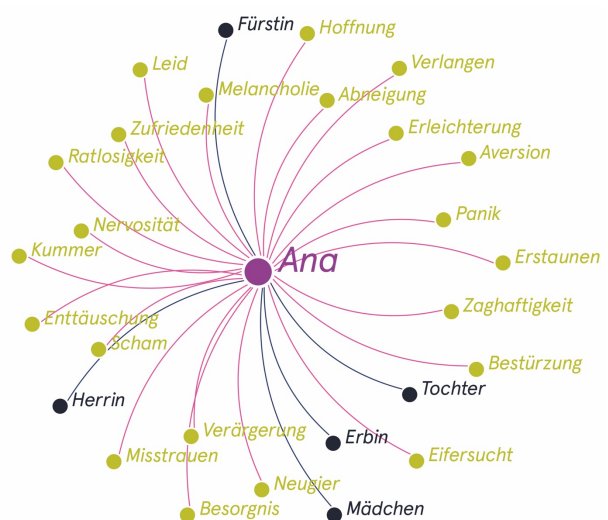


Abb. 3: Beispielhafte Close-Reading-Auswertung der Emotions- und Rollenprofile von Thor und Ana

Ana ist die Heldin einer Trilogie (manuell annotiert wurde allerdings nur der erste Teil) und wird stark über Emotionen charakterisiert. Fünf Rollen stehen hier vierzig Emotionen gegenüber. Zwar wurden die Genderrollen automatisch und die Emotionen manuell annotiert, dennoch ist die Differenz hier signifikant. Im gesamten Korpus wurden 68 Emotionen und 83 Genderrollen ausgemacht, die sich auf Protagonist:innen beziehen. Zwar erreicht der Classifier nur eine Erkennungsgenauigkeit von 72% und annotiert somit höchstwahrscheinlich nicht alle Genderrollen, die mit einer Figur verbunden sind. Auf der anderen Seite handelt es sich aber um eine kontextsensitive automatische Annotation, d.h. die Anzahl der Genderrollen, die erkannt werden kann, ist potentiell unendlich. Die Protagonistin tritt als "Herrin", "Erbin" und "Fürstin" in Erscheinung, gleichzeitig nimmt sie auch die im Korpus insgesamt sehr bedeutsame familiäre Rolle der Tochter ein. Dass Ana eine Protagonistin ist, die vergleichsweise stark über Emotionen charakterisiert wird, zeigt eine Gegenüberstellung mit den anderen Hauptfiguren im Korpus (vgl. Abb. 4). Die in Abb. 4 in Pink dargestellten Emotionsrelationen überwiegen die dunkelblauen Genderrollenzuschreibungen deutlich; stärker als dies bei anderen Protagonist:innen der Fall ist.

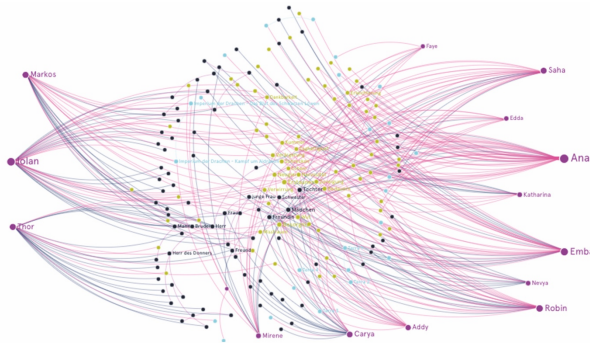


Abb. 4: Gegenüberstellung männlicher und weiblicher Hauptfiguren mit Emotionsrelationen und Genderzuschreibungen

Neben unterschiedlichen positiven Emotionen (Erheiterung, Dankbarkeit, Vertrauen, Zufriedenheit, Zuneigung, Verlangen, Zuversicht und Gelassenheit) steht eine deutlich höhere Anzahl negativer Emotionen (Abneigung, Trübsal, Leid, Schrecken, Kummer, Nervosität, Wut, Bedauern, Bestürzung, Verzweiflung, Widerwille, Scham, Panik, Ratlosigkeit, Aversion, Entsetzen, Verärgerung und Melancholie), die diese Figur charakterisieren und dem negativen genrespezifischen Emotionsprofil entsprechen.

Um zu prüfen, ob es sich hierbei um ein genderspezifisches Muster handelt, haben wir abschließend die anderen Romane im Korpus betrachtet. Um die größere Stichprobe analysieren zu können, haben wir die Romane nicht im Close-Reading-Verfahren annotiert, sondern lediglich Gender-Peaks – Passagen, in denen der Gender-Classifier besonders viele Annotationen hinzugefügt hat – und ein Fenster von sechs Sätzen pro Genderzuschreibung betrachtet (drei vor der Erwähnung einer Genderrolle innerhalb eines Peak-Abschnitts und drei danach). Um anschließend diejenigen Emotionsannotationen ausfindig zu machen, die Emotionen markieren, die den Protagonist:innen zugeschrieben wurden, haben wir erneut Kollokationsabfragen durchgeführt, die ausschließlich Emotionsannotationen im Wortumfeld von namentlichen Erwähnungen der Hauptfigur aufzeigen.

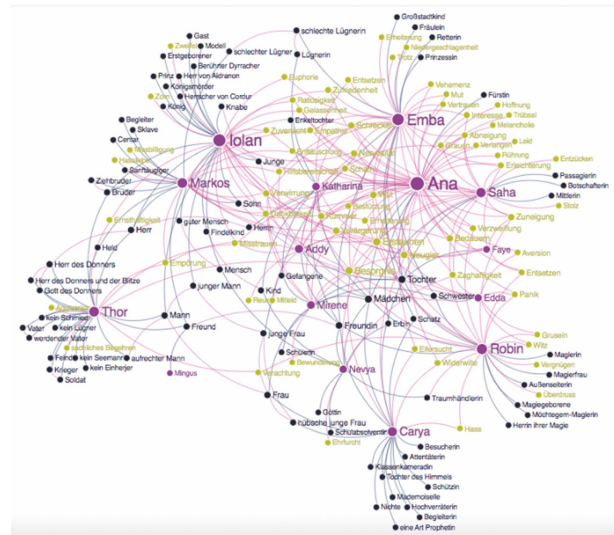


Abb. 5: Rollen und Emotionsprofile von Hauptfiguren in Jugend-Fantasy-Literatur

Der Graph in Abb. 5 zeigt ein interessantes Bild: Nur eine männliche und zwei weibliche Hauptfiguren werden stärker über Rollen als über Emotionen charakterisiert. Bei drei Hauptfiguren sind die Rollen- und Emotionsprofile relativ ausgeglichen. Sechs Protagonistinnen zeigen ein eindeutig stärker über Emotionen als über Rollen definiertes Profil. Es zeigt sich also eine leichte Tendenz zu einer stärkeren Rollenprofilierung der männlichen Hauptfiguren. Die Mehrzahl der Protagonistinnen wird stärker über Emotionen als über Rollen ausgestaltet. An dieser Stelle sind zwei methodenkritische Aspekte zu berücksichtigen: Erstens findet der hier vorgestellte Classifier nicht alle relevanten Entitäten; hier zeigt sich eine grundsätzliche Schwäche von NLP-Ansätzen zur Analyse von (literarischen) Texten. Zweitens führt der vorgestellte Mixed-Methods-Ansatz automatische, d.h. weniger zuverlässige, Annotationen mit manuellen zusammen. Da der Gender-Classifier mit einer Quote von 72% F1-Score nicht alle Vorkommnisse von Genderrollen annotiert, ist es möglich, dass die Figuren eigentlich mehr Genderrollen zugeschrieben bekommen, als hier gezeigt. Die Kontextsensitivität des Tools gewährleistet allerdings, dass sehr viele unterschiedliche Genderrollen automatisch annotiert werden, davon aber nicht immer unbedingt alle Vorkommnisse. Wir gehen darum davon aus, dass die Mehrheit der vorhandenen Genderrollen berücksichtigt werden konnte und der Gender-Classifier eine valide Tendenz der Verteilung aufzeigt. Auch ist bei der hier vorliegenden vergleichenden Analyse vor allem die Balance zu den anderen Texten bedeutsam.

Interessant ist auch, welche Emotionen mit den meisten weiblichen Hauptfiguren verknüpft sind. Erstaunen ist mit allen acht Protagonistinnen der Stichprobe verbunden. Besorgnis empfinden sechs der acht weiblichen Hauptfiguren. Bedauern ist bei fünf Charakteren zu finden und Neugier und Zuneigung bei jeweils vier von ihnen. Von diesen fünf am meisten verknüpften Emotionen ist nur eine der Basisemotion der Angst zuzuordnen, nämlich die Besorgnis. Dabei handelt es sich allerdings um eine relativ schwache Form von Angst. Das Klima des Grauens, das für das Genre der Phantastischen Literatur postuliert wurde und das die Gesamtbetrachtung der Emotionen in unserem Korpus bestätigen konnte, geht also nicht wesentlich von den überwiegend weiblichen Hauptfiguren aus. Diese wirken dem eher besorgt, mitfühlend und auch neugierig entgegen.

Ausblick

Bis hierhin erweist sich das geschilderte Verfahren als sinnvoll, um das Zusammenspiel von Emotionen und Genderrollen zu analysieren. Es eignet sich, um häufig vorkommende Genderrollen und Emotionsprofile zu ermitteln. Genderstereotype Muster zeichnen sich zwar in dieser Fallstudie schon ab, müssten aber durch die Analyse eines größeren Korpus noch bestätigt oder revidiert werden. Der beispielhafte Vergleich mit einem ebenfalls zunächst im Close-Reading betrachteten männlichen Hauptcharakter steht noch aus. Die Betrachtung des Gesamtkorpus gibt einen vorläufigen Hinweis darauf, dass männliche Hauptfiguren etwas weniger stark durch Emotionen und eher durch stereotype Rollenbilder charakterisiert werden. Um diese Tendenz weiter zu untersuchen, müsste allerdings in einer Anschlussstudie die Stichprobe erweitert werden, um mehr Protagonisten in die Untersuchung einbeziehen zu können.

Fußnoten

1. Der Graph kann hier eingesehen, durchsucht und analysiert werden: <https://graphcommons.com/graphs/6f285a83-15eb-4ed-d-830c-cd26f48b493b>.
2. Zusätzlich zur Automatisierung der Erkennung solcher eher stereotyper Genderrollen ist es ein Desiderat im Projekt m*w auch Genderrollen einzubeziehen, die weniger stereotyp sind und sich darum der Klassifizierung in eine dieser drei Kategorien entziehen. Ein entscheidender Schritt in Richtung dieses Desiderats ist es aber erst einmal erfassen zu können, wie stereotype Genderzuschreibungen in literarischen Texten eigentlich beschaffen sind.
3. Zum Clustering wird die in Graph Commons implementierte Louvain-Modularity-Methodik genutzt, die Knoten mit hoher Relationsdichte in den Fokus rückt (vgl. Graph Commons 2017).

Bibliographie

- Beauvoir, Simone de** (2018): *Das Andere Geschlecht*. Reinbek: Rowohlt.
- Arikan, Burak / Üstün, Zeyno / Kızılay, Ahmed / Badur, Aybars / Erikli, Fatih / Züncül, Özlem / Gilikoglu, Dara / Aldatmaz, Ayca / Dölec, Genk**: Graph Commons (o.J.): *Graph Commons*. URL: <https://graphcommons.com> [letzter Zugriff 29. November 2021].
- Bourdieu, Pierre** (2010): *Die Männliche Herrschaft*. Frankfurt am Main: Suhrkamp.
- Butler, Judith** (2016): *Das Unbehagen Der Geschlechter*. Frankfurt am Main: Suhrkamp.
- Connell, Raewyn** (1996). *Gender and Power*. Cambridge: Polity Press.
- Connell, Raewyn** (2015). *Der Gemachte Mann*. Wiesbaden: Springer.
- Ekman, Paul** (1972): "Universals and cultural differences in facial expression of emotion" in: Cole, J.K. (ed.): *Nebraska Symposium on Motivation*. Lincoln: University of Nebraska Press 207–283.
- Finkel, Jenny Rose / Grenager, Trond / Manning, Christopher** (2005): "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 363–70.
- Flüh, Marie** (2020, § 10): "Emotionsanalyse" in: Gius, Evelyn / Meister, Jan Christoph / Schumacher, Mareike / Gerstorfer, Dominik / Meister, Malte / Bläß, Sandra / Flüh, Marie / Horstmann, Jan / Jacke, Janina (eds.): *ForTEXT - Literatur digital erforschen*. <https://fortext.net/ressourcen/tagsets/emotionsanalyse> [letzter Zugriff: 13. Juli 2021].
- Flüh, Marie / Horstmann, Jan / Schumacher, Mareike** (im Erscheinen): "Distant Gender Reading. Genderaspekte in Fantasy-Jugendromanen von 2008 bis 2020" in: Weertje v. Willms (ed.): *Genderaspekte in der Kinder- und Jugendliteratur vom Mittelalter bis zur Gegenwart. Diachrone und synchrone Perspektiven*. Berlin: De Gruyter.
- Gius, Evelyn / Meister, Jan Christoph / Meister, Malte / Petris, Marco / Bruck, Christian / Jacke, Janina / Schumacher, Mareike / Gerstorfer, Dominik / Flüh, Marie / Horstmann, Jan** (2021): *CATMA 6 (Version 6.3)*. Zenodo. DOI: 10.5281/zenodo.1470118.
- Graph Commons** (2017): "Finding organic clusters in complex data-networks". In: *Graph Commons*. URL: <https://medium.com/graph-commons/finding-organic-clusters-in-your-complex-data-networks-5c27e1d4645d> [letzter Zugriff: 29. November 2021].
- Heilman, Elisabeth E.** (2003): "Blue Wizards and Pink Witches: Representations of Gender Identity and Power" in: dies. (eds.): *Critical Perspectives on Harry Potter*. New York: Routledge, 221–241.
- Schumacher, Mareike / Flüh, Marie** (2020): "Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen: Genderstereotypen und -bewertungen in der Literatur des 19. Jahrhunderts" in: Schöch, Christof (ed.): *DHd2020: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts* 162–167. URL: <https://zenodo.org/record/3666690#.X37-FFICTus> [letzter Zugriff 13. Juli 2021].
- Schumacher, Mareike** (2021): *StanfordNER Gender-Classifer (Version 0.1)*. Zenodo. <http://doi.org/10.5281/zenodo.3667462>.
- Schwarz-Friesel, Monika** (2007): *Sprache und Emotion*. Tübingen: Narr Francke.
- Sutton, Charles, / Andrew McCallum** (2010): *An Introduction to Conditional Random Fields*. ArXiv:1011.4088 [Stat], November 17, 2010. <http://arxiv.org/abs/1011.4088>.
- Todorov, Tzvetan** (2018): *Einführung in die fantastische Literatur*. Berlin: Wagenbach.
- Winko, Simone** (2003): *Kodierte Gefühle. Zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900*. Berlin: Erich Schmidt.
- Winko, Simone** (2020): „Literaturwissenschaftliche Emotionsforschung“ in: Kappelhoff, Hermann / Bakels, Jan-Hendrik / Lehmann, Hauke / Schmitt, Christina (eds.): *Emotionen. Ein interdisziplinäres Handbuch*. Stuttgart: Metzler 397–407.

Kategorientheoretische Ontologieentwicklung und Wissensmodellierung für die Digital Humanities

Gerstorfer, Dominik

gerstorfer@linglit.tu-darmstadt.de
TU Darmstadt, Germany

Einleitung

Formale Modellierung von Wissen spielt in den Digital Humanities eine nicht zu unterschätzende Rolle. Daten müssen organisiert, kategorisiert, gespeichert, abgerufen und verarbeitet werden. Hierzu werden häufig Datenbank- und Ontologiesprachen wie SQL und RDF/OWL oder auch semantische Netzwerke eingesetzt, die als *informatische* Systeme robust und etabliert sind. Ihre Verwendung setzt zumeist nach der eigentlichen geisteswissenschaftlichen Arbeit an, d.h. erst, wenn die konzeptuelle Arbeit abgeschlossen ist, werden die so gewonnenen Ergebnisse formalisiert. Dieses Vorgehen birgt das Risiko, dass sich die, im ersten Arbeitsschritt gewonnenen Ergebnisse nicht ohne Weiteres formalisieren lassen. Die Gründe hierfür sind vielfältig und reichen von unzureichender Rigidität der informellen Analyse bis zur mangelnden Expressivität des verwendeten Formalismus. Hinzu kommt, dass bestehende Formalismen die kreative Beschäftigung mit den Inhalten erschweren und die Beherrschung der technischen Systeme erfordert häufig eine steile Lernkurve.¹

In diesem Beitrag werden *ontology logs* (*ologs*) als Framework für konzeptuelle Arbeit und Wissensmodellierung in den Digital Humanities vorgestellt. *Ologs* sind im Bereich der angewandten mathematischen Kategorientheorie als Framework entwickelt worden, mit dem Ziel ein gleichermaßen mächtiges wie benutzerfreundliches Modell zur Wissensrepräsentation bereitzustellen, das in der Lage ist, die geisteswissenschaftliche Forschung zu unterstützen und leichter an mathematisch-informatische Theorien und Techniken anschlussfähig zu machen.

Kategorientheorie

Die mathematische Kategorientheorie wurde in den 1940er-Jahren von Saunders Mac Lane und Samuel Eilenberg entwickelt, um verschiedene mathematische Felder und Theorien zu vergleichen. Die grundlegende Einsicht hinter der Kategorientheorie formuliert Mac Lane (1998: 1) wie folgt: “Category theory starts with the observation that many properties of mathematical systems can be unified and simplified by a presentation with diagrams of arrows.”

Eine Kategorie C besteht aus einer Sammlung von Objekten $Ob(C)$ und Pfeilen (\rightarrow) zwischen Objektpaaren $A, B \in Ob(C)$, sodass $f: A \rightarrow B$. Des Weiteren existieren Pfeile zwischen zwei Kategorien C und D , sodass $C \rightarrow D$ eine strukturerhaltende Abbildung ist, die Funktor genannt wird. Diese einfache Systematik und die Möglichkeit, Pfeile zu kombinieren eröffnen weitgehende Abstraktionsmöglichkeiten, die so weit gehen, große Teile der Mathe-

matik (Algebra, Topologie, Mengentheorie, Grapfentheorie, Gruppentheorie, etc.) zu fundieren.

In der Regel wird die Kategorientheorie in der Mathematik verwendet, um Theorien zu analysieren und weiter zu verallgemeinern. Man könnte auch davon sprechen, dass es sich um Metamathematik handelt, die Mathematik als Anwendungsgegenstand hat.

“Eilenberg and Mac Lane introduced very abstract tools into mathematics, which seemed even too abstract. Nevertheless, they motivated their work with both technical merits, which allow for an effective study of the phenomenon of naturality, and conceptual advantages. They noted that the proposed conception is so general that it allows for the detection of the same structures in fundamentally different fields of mathematics. By finding new analogies between different fields of mathematics it suggests new results. Thanks to the fact that categorical glasses allow for the observation of the same structures in both topology and algebra, these glasses allow for a unifying view of mathematics. Already in 1945 it was clear that CT had the power to unify mathematics.” (Landry 2017: 3)

Und obwohl die Kategorientheorie als *allgemeine Theorie mathematischer Strukturen* ihren Anfang nahm, beschränken sich ihre Anwendungsmöglichkeiten nicht auf die Mathematik, sie kann auf viele weitere Gegenstandsbereiche appliziert werden, zu denen nun auch die Digital Humanities gehören.

Angewandte Kategorientheorie

Die angewandte Kategorientheorie oder *applied category theory* (ACT) ist eine neuere Entwicklung,² mit dem Ziel kategorientheoretische Ideen aus der Mathematik in andere wissenschaftliche Disziplinen zu übertragen. So heißt es programmatisch auf der ACT-Website im Proposal zu einem Workshop mit dem klingenden Titel „Towards An Integrative Science“:

„[...] we should treat the use of categorical concepts as a natural part of transferring and integrating knowledge across disciplines. The restructuring employed in applied category theory cuts through jargon, helping to elucidate common themes across disciplines. Indeed, the drive for a common language and comparison of similar structures in algebra and topology is what led to the development category theory in the first place, and recent hints show that this approach is not only useful between mathematical disciplines, but between scientific ones as well.“ (ACT2018)

Dies ist unter anderem in der Physik (Abramsky / Coecke 2007; Baez / Stay 2010), Linguistik (Coecke et al. 2010), den Neurowissenschaften (Brown / Porter 2008), Informatik (Ehrig et al. 2001) und der Philosophie (Landry 2017) geschehen.

Gerade in der Informatik spielt die Kategorientheorie eine ausgezeichnete Rolle, da formale Logik und Mengentheorie in der theoretischen Informatik und funktionale Programmiersprachen wie Haskell oder Module wie Catlab.jl für Julia in der angewandten Informatik durch sie verbunden sind.

In den oben genannten Fällen hat sich gezeigt, dass die angewandte Kategorientheorie einige für die Digital Humanities attraktive Eigenschaften aufweist: Der hohe Abstraktionsgrad und der Umstand, dass nur Objekte und Pfeile in einer Kategorie vorkommen, erlauben einen leichten Einstieg in kategorientheoretische Modellierung und der Einsatz von kommutativen Diagrammen zur Analyse macht die Verwendung benutzerfreundlich. Gleichzeitig können Modellierungen aufgrund der Modularität

und Kompositionalität der Theorie auch sehr komplexe Sachverhalte repräsentieren, ohne selbst unüberschaubar zu werden. Darüber hinaus bietet die Kategorientheorie vielfältige Anschlussmöglichkeiten an andere mathematische Bereiche. Sollte sich herausstellen, dass formale Logik oder Grafentheorie benötigt wird, ist es ein Leichtes die nötigen Übergänge herzustellen.

In diesem Sinne kann die Kategorientheorie in den Digital Humanities als leichtgewichtiges Modellierungstool eingesetzt werden, das nach Bedarf erweitert und skaliert werden kann. Dies soll nun anhand der von Spivak und Kent (2012) entwickelten *Ontology Logs* illustriert werden.

Was sind und was können *ologs*?

Ologs sind eine konkrete Anwendung der Kategorientheorie zur Wissensmodellierung. Der Ausgangspunkt ist natürliche Sprache, darum ist dieses Modell besonders für die textbasierten Digital Humanities geeignet. Ein *olog* besteht zunächst nur aus einer Sammlung gelabelter Objekte und Pfeile, die funktionale Relationen zwischen den Objekten ausdrücken. Diese einfachen *ologs* können bei Bedarf um einfache und komplexe *Typen* erweitert werden, d.h. um die Charakterisierung einer Menge, in der das Objekt Element ist. Ebenso können die Pfeile weiter spezifiziert werden als *Aspekte* oder *Attribute*.

Zu beachten ist, dass *ologs* immer perspektivisch an einen bestimmten Standpunkt gebunden sind, d.h. ausgehend von einer individuellen Lesart wird eine Konzeption entwickelt und Schritt für Schritt mit mehr Informationen angereichert, bzw. in Zusammenarbeit mit anderen Lesarten abgeglichen. Die Konstruktion von *ologs* stellt sicher, dass das Ergebnis strukturell valide ist und nicht faktisch korrekt. Dies ermöglicht es, im Modell mit Diskrepanzen umzugehen und Gemeinsamkeiten und Differenzen zwischen Modellierungen mit den, durch die Kategorientheorie bereitgestellten mathematischen Werkzeugen zu bearbeiten. Die Darstellung als kommutative Diagramme erleichtern die Kommunikation über unterschiedliche Interpretationen des Untersuchungsgegenstandes und unterstützen den geisteswissenschaftlichen Reflexionsprozess.

Da es jederzeit möglich ist ein bestehendes *olog* um weitere Pfeile und Objekte zu erweitern, können neue Informationen – wie zusätzliche Aspekte und Attribute – hinzugefügt werden, ohne das bereits entwickelte Schema von neuem aufbauen zu müssen.

Da *ologs* von Grund auf modular sind, können genaue Schnittstellen durch Funktoren spezifiziert werden, die kollaboratives Arbeiten erleichtern. *Ologs* können klar getrennte Abstraktionsebenen enthalten, die tief miteinander verlinkt sind, ebenso können mehrere *ologs* miteinander verbunden werden, um unterschiedliche Facetten des Untersuchungsgegenstands oder Sichtweisen auf ihn zu integrieren.

Abgrenzung der *ologs* zu anderen Wissensrepräsentationen

Ologs sind eng verwandt mit anderen Wissensrepräsentationen wie RDF/OWL, Datenbanken und semantischen Netzen, unterscheiden sich von ihnen jedoch in einigen wichtigen Punkten:

Im Vergleich zu RDF/OWL zeichnen sich *ologs* durch eine höhere Expressivität aus, da in *ologs* Kommutativität ausgedrückt werden kann. Aussagen der Art “Die Schwester meines Vaters ist meine Tante” können also ausgedrückt werden, da die Äquivalenz

der beiden Pfade $\vdash \text{ich} \rightarrow \vdash \text{Tante}$ und $\vdash \text{ich} \rightarrow \vdash \text{Vater} \rightarrow \vdash \text{Schwester} \rightarrow$ spezifiziert werden können, sodass $\vdash \text{Tante}$ und $\vdash \text{Schwester}$ Label des gleichen Objekts (Entität) sind.

Im Vergleich mit Datenbanken zeichnen sich *ologs* durch höhere Flexibilität aus, sie sind einfacher zu lesen, weniger präskriptiv und lassen sich leichter um neue Informationen erweitern. Dennoch lassen sich *ologs* direkt als Datenbanken implementieren, indem Objekte und Pfeile den Reihen und Spalten von Tabellen zugewiesen werden.

Semantische Netze und *ologs* sind sich sehr ähnlich, unterscheiden sich jedoch hinsichtlich der Robustheit: In semantischen Netzen müssen bei Veränderungen oft viele Links synchronisiert werden, was zu Fehlern führen kann, während bei *ologs* neue Pfade ohne Veränderung des bestehenden kreierte werden können.

Vortrag

Der Vortrag ist zweigeteilt: Im ersten Teil werden *ologs* mit den zugehörigen Formalismen und Diagrammen eingeführt und das Potential der kategorientheoretischen Ontologieentwicklung und Wissensmodellierung für die Digital Humanities herausgearbeitet. Besonderes Gewicht wird hierbei auf der präzisen Formulierung perspektivisch gebundener Sichtweisen auf Untersuchungsgegenstände, flexible Erweiter- und Kombinierbarkeit von *ologs* und den kollaborativen Möglichkeiten der Modellierung liegen. Darüber hinaus wird die Unterstützung des geisteswissenschaftlichen Denkens durch *ologs* und damit verbundene *rules of good practice* untersucht werden.

Im zweiten Teil werden dann die im ersten Teil gewonnen Einsichten anhand von Beispielen erläutert und gezeigt, wie die Wissensmodellierung durch *ologs* in den Workflow textbasierter Digital Humanities Projekte – wie zum Beispiel CATMA³ – integriert werden kann.

Das übergreifende Ziel des Vortrags ist es zu zeigen, dass eine mathematisch-theoretische Fundierung der Ontologieentwicklung und Wissensmodellierung sowohl das geisteswissenschaftliche Denken und Arbeiten unterstützen kann, als auch die technische Entwicklung und informatische Implementation treiben kann. Die Kategorientheorie hat in der Mathematik und Informatik sowie in Naturwissenschaften wie Physik, Chemie und Genetik bereits erfolgreich unter Beweis gestellt, ein geeignetes Denkwerkzeug zu sein. Der größte Nutzen in diesen Feldern ist durch Systematisierung und Vergleichbarkeit mit bzw. Anschlussfähigkeit an andere Forschungsgebiete entstanden. Diese Effekte gilt es auch für die Digital Humanities nutzbar zu machen.

Fußnoten

1. Ein kurzer Blick in “Ontology Development 101: A Guide to Creating Your First Ontology” (https://protege.stanford.edu/publications/ontology_development/ontology101.pdf) vermag das zu bestätigen.
2. Wenngleich Kategorientheorie schon immer auch angewendet wurde, so hat sich in den letzten Jahren eine lebhafte *scientific community* gebildet (vgl. <https://www.appliedcategorytheory.org> und <https://golem.ph.utexas.edu/category/>), seit 2018 gibt es auch ein OpenAccess Journal (<https://compositionality-journal.org>)
3. CATMA – Computer Assisted Text Markup and Analysis (<https://www.catma.de>)

Bibliographie

ACT2018 (2018): “Towards An Integrative Science”, <http://www.appliedcategorytheory.org/workshops/> [letzter Zugriff 15. Juli 2021].

Abramsky, Samson / Coecke, Bob (2007): “A categorical semantics of quantum protocols”, <https://arxiv.org/pdf/quant-ph/0402130.pdf> [letzter Zugriff 15. Juli 2021].

Baez, John C. / Stay, Mike (2010): “Physics, Topology, Logic and Computation: A Rosetta Stone”, <https://arxiv.org/pdf/0903.0340.pdf> [letzter Zugriff 15. Juli 2021].

Brown, Ronald / Porter, Timothy (2008): “Category Theory and Higher Dimensional Algebra: potential descriptive tools in neuroscience”, <https://arxiv.org/pdf/math/0306223.pdf> [letzter Zugriff 15. Juli 2021].

Coecke, Bob / Sadrzadeh, Mehrnoosh / Clark, Stephen (2010): “Mathematical Foundations for a Compositional Distributional Model of Meaning”, <https://arxiv.org/pdf/1003.4394.pdf> [letzter Zugriff 15. Juli 2021].

Ehrig, Hartmut / Mahr, Bernd / Große-Rhode, Martin / Cornelius, Felix / Zeitz, Philip (2001): *Mathematisch-strukturelle Grundlagen der Informatik*. Berlin: Springer.

Landry, Elaine M. (2017): *Categories for the working philosopher*. Oxford: Oxford University Press.

Mac Lane, Saunders (1998): *Categories for the working mathematician*. New York: Springer.

Spivak, David I. / Kent, Robert E. (2012): “Ologs: A Categorical Framework for Knowledge Representation”, in: *PLoS ONE* 7(1): e24274. <https://doi.org/10.1371/journal.pone.0024274>

Lesen, was wirklich wichtig ist

Die Identifikation von Schlüsselstellen durch ein neues Instrument zur Zitatanalyse

Arnold, Frederik

frederik.arnold@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Fiechter, Benjamin

fiechtbe@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Einleitung

Literaturinterpretationen heben in der Regel einige wenige Bestandteile des analysierten Primärtextes hervor, die für die jeweilige These und die damit verbundene Interpretation besonders wichtig erscheinen, und interpretieren sie mehr oder weniger ausführlich. Diese Passagen fassen wir als *Schlüsselstellen*, worunter wir solche Stellen eines Textes (häufig kurze Passagen oder Absätze) verstehen, die im Kontext einer Interpretation besonders

relevant sind. Schlüsselstellen, aber auch allgemein Stellen, wurden bislang weder aus einer interpretationstheoretischen noch aus einer interpretationspraktischen Perspektive systematisch untersucht.

Mithilfe des unten beschriebenen Konzepts, seiner Umsetzung als Algorithmus (*Lotte*) und einer zugehörigen Visualisierung als Webseite (*Annette*), die die Erkundung literarischer Texte ermöglichen, gelingt es uns nachzuvollziehen, welche Stellen über den individuellen Interpretationsansatz hinaus von Bedeutung sind und somit aus einer über den Zugriff von Expert*innen vermittelten Sichtweise das literarische Werk selbst konstituieren bzw. für dessen Lektüre maßgeblich sind. Unser Ansatz ermöglicht durch die Explorationsumgebung *Annette* den nahtlosen Wechsel zwischen *Close* und *Distant Reading* sowie eine sowohl auf den Primärtext als solchen als auch auf einzelne Sekundärtexte bezogene Perspektive. In diesem Sinn schließt sich unsere Herangehensweise an den von Martin Mueller 2012 eingeführten Begriff des *Scalable Reading*¹ an und bietet ein neues Konzept und dessen praktische Umsetzung zur Analyse von Korpora, die sich auf einen gemeinsamen Referenztext beziehen. Wir stellen somit ein neues Instrument zur Strukturidentifikation vor, das auf der Basis der Zitatanalyse arbeitet, aber auch über unseren Verwendungskontext hinaus von Interesse ist. Dabei lassen sich gleichermaßen das Verhalten der Interpret*innen und die Bezugnahmen der Korpus-texte auf den Primärtext als auch das Verhältnis der Korpus-texte untereinander untersuchen.

Im Folgenden gehen wir zunächst kurz auf verwandte Arbeiten ein, erläutern dann den Kontext von Konzept und Instrument, ehe wir den Algorithmus und die Webseite zur Exploration vorstellen; dabei wird der Fokus weniger auf technische Details des Algorithmus als auf dessen praktische Anwendung gelegt. Daran schließen sich Überlegungen zur Konzeption und eine Verortung in den Digital Humanities sowie ein Ausblick auf mögliche Erweiterungen des Instruments an.

Projektkontext

Die hier vorgestellte Herangehensweise ist ein zentrales Instrument, um im Rahmen des Projekts *Was ist wichtig? Schlüsselstellen in der Literatur*² (Humboldt-Universität zu Berlin, Teil des DFG-Schwerpunktprogramms *Computational Literary Studies*, SPP 2207)³, Schlüsselstellen zu identifizieren. Diese Schlüsselstellen sind stets nur in Relation zu einem bestimmten Kontext besonders wichtig, z. B. für den Primärtext selbst, für Teile oder Teilaspekte desselben, für ein schriftstellerisches Gesamtwerk oder für übergeordnete Thematiken, Motive etc. Inwiefern dabei die Perspektive des*der jeweiligen Literaturwissenschaftler*in entscheidend ist oder die Beschaffenheit des Primärtextes selbst, gehört zu den offenen Fragen. Auch die soziale Dimension des Wissenschaftssystems (“Was muss zwangsläufig zitiert werden, um die Kenntnisnahme der zuvor publizierten Beiträge abzubilden?”) ist bislang schwer greifbar. Trotz der bis jetzt nicht vorgenommenen Definition wird der Terminus “Schlüsselstelle” wörtlich, in Abwandlungen, die ebenfalls mit der Metapher “Schlüssel” arbeiten (z. B. “Schlüsselszene”), oder in ähnlichen Formulierungen (z. B. “Angelpunkt”) regelmäßig verwendet. Noch deutlich weiter verbreitet ist der Verweis auf “Stellen”, die in der Regel ein mehr oder weniger klar definiertes Teilstück des Textes meinen,⁴ wobei besonders die Länge solcher “Stellen” einer großen Varianz unterliegt. Auch für diesen Begriff existiert keine wissenschaftliche Definition, weder allgemein noch für die textbasierte Litera-

turwissenschaft, sodass eine Abgrenzung zu “Passage”, “Szene”, “Abschnitt” oder “Episode” nicht immer sinnvoll erscheint.

Exemplarisch haben wir die Untersuchung mit zwei kanonischen Texten der deutschen Literatur begonnen, zu denen zahlreiche Interpretationen vorliegen: *Die Judenbuche* (1842) von Annette von Droste-Hülshoff und *Michael Kohlhaas* (1808/1810) von Heinrich von Kleist. Zu beiden Texten verfügen wir über je ein Korpus von jeweils ca. 50 deutschsprachigen Interpretationstexten aus dem Projekt *Das Herstellen von Plausibilität in Interpretationstexten. Untersuchungen zur Argumentationspraxis in der Literaturwissenschaft*⁵ (Georg-August-Universität Göttingen). Diese Sekundärtexte sind zwischen 1995 und 2015 erschienen und wurden primär hinsichtlich ihrer Vergleichbarkeit und Diversität ausgewählt; die zeitliche Begrenzung ergibt sich aus dem Anspruch des Göttinger Projekts, die gegenwärtige Interpretationspraxis zu erforschen.⁶ Anhand des Korpus zur *Judenbuche* wollen wir im Folgenden zeigen, wie uns ein neues Instrument bei der Identifizierung von Schlüsselstellen unterstützt.

Lotte — Ein Werkzeug zur Erkennung von Textwiederverwendung

Lotte ist ein in Python implementierter Algorithmus zur Erkennung von Zitaten. Gegeben einen Quelltext und einen Zieltext, findet er alle Instanzen ab einer Länge von fünf Wörtern, in denen der Zieltext (in unserem Fall eine Literaturinterpretation) einen Teil des Quelltextes, also des literarischen Textes, enthält.⁷ Somit lassen sich alle wörtlichen Übernahmen ab einer bestimmten Länge von einem Text in einen anderen erkennen, ohne dass in dieser Hinsicht Vorarbeit geleistet werden muss; Voraussetzung sind lediglich zwei Textdateien, die sich sinnvollerweise aufeinander beziehen sollten (möglich ist natürlich auch, eventuell vorhandene Übernahmen erst durch Lotte zu ermitteln). Lotte basiert auf *Sim_text* von Grune und Huntjens (1989) und erweitert den Algorithmus um Funktionalitäten für die korrekte Behandlung spezifischer Eigenschaften von Zitaten. Für technische Details und eine ausführliche Evaluierung sei an dieser Stelle auf Arnold und Jäschke (2021) verwiesen.

Annette — Eine Webseite zur Visualisierung und Erkundung

Die Idee der Verwendung von (interaktiven) Visualisierungen in den Digital Humanities zur Unterstützung der Arbeit mit Texten aller Art ist nicht neu. Es gibt eine große Anzahl von Ansätzen zur Visualisierung von Strukturen, Häufigkeiten, Mustern etc. – sowohl zur Unterstützung beim Distant Reading als auch beim Close Reading bzw. zur Kombination beider Perspektiven, wie auch wir sie in dieser Arbeit vorstellen. Für einen ausführlichen Überblick sei verwiesen auf Jänicke et al. (2015a). Auch für die Visualisierung von wiederverwendetem Text, wie beispielsweise Zitaten, gibt es verschiedene Ansätze (Vgl. Jänicke et al. 2015b). Im Unterschied zu Annette liegt der Fokus dieser Ansätze auf der Visualisierung der Art und Häufigkeit des wiederverwendeten Texts oder der Alignierung von verschiedenen Varianten des gleichen Texts, zum Beispiel verschiedener Übersetzungen.

Eine Webseite zur Visualisierung von (wörtlichen) Zitaten aus Shakespeares Werken wurde von Miller vorgestellt.⁸ Sie veranschaulicht, wie oft jede Zeile Dialog aus jedem dramatischen

Werk Shakespeares in der Zeitschriftensammlung von JSTOR zitiert wurde. Die Website beschränkt sich auf die Visualisierung der Zitierhäufigkeit der einzelnen Zeilen und bietet keine Funktion zur Erkundung der Quellen der Zitate. Für eine Visualisierung von Zitaten aus verschiedenen Quellen, wie wir sie vorschlagen, sind die beschriebenen Ansätze nicht geeignet.

Für Annette werden die von Lotte gefundenen Übereinstimmungen weiterverarbeitet, um Schlüsselstellen zu identifizieren. Hierfür kombinieren wir überlappende Übereinstimmungen zu einer *Stelle* und erzeugen minimale nicht überlappende *Segmente* mit Häufigkeitsangaben. Das Ergebnis dieses Segmentierungsprozesses wird verwendet, um den literarischen Text und die wissenschaftlichen Texte zu visualisieren, wie im Folgenden beschrieben.

Ein Screenshot der Webseite⁹ ist in Abbildung 1 dargestellt. Links zeigt eine Heatmap des gesamten literarischen Textes die Verteilung der zitierten Passagen. Je dunkler der Text ist, desto häufiger wird er zitiert und als desto wichtiger wird er angenommen. Ist er weiß, wird er gar nicht zitiert.

Rechts neben der Heatmap ist der literarische Text selbst dargestellt. Die Grauskala wird dadurch bestimmt, wie viele Interpretationstexte einen Teil einer Stelle oder die Stelle insgesamt zitieren. Die Farbe ist dabei für eine gesamte Stelle immer die gleiche. Die Schriftgröße wird dadurch bestimmt, wie oft ein minimales Segment zitiert wird, somit kann sie auch innerhalb einer Stelle variabel sein.

Rechts unten neben dem literarischen Text wird eine Liste aller Interpretationstexte gezeigt. Ganz rechts werden die zehn häufigsten Passagen aufgelistet. Darüber wird bei einer entsprechenden Auswahl, wie unten beschrieben, der gesamte Text einer Interpretation angezeigt.

Ausgehend vom Startbildschirm kann der*die Benutzer*in zwischen verschiedenen Zugriffsmöglichkeiten wählen. Die erste Möglichkeit ist die Auswahl einer Stelle, indem diese im Primärtext angeklickt wird. Anstelle der Gesamtliste werden nun rechts davon nur noch diejenigen Interpretationstexte angezeigt, die zu der ausgewählten Stelle beitragen, zusammen mit einer kurzen Vorschau des Textes. Durch Anklicken eines der Interpretationstexte können wir eine bestimmte Interpretation anzeigen lassen, deren Text oben rechts dargestellt wird. Wir können dann diesen Text durchgehen und andere zitierte Passagen auswählen. Rechts unten wird angezeigt, wie oft die ausgewählte Stelle zitiert wird und von wie vielen Interpretationstexten. Darunter finden wir die zehn meistzitierten Segmente dieser Stelle.

Als weitere Möglichkeit kann über die Liste der zehn am häufigsten zitierten Stellen auf die möglicherweise relevantesten, jedenfalls am breitesten rezipierten Stellen zugegriffen werden. Nach Auswahl einer dieser Stellen passt sich wie oben beschrieben die Liste der Interpretationstexte an und es kann von hier aus weiter diese Liste oder ein einzelner Interpretationstext untersucht werden.

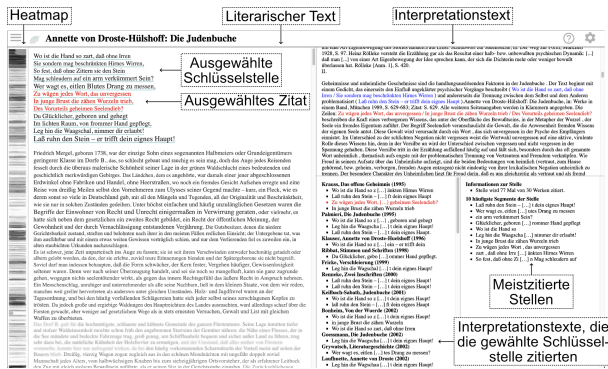


Abb. 1: Screenshot von Annette

Konzeptionelles

Durch die beschriebene Funktionsweise werden im Primärtext Stellen konstituiert, die teilweise sehr häufig, teilweise nur wenige Male zitiert werden. Im Fokus unserer Projektarbeit stehen die besonders häufig zitierten Stellen, da sie am ehesten die Schlüsselqualität einer Stelle anzeigen. Dabei darf die reine Quantität selbstverständlich nicht absolut gesetzt werden; um ein Korpus im Querschnitt zu überblicken, scheint uns die Häufigkeit von zitierten Stellen aber ein wichtiger Anhaltspunkt zu sein. Das Konzept, nach dem diese Stellen konstituiert werden, ist dabei die Pointe an unserem Ansatz, da es weit über das bloße Auffinden von Zitaten hinausgeht. Die auf der Webseite visualisierten Stellen ergeben sich nämlich häufig erst durch die Überlappung verschieden langer Segmente bzw. Zusammenfassung verschieden langer Zitate aus unterschiedlichen Interpretationstexten (s. o.). Somit sind Lotte und Annette im Zusammenspiel umso wirkungsvoller, je mehr Texte das Korpus umfasst.

Umgekehrt können aber auch jene Stellen von besonderem Interesse sein, die überhaupt nicht zitiert werden. Bei einer überschaubaren Textlänge, wie sie *Die Judenbuche* aufweist, liegt die Frage nahe, warum z. B. die Zeit nach Friedrich Mergels Heimkehr in den Interpretationstexten kaum eine Rolle spielt – zumindest auf der Ebene der wörtlichen Zitate. In diesem Sinn sagt die Webseite auch etwas über die Struktur der in die Interpretationen übernommenen Textstellen einerseits, andererseits aber auch des literarischen Textes selbst aus, indem visualisiert wird, wie sich die Struktur des literarischen Textes im Spiegel der Interpretationstexte darstellt. Darüber hinaus bietet es sich z. B. auch an, literarische Texte selbst miteinander zu vergleichen, um hier Verweise, Bezugnahmen, Hommagen oder aus anderen Gründen bewusst oder unbewusst übernommene Formulierungen aufzudecken – oder aber Plagiate nachzuspüren.

Diese Funktionsweise eröffnet auch den Horizont für ein integratives Scalable Reading, das wir im Anschluss an Thomas Weitin als Überwindung der “Frontstellung von *close* und *distant reading*” verstehen.¹⁰ So kann zunächst, wie oben beschrieben, der Primärtext ebenso “klassisch” gelesen werden wie die einzelnen Interpretationstexte,¹¹ aus denen unmittelbar über die wörtlichen Übernahmen zum jeweiligen Äquivalent im anderen Text gesprungen werden kann. Andererseits kann über die angezeigten am häufigsten zitierten Stellen und die Visualisierung als Heatmap auch ein quantitativer Zugang gewählt werden, der durch die sich anpassenden Visualisierungen die Nutzer*innen in die Lage ver-

setzt, zu entscheiden, wie *close* oder *distant* mit den Texten umgegangen werden soll.

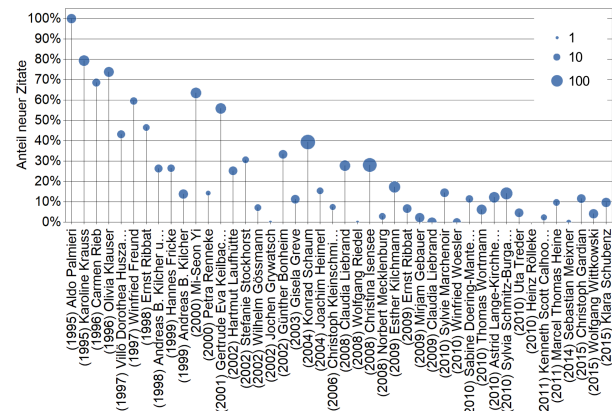
Die Vision für Annette

Ein erstes Ziel ist die Einführung weiterer Visualisierungen, die einen umfangreicheren Zugang zu den Texten bieten sollen. Im Folgenden wollen wir eine davon genauer vorstellen, die sich mit der Entwicklung der zitierten Stellen über die Zeit befasst, sowie abschließend ein paar weitere kurz skizzieren.

Darüber hinaus wollen wir in Zukunft Zitate, die kürzer als fünf Wörter sind, berücksichtigen. Hierfür muss in erster Linie Lotte angepasst werden. In Annette könnten diese dann so visualisiert werden, dass beispielsweise untersucht werden kann, ob kurze Zitate neue Informationen bringen und ob diese ähnlich zu langen Zitaten verteilt sind oder ob sich hier Unterschiede zeigen.

Entwicklung der zitierten Stellen über die Zeit

Das Ziel dieser Visualisierung ist die Darstellung des Anteils neuer Zitate eines Interpretationstexts im Verhältnis zu allen Zitaten früherer Interpretationstexte. Die Abbildungen 2 und 3 zeigen diese Visualisierung für *Die Judenbuche* und *Michael Kohlhaas*. Die horizontale Achse zeigt jeweils alle Interpretationstexte sortiert nach Erscheinungsjahr. Auf der vertikalen Achse ist der Anteil bisher nicht zitierter Zeichen unter allen zitierten Zeichen des Interpretationstexts aufgetragen. Der älteste Interpretationstext hat somit immer einen Anteil von 100%. Der Durchmesser eines Kreises gibt Auskunft über die Gesamtzahl der Zitate des Interpretationstexts aus dem Primärtext.

Abb. 2: *Die Judenbuche*: Anteil neuer Zitate pro Interpretationstext

Bei genauerer Betrachtung von Abbildung 2 lässt sich erkennen, dass die drei Texte mit den wenigsten neuen Zitaten auch insgesamt nur sehr wenig zitieren. Gleichzeitig gibt es auch Texte, die verhältnismäßig wenig zitieren und dennoch einen hohen Anteil an neuen Zitaten aufweisen.

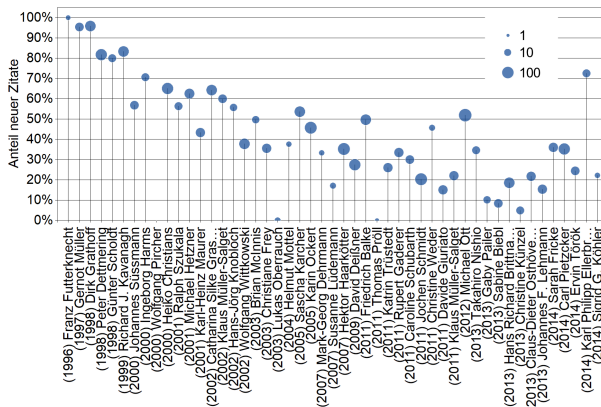


Abb. 3: Michael Kohlhaas: Anteil neuer Zitate pro Interpretationstext

Abbildung 3 zeigt analog die Auswertung für Michael Kohlhaas. Im Unterschied zu Abbildung 2 ist hier zu sehen, dass es immer wieder Interpretationstexte gibt, die viel zitieren und gleichzeitig einen hohen Anteil neuer Zitate aufweisen. Auch im Hinblick auf den Kurvenverlauf fallen Abweichungen gegenüber Abbildung 2 auf; ob sie allein der größeren Textmenge des Primärtextes geschuldet sind, ist offen.

Diese Diagramme sollen in interaktiver Form auf der Webseite verfügbar gemacht werden. Es wird dann zum Beispiel möglich sein, einen Interpretationstext auszuwählen und sich anzeigen zu lassen, welche zitierten Stellen neu sind. Diese können dann auch mit allen früheren und zukünftigen Zitaten verglichen werden.

Ideen für weitere Visualisierungen

Für die Zukunft sind noch weitere Visualisierungen angedacht. So soll zum Beispiel für alle Interpretationstexte ermittelt und visualisiert werden, ob diese in der Reihenfolge der Zitate dem literarischen Werk folgen. Weiterhin soll die Länge von Zitaten analysiert werden, was dann beispielsweise in die Bestimmung von Schlüsselstellen einfließen kann. Außerdem wollen wir untersuchen, welcher Anteil eines Werks mengenmäßig zitiert wird und wie sich die Zitate über den Interpretationstext verteilen.

Fazit

Im Fokus unseres Beitrags stand ein neues Konzept und dessen Umsetzung zur Identifikation von Schlüsselstellen in literarischen Texten. Neben der praktischen Realisierung als Algorithmus (Lotte) und der Visualisierung zur Erkundung von literarischen Texten (Annette) lag der Schwerpunkt auf den Möglichkeiten und zukünftigen Visionen für Annette. Darüber hinaus haben wir die Überlegungen zur Schlüsselstelle und zur Konstituierung von Stellen vorgestellt und als Beitrag zur Realisierung einer Scalable Reading-Methodik kontextualisiert, die auf Strukturidentifikation abzielt.

Danksagung

Die Arbeit wurde durch das DFG-Schwerpunktprogramm (SPP) 2207 *Computational Literary Studies* in dem von Robert Jäschke und Steffen Martus geleiteten Projekt *Was ist wichtig?*

Schlüsselstellen in der Literatur gefördert (Förderkennzeichen 424207720).

Fußnoten

1. Zwar erstmalig 2011 in seinem Blog verwendet (vgl. Mueller 2011), aber erst 2012 als Konzept eingeführt (vgl. Mueller 2012), dann in einem 2013 publizierten, aber bereits 2008 gehaltenen Vortrag angewandt (vgl. Mueller 2013). Vgl. auch den Überblickseintrag von 2020 (Mueller 2020). Ebenfalls 2013 hat Ryan Cordell den Begriff *Zoomable Reading* vorgeschlagen, der sich bis hin zur Metapher des Zoomens mit Muellers Konzept deckt (vgl. Cordell 2013). Im deutschsprachigen Raum hat Thomas Weitin den Begriff und seine deutsche Übersetzung bekannt gemacht (vgl. Weitin 2015 und Weitin 2017).
2. Homepage: <https://www.projekte.hu-berlin.de/de/schluesselstellen>. Zugriff am 08.11.2021.
3. Homepage: <https://dfg-spp-cls.github.io/>. Zugriff am 08.11.2021.
4. Vgl. beispielsweise den Eintrag im *Wörterbuch der deutschen Gegenwartssprache*, der die "Stelle" unter Bedeutungsvariante 1 („lokalisierter Ort, Punkt, Platz“) neben anderen Bedeutungen als „kürzeres Teilstück, kürzerer Abschnitt“ „in einem Schriftwerk, Theaterstück, Film“ bzw. „in einem Musikstück“ definiert (<https://www.dwds.de/wb/Stelle#d-1-1-3>). Zugriff am 08.11.2021).
5. Homepage: <https://uni-goettingen.de/de/587821.html>. Zugriff am 08.11.2021.
6. Vgl. <https://www.uni-goettingen.de/de/profil+und+ziele/588365.html>. Zugriff am 08.11.2021.
7. Quellcode: <https://scm.cms.hu-berlin.de/schluesselstellen/lotte>. Zugriff am 08.11.2021.
8. Vgl. Homepage: <http://shakespeare.visualizingbroadway.com/index.html>. Zugriff am 08.11.2021.
9. Verfügbar unter: <https://hu.berlin/annette>. Zugriff am 08.11.2021.
10. Weitin 2017, S. 2. Auch wenn unser Konzept durch den umstandslosen Wechsel zwischen den beiden Stufen close und distant bzw. die Möglichkeit, gleichzeitig nah und distanziert zu analysieren, skalierbar ist, kann es die Vorstellung eines tatsächlich stufenlosen Skalierens, die die Metapher des Zoomens bei Mueller nahelegt und die Weitin als "irreführend" bezeichnet, nicht gänzlich erfüllen.
11. In der oben verlinkten Version der Webseite kann aus urheberrechtlichen Gründen nur auf eine Version zugegriffen werden, bei der die Interpretationstexte unkenntlich gemacht wurden.

Bibliographie

- Arnold, Frederik / Jäschke, Robert** (2021): "Lotte and Annette: A Framework for Finding and Exploring Key Passages in Literary Works", in: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities at ICON 2021*.
- Art. "Stelle"** (1976): In: *Wörterbuch der deutschen Gegenwartssprache*. 5. Band. Bereitgestellt durch das Digitale Wörterbuch der deutschen Sprache, <https://www.dwds.de/wb/Stelle>. [letzter Zugriff 08.11.2021].
- Cordell, Ryan** (2013): "“Taken Possession of”: The Reprinting and Reauthorship of Hawthorne’s “Celestial Railroad” in the Antebellum Religious Press", in: *Digital Humanities Quarterly*

7.1 <http://digitalhumanities.org/dhq/vol/7/1/000144/000144.html> [letzter Zugriff 08.11.2021].

Droste-Hülshoff, Annette von (1978): “Die Judenbuche : Ein Sittengemälde aus dem gebirgigten Westphalen“, in: Droste-Hülshoff, Annette von: *Historisch-kritische Ausgabe : Werke, Briefwechsel* . Hrsg. von Winfried Woesler. 5. Band: Prosa. 1. Teil: Text. Bearbeitet von Walter Hüge. Tübingen: Niemeyer, 1–42.

Grune, Dick / Huntjens, Matty (1989): *Detecting copied submissions in computer science workshops* , https://dick-grune.com/Programs/similarity_tester/Paper.pdf [letzter Zugriff 08.11.2021].

Kleist, Heinrich von (1990): “Michael Kohlhaas“, in: Kleist, Heinrich von: *Sämtliche Werke und Briefe in vier Bänden* . Hrsg. von Klaus Müller-Salget u. a. 3. Band: Erzählungen, Anekdoten, Gedichte, Schriften. Frankfurt am Main: Deutscher Klassiker Verlag 11–142.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015a): “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges“, in: *EuroVis (STARs)* , S. 83–103.

Jänicke S., Efer T., Büchler M., Scheuermann G. (2015b): “Designing Close and Distant Reading Visualizations for Text Re-use“, in: *Computer Vision, Imaging and Computer Graphics – Theory and Applications* . Springer, Berlin u. a., S. 153–171, https://doi.org/10.1007/978-3-319-25117-2_10 . [letzter Zugriff 08.11.2021].

Miller, Derek: *To Quote or Not to Quote* , [o. J.], <http://shakespeare.visualizingbroadway.com/index.html> . [letzter Zugriff 08.11.2021].

Mueller, Martin (2011): “The Top Fifty n-gram heavy play links“, in: *Scalable Reading* (Blog), <https://sites.northwestern.edu/scalablereading/2011/04/01/the-top-fifty-n-gram-heavy-play-links/> [letzter Zugriff 08.11.2021].

Mueller, Martin (2012): “Very briefly: scalable reading“, in: *Scalable Reading* (Blog), <https://sites.northwestern.edu/scalablereading/2012/06/01/very-briefly-scalable-reading/> [letzter Zugriff 08.11.2021].

Mueller, Martin (2013): “Morgenstern’s Spectacles or the Importance of Not-Reading“, in: *Scalable Reading* (Blog) <https://sites.northwestern.edu/scalablereading/2013/01/21/morgensterns-spectacles-or-the-importance-of-not-reading/> [letzter Zugriff 08.11.2021].

Mueller, Martin (2020): “Scalable Reading“, in: *Scalable Reading* (Blog), <https://sites.northwestern.edu/scalablereading/2020/04/26/scalable-reading/> [letzter Zugriff 08.11.2021].

Weitin, Thomas (2015): *Thinking slowly: Literatur lesen unter dem Eindruck von Big Data* (= LitLingLab Pamphlet, 1). http://digitalhumanitiescenter.de/pamphlets/k13-01_weitin-thinking-slowly.pdf [letzter Zugriff 08.11.2021].

Weitin, Thomas (2017): “Scalable Reading“, in: *Zeitschrift für Literaturwissenschaft und Linguistik (LiLi)* 47,1: 1–6. <https://link.springer.com/article/10.1007/s41244-017-0048-4> [letzter Zugriff 08.11.2021].

Liebblingsgegenden, Fenster und Mauern

Zur emotionalen Enkodierung von Raum in Deutschschweizer Prosa zwischen 1850 und 1930

Herrmann, J. Berenike

berenike.herrmann@uni-bielefeld.de
Universität Bielefeld, Germany

Grisot, Giulia

giulia.grisot@uni-bielefeld.de
Universität Bielefeld, Germany

Einleitung

Raum ist eine wichtige Dimension von ‘Kultur’, nicht zuletzt in literarischen Artefakten. Definiert als “area, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors” (Council of Europe, 2000, S. 2) impliziert besonders die Landschaft, das ‘Gebiet, wie wahrgenommen’, einen oftmals vergleichenden En- und Dekodierungsakt.

Wer die Natur aufrichtig schätzt, hat seine Liebblingsgegenden, in welche er immer wieder zurückkehrt, selbst wenn er inzwischen überlegene landschaftliche Bilder kennengelernt haben sollte. (Carl Spitteler, *Xaver Z’Gilgen*, 1891)

Die Räume der deutschschweizer Literatur sind wie bei Spitteler offenbar regelmäßig solche “Lieblingsgegenden”, die dann doch Untiefen offenbaren, wie die Dörfer Gotthelfs, die Kleinstädte Kellers (*Seldwyla*) und Frischs (*Güllen*). Aber da ist auch der alpine Naturraum (Meyers *Jürg Jenatsch*, Heers *An heiligen Wassern*), sowie urbane und auch nichtschweizer Räume, wie etwa in Spyris Frankfurt am Main.

Es lief von einem Fenster zum anderen und dann wieder zum ersten zurück; aber immer war dasselbe vor seinen Augen, Mauern und Fenster und wieder Mauern und dann wieder Fenster. Es wurde Heidi ganz bange. (Johanna Spyri, *Heidis Lehr- und Wanderjahre*, 1880)

Spittelers scheinbar zahme Alpen, Spyris urbanes Gefängnis und schließlich Hölderlins erhabener “furchtbarherrlicher Haken” des Hochgebirgsaufstiegs (*Kanton Schweiz*, 1792) sind dabei ‘Kultur’ in doppeltem Sinne. Sie sind zum einen archivierte Beschreibungen von (fiktionalen) Raum, die zum kulturellen Gedächtnis gehören. Doch sind sie auch ‘Kultur’ im Luhmannschen Sinne (vgl. Schaffrick, 2017) – eine Beobachtungsoperation, mit der die beschriebenen Räume in Relation zu anderen (fiktional) enkodierten Räumen gesetzt werden.¹

Forschungsfrage und Vorgehen

Unser Beitrag möchte die beiden beschriebenen Ebenen von Kultur mit der des ‘digitalen Gedächtnisses’ zusammenbringen, indem wir computationale Verfahren auf literarische Texte (als

‘Schweizer digitales Kulturerbe’) anwenden, um die affektive Enkodierung dargestellter Raumtypen (als ‘Reflektion der Reflektion’) zu untersuchen.

Ausgehend vom übergreifenden Forschungsinteresse einer Komparatistik der deutschsprachigen Länder möchte unser Beitrag erste Ergebnisse berichten über die emotionale Enkodierung von fiktionalem Raum. Anhand des DCHLi (Deutschschweizer Literaturkorpus), zurzeit als Pilotkorpus mit 76 Texten, und ausgehend von einem semiotischen Zugang zu textuell enkodierten Emotionen (z.B. Schiewer, 2007; vgl. Anz, 2007; Winko, 2022) und Raumanalyse (Balshaw, & Kennedy, 2000; Bologna, 2020)

legen wir die in gängigen Sentiment-Diktionären vorgehaltenen Affekt-Kategorien zwischen dimensional (Valenz, Arousal) und diskreten Emotionen (“Angst”, “Freude”, “Wut”, “Trauer”, “Ekel”) an. Wir fragen:

Welche unterschiedlichen Typen von Landschaft und Raum gibt es in der fiktionalen deutschschweizer Prosa zwischen 1854 und 1930, und wie sind diese jeweils emotional enkodiert?

Unsere quantitativen Befunde sollen Bezüge herstellen zu ikonischen Kultur/Natur-Dichotomien im Erbe der Romantik, zu historischen Stadt/Land-Konstellationen, aber auch zu einem nationalliterarischen Rahmen mit vielbeklagtem Schweizer “Mythos” (Böhler, 2010) einerseits und identifikatorischen (oftmals Alpen-orientierten) Angeboten (Zimmer, 1998) für die “imagined community” (Anderson) der sogenannten Willensnation andererseits.

Daten und Methode

Unser DCHLi Pilotkorpus umfasst derzeit 76 fiktionale Prosatexte von AutorInnen, die der deutschschweizer “Nationalliteratur” zugeordnet werden und die zwischen 1854 und 1930 zuerst publiziert wurden (N= 2,025,529 Wörter). DCHLi enthält das wachsende Deutschschweizer ELTeC-gsw (Grisot & Herrmann, 2021) das wiederum Teil der European Literary Text Collection (ELTeC, Odebrecht et al., 2021) ist.

Author	Title	Publication year	Token number
Jakob Christoph Heer	An heiligen Wassern	1898	45757
Ulrich Kiebler	Aus Berg und Tal. Charakterbilder aus dem schweizer. Bauernleben	1903	14652
Johanna Spyri	Aus dem Leben	1902	21334
Marie Walden	Aus der Heimat	1884	49337
Ida Frohnmeyer	Aus Kinderland	1912	18896
Ida Frohnmeyer	Aus stillen Gassen	1921	12548
Hugo Marti	Balder. Sieben Nächte	1923	14555
Heinrich Federer	Berge und Menschen	1911	75130
Hugo Marti	Das Haus am Haff	1922	14844
Ludwig Rubiner	Das himmlische Licht	1916	2933
Meinrad Lienert	Das Hochmutsnährchen	1911	19447
Maria Waser	Das Jätvreni	1917	5967
Hugo Marti	Das Kirchlein zu den sieben Wundern	1922	15124
Gottfried Keller	Das Sinngedicht	1882	47014
Heinrich Federer	Das Wunder in Holzschuhen	1919	1454
Meinrad Lienert	Der doppelte Matthias und seine Töchter	1929	43299
Heinrich Federer	Der Fürchtemacher	1919	7691
Felix Moeschlin	Der glückliche Sommer	1920	32312
Conrad Ferdinand Meyer	Der Heilige	1880	20405
Gottfried Keller	Der grüne Heinrich. Bd. 1	1854	32965
Gottfried Keller	Der grüne Heinrich. Bd. 2	1854	38186
Gottfried Keller	Der grüne Heinrich. Bd. 3	1854	29603
Gottfried Keller	Der grüne Heinrich. Bd. 4	1855	38806
Hugo Marti	Der Kelch. Gedichte	1925	2217
Jakob Christoph Heer	Der König der Bernina	1900	36637
Meinrad Lienert	Der König von Euland	1928	22183
Jakob Christoph Heer	Der lange Balthasar	1915	26117
Ludwig Rubiner	Der Mensch in der Mitte	1917	19757
Meinrad Lienert	Der Pfeiferkönig. Eine Zürcher Geschichte	1909	26900
Lisa Wenger	Der Rosenhof	1915	38191
Jakob Christoph Heer	Der Wetterwart	1905	49962
Maria Waser	Die Geschichte der Anna Waser	1913	57319
Conrad Ferdinand Meyer	Die Hochzeit des Mönchs	1884	13906
Conrad Ferdinand Meyer	Die Richterinnen	1885	10964
Lisa Wenger	Die Wunderdoktorin	1910	39897
Jakob Bosshart	Ein Rufer in der Wüste	1921	53657
Lisa Wenger	Er und Sie und das Paradies	1918	37666
Helene Welti	Famulus der seltsame Pudel	1925	13438
Jakob Christoph Heer	Felix Notvest	1901	38026
Alexander Castell	Fieber. Drei Novellen	1916	19203
Carl Spitteler	Friedli der Kolderli	1891	4294
Heinrich Federer	Gebt mir meine Wildnis wieder	1918	9274
Conrad Ferdinand Meyer	Gedichte	1882	18717
Johanna Spyri	Heidi kann brauchen, was es gelernt hat	1881	17252
Johanna Spyri	Heidi's Lehr- und Wanderjahre	1880	24463
Johanna Spyri	Heimatlos	1878	27772
	Herrn Dames Aufzeichnungen oder Begebenheiten aus einem merkwürdigen Stadtteil	1913	16353
Fanny Gräfin zu Reventlow	Im Rhonethal	1880	12508
Carl Spitteler	Imago	1906	23000
	Jahresring. Ein poetischer Roman voll Nordlandzauber	1925	14939
Hugo Marti	Jungfer Therese	1913	41118
Heinrich Federer	Laubgewind	1908	38992
Jakob Christoph Heer	Lyrische Dichtungen	1923	6001
Heinrich Leuthold	Martin Salander	1886	46619
Gottfried Keller	Pilatus. Eine Erzählung aus den Bergen	1912	39275
Heinrich Federer	Regina Lob	1925	36778
Ina Jens	Rosmarin. Weitere Erlebnisse aus Majas Kindheit	1930	11935
Theobald Baerwert	Rosswiler Geschichten und anderes	1918	14919

Tab. 1: Auszug aus dem DCHLi Pilotkorpus

Ausgehend vom derzeit *de facto* Standard der diktionsbasierten Sentimentanalyse innerhalb der DH (vgl. Kim & Klinger, 2019) nutzen wir zur Co-Identifizierung von räumlichen Entitäten und Affekt acht für das Deutsche gängige frei verfügbare Sentiment-/Emotions-Diktionäre, sowie Ressourcen mit geopolitischen und räumlichen Informationen. Ähnlich wie Heuser et al. (2016) identifizieren wir zunächst regelbasiert lexikalisch ‘räumliche Entitäten’ als “Seedwords”, und analysieren daraufhin innerhalb einer Spanne von + - 50 Wörtern um das seedword den enkodierten Affekt.

Spatial entities

Im ersten Schritt erstellten wir ein möglichst umfassendes und feingranuliertes Diktionär “räumlicher Entitäten”, das auf höchster Taxonomie-Ebene die Kategorien RURAL und URBAN zusammenfasst, die sich wiederum in fünf Subkategorien ‘natural entity’, ‘rural entity’, und ‘geographic entity’, sowie ‘urban entity’ und ‘geopolitical entity’ auffächern (vgl. Wartmann et al. 2018, p. 1580; siehe Abb. 1).

URBAN entities

- "urban entities": i.e. spatial terms relating to the city, its buildings and infrastructures (e.g. *Bahnhof*, *station*, *Kreuzung*, *cross*, *Palast*, *palace*); (openthesaurus, wiktionary, opendata)
- "geopolitical entities": proper names of cities and villages in Switzerland, Austria, Germany, France and Italy
- <http://www.geonames.org/>
- <https://www.swisstopo.admin.ch/en/home/meta/supply-structure/freely-available.html>

RURAL entities

- "rural entities": i.e. spatial terms relating to – or characteristic of – the countryside, in particular related to human settlements or infrastructures, as opposed to those of the city (e.g. *Wanderweg*, *footpath*, *Feld*, *field*, *Hütte*, *hut*, *shack*)
- "natural entities": terms describing spatial elements as found in nature, not involving anything made or done by people (e.g. *Baum*, *tree*, *Bach*, *brook*, *Felsen*, *rock*)
- "geographical entities": proper names of natural locations such as mountains, rivers, valley, lakes (e.g. *Matterhorn*, *Mont Blanc*, *Donaue*)

Subcategory	Entity	Sum
Geographical 7,933	RURAL	8,684
Natural 478		
Rural 273		
Geopolitical 2,325	URBAN	2,596
Urban 271		

Abb. 1: Taxonomie der räumlichen Entitäten mit Gesamtanzahl der Elemente (Stand 9. Juli 2021).

Hier wurden unter Rückgriff auf Ressourcen wie Openthesaurus und das Schweizer Idiotikon historisch wie sprachlich relevante Elemente berücksichtigt (i.e. *Weiher*, *Weg*, *Hütte*, *Berg*, *See*, *Straße*, *Gebäude*, *Dom*; *Wiler*, *Bergli*). Für die geopolitischen (i.e. *Basel*, *Zürich*, *Berlin*, *Rom*) and geographischen Elemente (i.e. *Matterhorn*, *Rigi*, *Rhein*) nutzen wir als digitale Ressourcen unter anderem Wikidata, Ortsnamen und Swisstopo. Die resultierenden Listen wurden händisch nachkorrigiert und werden auf GitHub (wie auch der gemeinfreie Teil des Korpus sowie der Code) frei zugänglich publiziert werden.

Sentiment und Emotion

Im zweiten Schritt erstellten wir für einen systematischen Vergleich ein Repositorium mit acht der frei verfügbaren Sentiment-Diktionäre (ADU, BAWL, Germanlex, LANG, Klinger, Plutchik, SentiWS, SentiArt, siehe Tabelle 2).

Tab. 2: Sentiment lexicons

BAWL-R (Vö et al., 2009)
LANG (Kanske & Kotz, 2010)
Plutchik (Stamm, 2014)
Klinger (Klinger et al., 2016)
Adu (Hölzer et al., 1992)
Germanlex (Clematide et al., 2010)
SentiWS (Remus et al., 2010)
SentiArt (Jacobs, 2019)

Deren unterschiedlichen Formate wurden für die automatische Sentiment-Annotation in einer processing pipeline vereinheitlicht. Abbildung 2 zeigt die lexikalische Abdeckung der acht Diktionäre auf dem DCHLi, die angibt, wie viele der Wörter von der jeweiligen Ressource erkannt wurden (geordnet in abfallender Reihenfolge).

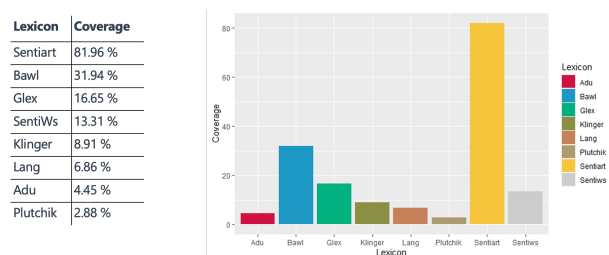


Abb. 2: Lexikalische Abdeckung der Sentiment-Diktionäre im Vergleich

Im dritten Schritt errechneten wir das semiotische Emotionspotenzial innerhalb der räumlichen Seedword-Spannen. Abbildung 3 zeigt den Abbildungsprozess der räumlichen Entitäten auf das Korpus beispielhaft für das BAWL-R-Diktionär: Sobald eine Entität identifiziert ist, werden innerhalb einer Gesamtspanne von 101 Wörtern je 50 Wörter vor und nach der Entität für die Berechnung von Emotions- und Sentimentwerten einbezogen (ohne Stoppwörter).

token	lemma	type_grouped	title_full	author_full	pub_date	valence	arousal
umgehen	umgehen	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	NA	NA
Blitz	Blitz	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	0.500	3.905
Donner	Donner	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	-0.200	3.947
knallte	knallen	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	NA	NA
Krach	Krach	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	NA	NA
Berge	Berg	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	NA	NA
stürzen	stürzen	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	-1.618	3.357
heut	heuen	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	NA	NA
übereinander	übereinan	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	NA	NA
rief	rufen	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	0.088	2.684
Amstalden	Amstalden	NAT_RUR	Der Fürchtemacher	Heinrich Federer	1919	NA	NA
SUM by span		type_grouped	title_full	author_full	span_id	arousal	valence
		NAT_RUR	Der Fürchtemacher	Heinrich Federer	geonat_20	85.337	94.047
		URBAN_LOC	heirich04	keller	urban_568	82.40	125.59

Abb. 3: Schematisches Beispiel für den Abbildungsprozess des Sentiment- / Entitäten-Matchings (BAWL-R).

Mittels dieses Verfahrens kann die Repräsentation von Emotionen (Valenz, Diskrete Emotionen) und ihre Ausprägung (Arousal) bezüglich des fiktionalen Raums näherungsweise untersucht werden, wobei uns zunächst die potenzielle Differenz in der Emotionsrepräsentation zwischen ländlichen und städtischen Räumen interessierte.

Ergebnisse und Diskussion

Wir verwendeten R (Version R 4.1.0, R Core Team, 2021) um mittels mixed linear models den Effekt des Entitätstyps (rural, urban) auf die jeweiligen Sentiment-Werte zu beobachten, mit AutorIn und Titel als randomisierte Faktoren. Verwendete Pakete waren v.a. tidyverse (Wickham et al., 2019); lmerTest (Kuznetsova et al., 2017), und tm (Feinerer, 2020).

Wir konnten statistisch signifikante Effekte des Entitätstyps u.a. auf die Valenz/Polarität in verschiedenen Diktionären beobachten, wobei LANG und BAWL "positives Sentiment" häufiger für "rural" Passagen aufwiesen, Germanlex jedoch den entgegengesetzten Befund (Tabelle 3).

Lexicon	RURAL	URBAN	Results of lmer (linear mixed model) Effect of entity type on values, with author and title as random factors			
			Estimates	SE	z	p
BAWL (31.94 %)						
• valence	+	-	-0.68 *	0.27	-2.56	0.011
• arousal	-	+	0.38 *	0.19	2.06	0.039
• imageability	+	-	-0.93 **	0.34	-2.76	0.006
LANG (6.86 %)						
• valence	+	-	-2.32 ***	0.19	-12.17	<0.001
• arousal	+	-	-1.56 ***	0.15	-10.71	<0.001
• concreteness	+	-	-0.46 **	0.15	-3.06	0.002
SentiWS (13.31 %)						
• polarity	ns.	ns.	-0.05	0.05	-0.99	0.324
Germanlex (16.65 %)						
• polarity	-	+	0.26 ***	0.05	5.21	<0.001

Tab. 3: Statistischer Vergleich durchschnittlichen Sentiments (Valenz, Arousal, Polarität) über DCHLit für Textpassagen um *rural* und *urban* (BAWL, LANG, SentiWS, Germanlex, Coverage in Klammern). “+” und “-” in Spalten 2 und 3 bezeichnen einen jeweils positiv oder negativ signifikant abweichenden Wert, während ns. (“nicht signifikant”) auf die Abwesenheit eines signifikanten Effekts des Entitätstyps hinweist.

Emotion	Adu (4.45 %)		Klinger (8.91 %)		Plutchik (2.88 %)		SentiArt (81.96 %)		Emotion
	RURAL	URBAN	RURAL	URBAN	RURAL	URBAN	RURAL	URBAN	
Joy	+	-	ns.	ns.	ns.	ns.	+	-	Joy
Fear	ns.	ns.	+	-	ns.	ns.	ns.	ns.	Fear
Sadness	ns.	ns.	ns.	ns.	ns.	ns.	+	-	Sadness
Surprise	ns.	ns.	ns.	ns.	ns.	ns.	-	+	Surprise
Disgust	ns.	ns.	ns.	ns.	ns.	ns.	+	-	Disgust
Anger	ns.	ns.	ns.	ns.	ns.	ns.	+	-	Anger
Depression	+	-							Depression
Love	-	+							Love
AAZ							+	-	AAZ

Tab. 4: Statistischer Vergleich durchschnittlicher diskreter Emotionswerte über DCHLit für Textpassagen um *rural* und *urban* (ADU, Klinger, Plutchik, SentiArt, Coverage in Klammern). Legende s. Tabelle 3.

Für die diskreten Emotionen berücksichtigte das mixed model jede einzelne Emotion als zusätzlichen “fixed factor” (Tabelle 4). Abweichungen zwischen den Diktionären konnten wieder beobachtet werden, wobei SentiArt signifikante Differenzen für fünf Basisemotionen (ausser *Angst*) detektiert. *Freude*, *Trauer*, *Ekel* und *Wut* sind dabei häufiger in “rural” Passagen zu finden, während Überraschung häufiger in “urban” Passagen auftritt.

Obwohl die Sentimentdetektion, das räumliche Matching der Emotionen und die Korpusgröße weiter verbessert werden sollen, interpretieren wir die vorliegenden Daten vorsichtig dahingehend, dass Textpassagen mit ländlichen und Natur-Referenzen in unserem Korpus häufiger positiv enkodiert sind. Es scheint, dass diese «ruralen» und «Natur-» Räume im Vergleich insgesamt mehr unterschiedliche und möglicherweise reichhaltigere Emotionen repräsentieren.

Angeht die Zusammensetzung des vorliegenden Korpus kann dies nicht nur auf eine topische Assoziation von positiv enkodierter Natur vs. negativ enkodierter Stadt/industrialisierter Zivilisation bezogen werden, sondern scheint auch Landschaft und Natur als vornehmlichen Schauplatz der Diegese abzubilden. Schlägt man den Bogen weiter, und projiziert noch hypothetisch auf die Grundgesamtheit der Deutschschweizer Prosa (Herrmann et al., 2021), könnte der Vorschlag, dass Deutschschweizer Literatur in dieser Zeit vornehmlich auf dem Lande und in der Natur stattfindet, im Luhmannschen Sinne als ‘kultureller’ Differenzvorschlag verstanden werden: ein Identifikationsangebot, das ‘Schweiz’ ebendort, und nicht anderswo, verortet. Wohlgermerkt wäre gerade unter solchen Bedingungen die evidente Rolle von

Technik, Infrastruktur, Handel und Industrialisierung mitzumodellieren.

Wir schließen mit einer unabdingbaren methodologischen Notiz. In der vorliegenden Studie war es unsere Absicht, diktionsbasierte Sentimentanalyse als im Feld der DH gegenwärtig noch kanonischen Ressourcentyp in Anschlag zu bringen (Kim & Klinger, 2019). Die niedrige lexikalische Abdeckung für die meisten Diktionäre, die im Umlauf sind (Abb. 2), zeigt auf, dass hier neue Ressourcen und ein erweitertes Methodenbewusstsein nötig sind. Untersucht man die Reliabilität und Domänenspezifität der einzelnen Diktionäre genauer, wie wir es getan haben, wird schnell deutlich, dass es sich für die DH lohnt, den Anschluss an den *State of the Art* des Affective Computing aktiv zu verfolgen.

Die Verwendung von vektorraumbasierten Diktionären wie SentiArt, aber besonders die Domänenadaptation des avancierten maschinellen Lernens, auch auf feinjustierten Annotationen (Kim & Klinger, 2018; Hoang et al., 2019), sind notwendig, um Nuancen, Objekte und Bedingungen von fiktional enkodiertem Affekt sicher zu detektieren. So bereiten wir derzeit manuelle Annotationen zur Implementierung in einer *deep learning* Architektur vor und rechnen mit aussagekräftigen Ergebnissen zum Zeitpunkt des Vortrages. Zudem erweitern wir derzeit die Raumentitätszuordnung bezüglich von Elementen des Interiors, da wir davon ausgehen, dass diese im Allgemeinen und Spezifischen in urbanen Settings häufiger auftreten. Diese Annahme prüfen wir in explorativen Studien.

Fußnoten

1. “Seit dem Ende des 18. Jahrhunderts besetzt der Begriff der Kultur den Platz, an dem Selbstbeschreibungen reflektiert werden” (Luhmann, 1997, S. 880). Wer etwas als “Kultur” thematisiert, richtet ein muster(er)findendes, vergleichendes *bird’s eye* auf bestimmte Gepflogenheiten des Lebens. ‘Kultur’ ist also Vergleichsoperation im Modus der Beobachtung zweiter Ordnung, und literarische Texte sind dafür Musterkandidaten: in der zerdehnten und oft mehrreihigen Kommunikationssituation zwischen Autorinstanz, Erzählinstanzen und Lesenden wird das Beobachten erster Ordnung im Erzählen ganz besonders beobachtbar.

Bibliographie

- Anz, T. (2007). Kulturtechniken der Emotionalisierung: Beobachtungen, Reflexionen und Vorschläge zur literaturwissenschaftlichen Gefühlsforschung. In *Im Rücken der Kulturen*. - Paderborn: Mentis-Verlag, pp. 207–39.
- Balshaw, M., & Kennedy, L. (2000). *Urban space and representation*. Pluto.
- Bologna, F. (2020). A Computational Approach to Urban Space in Science Fiction. *Journal of Cultural Analytics*. <https://doi.org/10.22148/001c.18120>
- Böhler, M. (2010). *Gefängnis Schweiz oder Bergnebel Seldwyla?* Max Niemeyer Verlag. <https://www.degruyter.com/document/doi/10.1515/9783484970526.1.45/html> (accessed 13 July 2021).
- Clematide, S. & Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. doi:10.5167/UZH-45506. <https://www.zora.uzh.ch/id/eprint/45506> (accessed 12 July 2021).
- Council of Europe. (2009). *European Landscape Convention*. Report and Convention Florence. *ETS*, 17(8).

Feinerer, I. & Hornik, K. (2020). tm: Text Mining Package. R package version 0.7-8. <https://CRAN.R-project.org/package=tm>

Grisot, G., & Herrmann, J.B. (Eds.) (2021). Swiss German Novel Collection (ELTeC-gsw), Version v1.0.0, July 2021. In: European Literary Text Collection (ELTeC). COST Action Distant Reading for European Literary History. <https://github.com/COST-ELTeC/ELTeC-gsw/blob/master/README.md>

Herrmann, J. B., Grisot, G., Gubser, S., & Kreyenbühl, E. (2021). Ein großer Berg Daten? Zur bibliothekswissenschaftlichen Dimension des korpusliteraturwissenschaftlichen DH-Projekts "High Mountains – Deutschschweizer Erzählliteratur 1880-1930". 027.7 Journal for Library Culture.

Heuser, R., Moretti, F. & Steiner, E. (2016). The Emotions of London. Literary Lab Pamphlet, 13. <https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf> (accessed 15 December 2020).

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using BERT. Proceedings of the 22nd Nordic Conference on Computational Linguistics, 187–196.

Hölzer, M., Scheytt, N. and Kächele, H. (1992). Das „Affektive Diktionär Ulm“ als eine Methode der quantitativen Vokabularbestimmung. In Züll, C. and Mohler, P. Ph. (eds), *Textanalyse: Anwendungen der computerunterstützten Inhaltsanalyse. Beiträge zur 1. TEXTPACK-Anwenderkonferenz*. (ZUMA-Publikationen). Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 131–54 doi:10.1007/978-3-322-94229-6_7. https://doi.org/10.1007/978-3-322-94229-6_7 (accessed 12 July 2021).

Jacobs, A. M. (2019). Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI*, 6 doi:10.3389/frobt.2019.00053. <https://www.frontiersin.org/article/10.3389/frobt.2019.00053/full> (accessed 8 September 2019).

Kanske, P., & Kotz, S. A. (2010). Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, 42(4), 987–991. <https://doi.org/10.3758/BRM.42.4.987>

Kim, E. & Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. Proceedings of the 27th International Conference on Computational Linguistics, 1345–1359.

Kim, E. & Klinger, R. (2019). A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *Zeitschrift Für Digitale Geisteswissenschaften*. 10.17175/2019_008.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *JOSS*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>

Luhmann, N. (1997). *Die Gesellschaft Der Gesellschaft*. Suhrkamp.

Odebrecht, C., Burnard L., & Schöch, C. (Eds.) (2021). European Literary Text Collection (ELTeC), version 1.1.0, April 2021. COST Action Distant Reading for European Literary History (CA16204). DOI: doi.org/10.5281/zenodo.4662444.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Schaffrick, M. (2016). Niklas Luhmann (1927–1998), *Kultur Als Historischer Begriff* (1995). *KulturPoetik*, 16(2), pp. 272–80.

Schiewer, G. L. (2007). Bausteine zu einer Emotionssemiotik: Zur Sprache des Gefühlsausdrucks in Kommunikation und affective computing. *Kodikas/Code. Ars Semeiotica: An International Journal of Semiotics*, 30(3–4), 235–257.

Stamm, N. (2014). Klassifikation und Analyse von Emotionswörtern in Tweets für die Sentimentanalyse.

Vö, M. L. H., Jacobs, A. M., & Conrad, M. (2006). Cross-validating the Berlin affective word list. *Behavior Research Methods*, 38(4), 606–609.

Wartmann, F. M., Acheson, E., & Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8), 1572–1592. <https://doi.org/10.1080/13658816.2018.1445257>

Wickham et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>

Winko, S. (2022). Literature and Emotion. In Schiewer, G.L., Altarriba, J., & Ng, B.C. (Hgg.), *Handbook on Language and Emotion. Handbooks of Linguistics and Communication Science*, HSK. Berlin; Boston: De Gruyter.

Zimmer, O. (1998). In search of natural identity: Alpine landscape and the reconstruction of the Swiss nation. *Comparative Studies in Society and History*, 40(4), 637–665.

Literaturgeschichtsschreibung datenbasiert und wikifiziert? Automatische Extraktion thematischer Statements aus französischen Primärtexten mithilfe von Topic Modeling, RDF und eines kontrollierten Vokabulars in LOD

Röttgermann, Julia

roettger@uni-trier.de
Universität Trier, Germany

Klee, Anne

klee@uni-trier.de
Universität Trier, Germany

Hinzmann, Maria

hinzmannm@uni-trier.de
Universität Trier, Germany

Schöch, Christof

schoech@uni-trier.de
Universität Trier, Germany

Welche Formalisierungs- und Modellierungsarbeit ist nötig, um Kulturen des kollektiven Gedächtnisses wie die Literaturgeschichtsschreibung als Daten abfragbar zur Verfügung zu stellen? Wir sehen aktuell einige Umbrüche in den Strategien der Gedächtnisinstitutionen, die sich zunehmend dem 'Linked Open Data'-Paradigma verpflichtet sehen.¹ Am Beispiel der Domäne französischer Literatur des 18. Jahrhundert verfolgt das Projekt "Mining and Modeling Text" einen ähnlich gearteten, jedoch im Bereich der Literaturgeschichtsschreibung neuen Ansatz einer datenbasierten, wikifizierten Arbeitsweise. Durch den Fokus auf eine spezifische literaturgeschichtliche Domäne entsteht ein besonders dichtes Netz von Aussagen, die über eine systematische

Extraktion von thematischen Aussagen aus bibliographischen Daten

Bei der zweiten Informationsquelle handelt es sich um bibliographische Nachweissysteme zur französischen Literatur 1751-1800 (s. Abb. 3). Im Fokus steht die *Bibliographie du genre romanesque français* (Martin et al. 1977), die die Grundgesamtheit der literarischen Produktion der entsprechenden Dekaden sorgfältig dokumentiert und Schlagworte zu thematischen Inhalten der Romane enthält, die jedoch nicht indexiert sind.⁶

Die Bibliographie bietet in Kombination mit den Ergebnissen des Topic Modelings die Möglichkeit eines Mensch-Maschine-Vergleichs – wurden die enthaltenen thematischen Schlagworte doch in den 1970er Jahren durch Lektüre und Zusammentragen von Informationen aus anderen Nachschlagewerken erhoben. Die Bibliographie wurde in mehreren Arbeitsschritten aufwendig erschlossen.⁷ Die Extraktion der thematischen Informationen stellt im Wechselspiel mit deren semantischer Modellierung eine besondere Herausforderung dar, da sie einerseits in sehr heterogener Form und andererseits nicht klar abgegrenzt zu weiteren Informationskategorien in der Bibliographie vorliegen.⁸ Zur Identifikation der häufigsten thematischen Aussagen, welche als Statements in das Wissensnetzwerk eingespeist werden, wurde das Korpusanalysetool TXM (vgl. Heiden 2010) genutzt.⁹ Jede dieser Aussagen wird auf ein Konzept unseres kontrollierten Themenvokabulars gemappt. So können die Strings automatisch extrahiert und als Statements formuliert werden. Aus dem Eintrag zu *Les enfans de la nature* von Pierre Blanchard in der Bibliographie (String aus 4./5. Spalte: <naufgabe, robinsonade, intrigue sentimentale; thèmes pédagogiques et philosophiques>) können beispielsweise die folgenden thematischen Statements abgeleitet werden:

[Les enfans de la nature] ABOUT [sentiment | Gefühl | sentiment]

[Les enfans de la nature] ABOUT [pedagogy | Pädagogik | pädagogie]

[Les enfans de la nature] ABOUT [philosophy | Philosophie | philosophie].

Technisch unterscheiden sich die RDF-Triple zu Themen je nach Datenquelle nicht, werden jedoch entsprechend ihrer Herkunft in Wikibase mit der Property *stated in* (*P14*) referenziert.

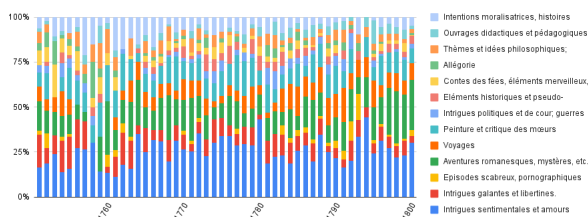


Abb. 3: Themenkategorien des französischen Romans 1751-1800 (Martin et al. 1977: xlvi–xlix).

In den bibliographischen Daten sind insgesamt knapp 2700 Items (Veröffentlichungen fiktionaler Prosa in französischer Sprache inklusive Übersetzungen) enthalten, von denen 349 das thematische Schlagwort 'voyage' enthalten.

Rolle des kontrollierten Vokabulars

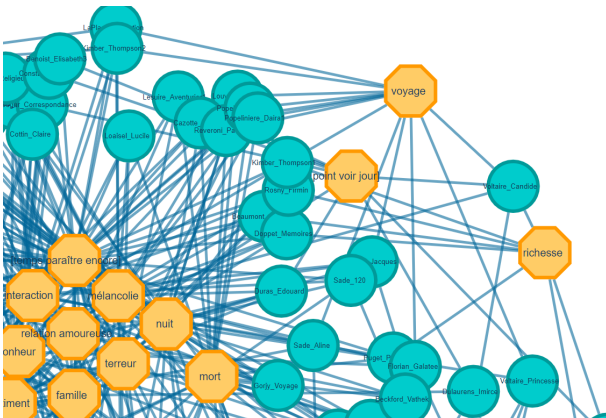
Wie lassen sich die thematischen Muster in der Primärliteratur mit den Daten aus den bibliographischen Nachweissystemen vergleichen? Ein wichtiger Modellierungsschritt ist zunächst das Erstellen eines kontrollierten Vokabulars aus thematischen Konzepten der französischen Aufklärung, auf das die Ergebnisse des Topic Modeling und die Bibliographie-Schlagworte gemappt werden.

Das Vokabular der Themenbegriffe besitzt eine hohe Relevanz für mehrere Teilprojekte: Es stellt zum einen die Labelbegriffe für die Topics aus dem Topic Model bereit, liefert daneben aber auch die Konzept-Items für die Objektposition solcher thematischer Statements, die aus der Sekundärliteratur und der Bibliographie extrahiert wurden.

An das Vokabular sind somit mehrere Anforderungen geknüpft: Die Begriffe müssen die Themenkonzepte der französischen Aufklärung abdecken, sollen ein gewisses Abstraktionslevel aufweisen, damit sie als kategorische Begriffe fungieren können und die Zusammenstellung der Begriffe sollte transparent und nachvollziehbar sein. Eine erste Grundlage bildet das Themeninventar des *Dictionnaire européen des Lumières* (Delon et al. 2007). Die Artikelstichwörter bieten eine gute Abdeckung an gesellschaftlich, politisch, ideengeschichtlich oder kulturell relevanten Themen der Epoche und stellen somit einen geeigneten Grundstock an möglichen Labels für die in den Romanen vorkommenden Themen. Dennoch enthält die Ressource Begriffe, die entweder zu spezifisch (z.B. 'pyrrhonisme') oder zu generisch (z.B. 'fonction') sind, um durch sie literarische Themen zu beschreiben, weshalb diese für das Vokabular nicht berücksichtigt wurden. Ergänzt wurden die Begriffe um solche Themenkonzepte, die bei der manuellen Annotation der Sekundärliteratur zusätzlich aufgedeckt wurden, um fehlende Konzepte beim Labeling der Topics sowie um thematische Schlagworte aus der Bibliographie (vgl. Martin et al. 1977), wenn diese anderweitig nicht repräsentiert waren. Das Vokabular ist nun konsolidiert, kann aber auch in Zukunft bei Bedarf erweitert werden.

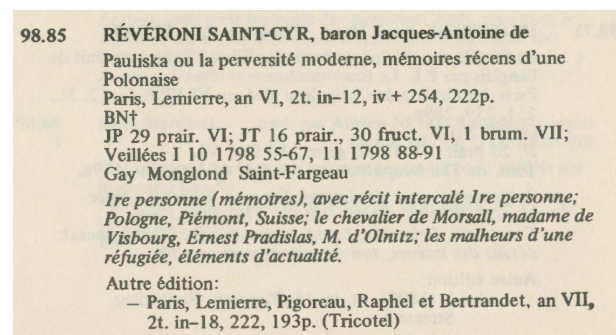
Um die multilinguale Vergleichbarkeit zwischen französischsprachigen Primärtexten und deutschsprachiger Sekundärliteratur zu gewährleisten, und im Sinne der Anschlussfähigkeit an und Interoperabilität mit anderen Datenbeständen, werden die Themenkonzepte auf einen Normdatensatz (Wikidata) gemappt, wodurch das kontrollierte Vokabular konsolidiert und multilingual erfasst ist (siehe Abb. 4).¹⁰

Unser Ziel: ein Wissensnetzwerk der Literaturgeschichtsschreibung



Dieser Graph lässt sich sodann auch über einen SPARQL-Endpoint abfragen (s. Abb.6). Ausgehend von der Beobachtung, dass das Themenkonzept "Reise" in den Bibliographie-Daten bei immerhin 14,7 % der Einträge vermerkt ist, ließe sich beispielsweise fragen, in welchen Werken auch laut Topic Modeling das mit dem Themenkonzept "Reise" verbundene Topic als dominantes Topic

Für das Werk *Jacques le fataliste* (1778) von Diderot stimmen Bibliographie-Daten und Topic Modeling-Ergebnisse im Hinblick auf das Themenkonzept “voyage” überein.



SPARQL-Abfragen zu den Ergebnissen des Topic Modelings und/oder der bibliographischen Schlagworten ermöglichen es, auch weniger bekannte Werke zu spezifischen Themen zu er-

mitteln. Zudem zeichnen sich Muster an Themenkomplexen im Zeitverlauf ab. Für das Thema “voyage” innerhalb der bibliographischen Daten zeigt sich eine (auch statistisch signifikante) ansteigende Entwicklung (vgl. Abb. 8).

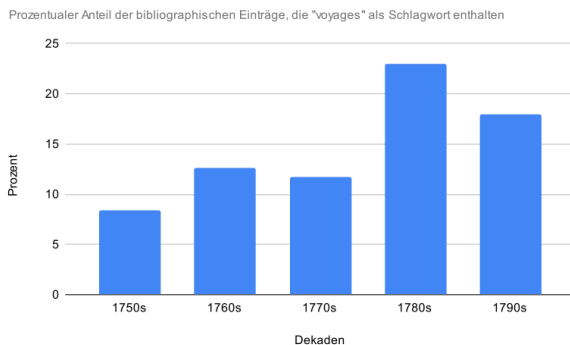


Abb. 8: Prozentualer Anteil der Werke aus der *Bibliographie du genre romanesque français 1751-1800* (Martin et al. 1977), die das Schlagwort “voyages” enthalten.

Eine Erklärung für den Anstieg der Themenkategorie “voyage” in den 1780er und 1790er Jahren könnte sein, dass viele Autor:innen die Handlung ihrer Werke aus politischen Gründen in andere Länder “verlegen” und zudem, dass der Themenkomplex der Reise im Kontext von Emigration im Zuge der politischen Ereignisse zunehmend in Romanen verhandelt wird.

Beispiele hierfür wären *Le roi Guiot* (1791) von Jean Vesque de Puttelange, dessen Protagonist in ferne Königreiche reist, um dort einen neugierigen Blick auf von der absolutistischen Monarchie abweichende politische Herrschaftssysteme zu werfen oder auch der erwähnte Roman *Pauliska ou la perversité moderne* (1798), in dem die Protagonistin durch Polen irrt. Das Thema Flucht und Emigration verweist auf die politische Realität in Frankreich nach der Französischen Revolution (vgl. Van Crugten-André 2001) und ist Anlass zu Reflexionen über Gesellschaftsformen (vgl. Pageaux 1968: 205–14).

Insgesamt ist das Thema “Reise” derzeit laut Topic Modeling im Romankorpus in 14,13% der Werke als dominantes Topic vertreten, in den bibliographischen Daten in 14,74% der Werke. Ein makrostruktureller Blick zeigt demnach in diesem Beispiel eine vergleichbare Größenordnung der thematischen Aussage über das gesamte Korpus hinweg, auch wenn in der Bewertung der Einzelwerke nicht immer Kongruenz besteht.

Fazit und Perspektiven

Das Projekt MiMoText modelliert die Geschichte des französischen Romans der zweiten Hälfte des 18. Jahrhunderts in Form von RDF-Tripeln in Wikibase als Knowledge Graph. Im Zuge eines Pilotprojekts wurden in einem ersten Schritt aus bibliographischen Daten und aus einem Romankorpus Relationen zwischen Werken und Themen extrahiert, die in Form von Tripeln in eine eigene Wikibase-Instanz eingelesen wurden. Zur Modellierung der Themen des französischen Romans der Aufklärung wurde ein kontrolliertes Vokabular erstellt, welches auf Wikidata gemappt wurde, um anschlussfähig an die Linked Open Data Cloud zu sein. Die Ergebnisse der Informationsextraktion aus den Romanen (mithilfe von Topic Modeling) und der Informationsextraktion aus

bibliographischen Daten können nun per SPARQL-Endpoint abgefragt werden.

Zu den nächsten Schritten gehört neben der Extraktion weiterer Statements über quantitative Romananalysen der Import von Themen-Statements aus dem dritten Typus von Informationsquellen (Fachliteratur) in unsere Wikibase-Instanz.¹²

Fußnoten

1. Als Beispiel sei hier die Initiative der GND genannt, die eigenen Daten in Wikidata oder zumindest in einer Wikibase-Instanz (die Software hinter Wikidata) zu integrieren: <https://blog.wikimedia.de/2020/03/04/wikibase-und-gnd/>, letzter Zugriff: 30.11.2021. Zum Begriff “Linked Open Data” vgl. (Berners-Lee et al. 2006: 1–130).
2. Wir nutzen die Quellen *Wikisource*, *Ebooks libres et gratuits*, *GoogleBooks*, *Rousseau Online* und *Frantext*. Diese Metadaten-tabelle dokumentiert die Korpuszusammensetzung und wird parallel zum Korpusaufbau laufend aktualisiert: <http://doi.org/10.5281/zenodo.5040855> /. https://github.com/MiMoText/roman18/blob/master/XML-TEI/xml-tei_metadata.tsv, letzter Zugriff: 30.11.2021.
3. https://github.com/MiMoText/roman18/blob/master/Python-Scripts/tei2txt_run.py, letzter Zugriff: 30.11.2021.
4. In der Regel werden die Topics durch ein Element des Themenvokabulars repräsentiert, in wenigen Fällen erscheint die Repräsentation durch zwei Themenkonzept-Label treffender.
5. Die Datengrundlage dieses Topic Modeling Durchgangs ist unter folgendem Release zu finden: (Klee/Röttgermann 2020).
6. Zu möglicherweise fehlenden Werken vgl. Dawson 1978. Dawson benennt auch das Desiderat eines Themenindex.
7. An das Scannen sowie OCR schlossen sich das Generieren von Trainingsdaten sowie die Auszeichnung aller Einträge über ein Machine Learning-Verfahren (CRF) in XML an. Diese bildeten die Datengrundlage für die anschließende Modellierung der Einträge in RDF (vgl. Lüscho 2020), bei der jedoch die einzelnen Keywords noch nicht semantisch modelliert worden sind.
8. Weitere Kategorien umfassen die Erzählform, den Ort der Handlung, die Figuren des Romans, die Tonalität/den Stil des Werks.
9. Ausgewählt wurden zum einen Strings mit mindestens acht Vorkommen in der Bibliographie und darüber hinaus solche mit einer besonderen Relevanz in den anderen Informationsquellen und für die literaturgeschichtliche Domäne insgesamt wie zum Beispiel der String ‘robinsonade’.
10. Zur Dokumentation der Liste vgl. Klee/Hinzmann 2021.
11. Als dominante Topics bezeichnen wir jeweils diejenigen Topics, die in einem Werk unter den 5 Topics mit den höchsten Wahrscheinlichkeiten sind (siehe oben).
12. Hierfür werden mit INCEpTION Aussagen in literaturgeschichtlichen Fachtexten annotiert, die in das Wissensnetzwerk eingespeist werden und zugleich als Trainingsdaten für die automatische Extraktion von Thementexten dienen.

Bibliographie

Berners-Lee, Tim / Hall, Wendy / Hendler, James A. / O’Hara, Kieron / Shadbolt, Nigel / Weitzner, Daniel J. (2006): “A Framework for Web Science”, in: *Foundations and Trends in Web Science* 1.1.: 1–130. 10.1561/1800000001.

Blei, David M. (2011): “Introduction to Probabilistic Topic Models”, in: *Communications of the ACM* 55.4.: 1–16.

Burnard, Lou (2014): *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. Marseille: Encyclopédie Numérique.

Dawson, Robert L. (1978). “The Martin, Mylne, Frautschi Bibliographie Du Genre Romanesque Français”, in: *Eighteenth-Century Studies*, 11.4., 497–508. 10.2307/2737969.

Delon, Michel (2007): *Dictionnaire européen des Lumières*. Paris: PUF.

Ehrlinger, Lisa / Wöß, Wolfram (2016): „Towards a Definition of Knowledge Graphs“, in: *SEMANTiCS (Posters, Demos, SuCESS)* <http://ceur-ws.org/Vol-1695/paper4.pdf>.

Heiden, Serge (2010): *The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme*. <http://halshs.archives-ouvertes.fr/halshs-00549764/en> [letzter Zugriff: 30.11. 2021].

Klee, Anne / Hinzmann, Maria (2021): *MiMoText/vocabularies* [Data Set]. https://github.com/MiMoText/vocabularies/blob/main/thematic_vocabulary.tsv [letzter Zugriff: 30.11.2021].

Klee, Anne / Röttgermann, Julia (2020): *Doing Topic Modeling on French 18th Century Novels in the Context of MiMoText Project* [Data Set] <https://github.com/MiMoText/topicmodeling> [letzter Zugriff: 30.11.2021].

Lüschow, Andreas (2020): “Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane”, in: *Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, 80–84 10.5281/zenodo.3666690.

Martin, Angus / Mylne, Vivienne / Frautschi, Richard L. (1977): *Bibliographie du genre romanescque français, 1751–1800*. London: Mansell.

McCallum, Andrew Kach (2002): *MALLET: A Machine Learning for Language Toolkit*. <http://mallet.cs.umass.edu>.

Pageaux, Daniel-Henri (1968): “Voyages romanescques au siècle des Lumières”, in: *Études littéraires*, 1.2.: 205–214. 10.7202/500020ar.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Barbelt, Nico / Wehner, Maximilian (2019): “OCR4all -- An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings”, in: *ArXiv:1909.04032* [Cs].

Röttgermann, Julia (2021): *Collection de romans français du dix-huitième siècle (1750–1800) / Eighteenth-Century French Novels (1750–1800) [dataset]*. Release v0.2.0 10.5281/zenodo.5040855.

Mithilfe von Machine Reasoning alchemische Decknamen entschlüsseln

Lang, Sarah

sarah.lang@uni-graz.at
Universität Graz, Austria

Einleitung

Aufgrund ihrer poetischen und auch manchmal geradezu abstrusen Sprachverwendung wurde die Alchemie lange Zeit nicht als Wissenschaft ernst genommen. Ihre eigentümliche Fachsprache ist durch eine Vielzahl eigentümlicher kryptographischer Stilmittel gekennzeichnet, darunter die sogenannten *Decknamen*. Dieser Beitrag schlägt einen Workflow vor, wie solche *Decknamen* im Kontext digitaler Edition halbautomatisch gefunden und annotiert werden können. Außerdem wird vorgeschlagen, sie in einem Wissensorganisationssystem inhaltlich zu erschließen, mit dem später deren Interpretation unterstützt werden kann.¹

Zielsetzung und Problemdefinition

Der Übergang von alchemischer Sprache zu chemischer Nomenklatur ist allgemein als zentrales Element der Chemiegeschichte anerkannt. Die *Méthode de nomenclature chimique* (1787) wird mitunter sogar erst als die Geburtsstunde der eigentlichen Chemie im Zuge einer ‘Chemical Revolution’ angesehen (Lefèvre 2018). Neuere Studien der Alchemiegeschichte zeigen jedoch, dass auch hinter der vormals häufig als obskur und sinnfrei bezeichneten Sprache der Alchemie valide chemische Erkenntnisse standen (Principe 2013). Die alchemische Sprache selbst bleibt aber wenig systematisch erforscht. Lange war die Forschung durch die von Umberto Eco popularisierte Theorie von der ‘hermetischen Semiose’ (Eco 2016) dominiert, also der Vorstellung, alchemische Decknamen lösten einen endlosen Semioseprozess aus, der nie zu einem Ende kommt, weil hinter dem sogenannten ‘alchemischen Geheimnis’ kein Inhalt stecke (Lippmann 1919; Principe 1992; Newman 1996). Neuere Studien im Zuge der ‘New Historiography of Alchemy’ haben allerdings gezeigt, dass gerade diese Grundannahme jeglicher früherer Theorien zur alchemischen Sprache vor deren Hintergrund nicht mehr haltbar ist (Newman / Principe 1998; Principe / Newman 2001; Martín-Torres 2011). Das Konzept der ‘alchemischen Sprache’ muss daher einer grundlegenden Revision unterzogen werden. Dieser Beitrag schlägt eine Möglichkeit vor, wie digitale Methoden dazu verwendet werden können.

Die verrätselte, poetische Sprache der Alchemie inspiriert seit jeher esoterische Interpretationen. Unter einer Vielzahl unterschiedlichster kryptographischer Stilmittel² sind vor allem die sogenannten Decknamen bekannt, die – oftmals in Form mythologischer Gestalten – von Alchemist:innen und Chymiker:innen anstatt chemischer Formeln verwendet wurden. Dieses Stilmittel kann als (mehr oder weniger metaphorisches) Wortsubstitutionsverfahren verstanden werden, bei dem Begriffen eine alchemische Fachbedeutung zugeschrieben wird, die mit deren sonstiger linguistischer Bedeutung nichts zu tun haben. Dabei entsteht Polysemie, wenngleich die ursprüngliche und die alchemische Bedeutung für gewöhnlich eine gewisse Übereinstimmung in ihren Eigenschaften aufweisen. Diese Sonderbedeutung dieser in linguistischer Hinsicht oftmals allgemein bekannten Wörter (z.B. „der grüne Löwe“) erschließt sich allerdings nur all jenen, die über das entsprechende enzyklopädische Kontextwissen verfügen.

Seit den 1990er Jahren hat sich infolge der bereits genannten Pioniersarbeiten von Lawrence Principe und William Newman in der Alchemieforschung die sogenannte ‘New Historiography of Alchemy’ durchgesetzt – ein Forschungsansatz, der versucht alchemische Decknamen chemisch zu lesen und die Lesung über experimentalarchäologische chemische Experimente zu

verifizieren.³ Dieser naturwissenschaftliche Zugang zu alchemischen Texten, die uns aufgrund ihrer eigenartigen Sprache heutzutage mitunter fremd, gar unwissenschaftlich anmuten, hat dazu beigetragen, nicht nur unsere Vorstellung von alchemistischen Sprechweisen, sondern das Alchemiebild im Allgemeinen zu revidieren, indem er chemische Leistungen von Alchemist:innen und Chymikerinnen feststellbar macht. Doch neben der chemischen Nachstellung historischer Rezeptvorschriften können auch digitale Methoden zur Erforschung alchemischer Decknamen beitragen: Das vorliegende Projekt hat sich mit der Frage nach der Funktionsweise alchemischer Sprache befasst. Es hatte zum Ziel, eine digitale Methode zur automatisierten semantischen Annotation und halbautomatisierten Disambiguierung des Stilmittels der sogenannten Decknamen zu entwickeln. Es beleuchtete das Konzept der alchemischen Sprache, indem es sie im Kontext eines digitalen Korpus mithilfe von Machine Reasoning analysierbar machte. Das Korpus der Druckwerke des deutschen Iatrochymikers Michael Maier (1568–1622) (Leibenguth 2002; Tilton 2003) wurde mithilfe automatisierter Annotation und Disambiguierung in Bezug auf seine Decknamenverwendung untersucht; einerseits mithilfe eines Semantic Web Wissensorganisationssystem unter der Verwendung von SKOS und RDFS sowie andererseits mithilfe automatisierter Annotation semantischer Ambiguität. Teil der Arbeit war neben der Entwicklung eines digitalen Analysetools auch, Einsichten in die Funktionsweise alchemischer Sprache zu vermitteln. Damit steht die Arbeit in der Tradition der durch William Newman und Lawrence Principe begründeten ‘New Historiography of Alchemy’, die sich zum Ziel gesetzt hat, durch Aufschlüsseln alchemischer *Decknamen* überholte Vorstellungen über Alchemie/Chymie und ihre Sprache zu revidieren.

Stand der Forschung

Michael Maiers Emblemwerk *Atalanta fugiens* (1617/18) erfreut sich seit den 1960er Jahren erhöhter Aufmerksamkeit der Forschungsgemeinschaft, doch der Rest seines Korpus bleibt weitgehend unerschlossen. Um die Jahrtausendwende kritisierte Erik Leibenguth große Teile der Forschungsgeschichte zu Maier als ‘Klitterpublizistik’.⁴ Angesichts des vermehrten Auftretens solcher Beanstandungen tritt die textuelle Tiefenerschließung des Maier’schen Œuvres als dringendes Desiderat der Maierforschung zutage. Doch stellt das Gesamtopus Maier’scher Drucke mit seinen rund 3500 Seiten lateinischen Texts eine ausgesprochen große Aufgabe dar, sofern dieses allein durch menschliches *close reading* bewältigt werden soll. Text Mining könnte die Antwort auf die beklagten Missstände sein.

Alchemische Texte werden mehr und mehr digitalisiert, so auch die Maiers.⁵ Die Digital Humanities beginnen sich der Alchemie zuzuwenden, was sich in Bezug auf Michael Maier besonders am jüngst abgeschlossenen *Furnace and Fugue*-Projekt gezeigt hat, bei dem eine multimediale digitale Edition der *Atalanta fugiens* erstellt wurde (Nummedal / Bilak 2019). Einige digitale Editionen wurden bereits für alchemische Texte umgesetzt, doch wurde das digitale Medium dabei zumeist lediglich als leicht zugängliche Plattform angesehen, über die Alchemieforschung schneller rezipiert werden kann und öffentlichkeitswirksamer wird. In seiner Prognose bezüglich der Zukunft der Alchemieforschung erwähnt Martínón-Torres zwar ‘digital resources’, aber nur in Bezug auf ‘primary sources’. Die Möglichkeit, digitale Methoden zur *Analyse* statt zur bloßen Bereitstellung alchemischer Texte zu verwenden, scheint ihm 2011 noch nicht denkbar (Martínón-Torres 2011, 233).

Furnace and Fugue enthält bereits einige digitale Tools, die eine dynamische Interaktion mit den Werkinhalten Maiers erlauben. Diese dienen aber vielmehr dazu, als interaktive Ergänzungen zu rein klassischen *close reading*-Ansätzen verwendet zu werden. Methoden der quantitativen Textanalyse bedient sich *Furnace and Fugue* allerdings noch nicht. Die Interpretation von Begriffen bleibt weiterhin ausschließlich den wissenschaftlichen Ansätzen vorbehalten. Vorliegendes Projekt setzte sich zum Ziel, digitale Methoden zur automatisierten Disambiguierung von alchemischen *Decknamen* im Korpus der Druckwerke Michael Maiers anzuwenden.

Methode und Forschungsfragen

Automatisierte Annotation wurde unter Hinzunahme eines formalen Wissensmodells zu einer quantitativen Textanalyse bestimmt durch eine Wissensressource im Vorhinein spezifizierter Begriffe alchemischen Fachvokabulars und ihrer Beziehungen untereinander verwendet.⁶ Das Wissensorganisationssystem *Knowledge Organization System* wurde mit dem Semantic Web-Vokabular SKOS (*Simple Knowledge Organization System*) und RDFS (*Resource Description Framework Schema*) kodiert. Teilautomatisiert gefunden werden konnten alchemische *Decknamen* mithilfe von Zipf’s law, da es sich dabei um Wörter vom Typ der *Realia* handelt.⁷ In weiteren Schritten mussten diesen im KOS Properties zugeordnet werden, die für die Interpretation relevante Zusatzinformationen enthalten.

Zur Funktionsweise alchemischer Sprache wurden in der Vergangenheit bereits einige Überlegungen beigetragen, doch sind diese meist theoretischer Natur und gehen über Forschungstypen kaum hinaus, weswegen sie für konkrete Textanalysen nicht fruchtbar zu machen sind (Schütt 1994; Duncan 1981). Bereits bestehende Ressourcen zu alchemischen Begrifflichkeiten waren zur Information über Alchemie, aber nicht zur Annotation konkreter alchemischer Texte gedacht, wie z.B. das Alchemie-Lexikon (Priesner / Figala 1998) oder der Alchemie-Thesaurus der HAB Wolfenbüttel, der der Verschlagwortung und Erschließung alchemischer Buchbestände diente (Frietsch 2017, 2021). Die bisher vorhandenen Ressourcen erlaubten zwar ein Einarbeiten oder Vorinformieren über Alchemie und gewisse alchemische Konzepte durch Einlesen, waren aber nicht für eine automatisierte Annotation konkreter alchemischer Texte geeignet. Die dort vorkommenden Einträge sind häufig allgemeine Überbegriffe, die einerseits zur automatisierten Annotation von Texten zu unkonkret sind, andererseits finden sich die darin enthaltenen Einträge nicht als Zeichenketten im konkreten Text.⁸

Out-of-the-box verfügbare Text-Mining- oder Distant-Reading-Methoden wie etwa Topic Modelling haben sich als nur sehr beschränkt zur Beantwortung alchemiegeschichtlicher Fragestellungen geeignet herausgestellt, besonders wenn es sich um die Analyse komplexer Decknamen handelt, wie z.B. im Fall des ‘Mercurius’. Methoden, die auf Wortzählungen (*bags of words*) basieren, sind nicht in der Lage, solche Konzepte adäquat zu disambiguieren, denn “quantitative research ... provides *data*, not interpretation” (Moretti 2005). Das erstellte Wissensorganisationssystem wird also verwendet, um Vorkommnisse der darin enthaltenen Begriffe im Korpus Maiers automatisiert zu annotieren. Dadurch entsteht ein Begriffsnetzwerk, das es wiederum erlaubte, Konkordanzen zu erstellen, in denen nicht nur der zu betrachtende Begriff als *Keyword in Context* zugänglich ist, sondern auch andere bereits im vorherigen Schritt annotierte Konzepte vorkommen. So entsteht ein ‘Decknamen-Kontext’: Dabei handelt es sich

um ein *Keyword in Context (KWIC)*, bei dem nicht der linguistische Kontext, sondern die umliegenden Decknamen aufgeschlüsselt werden. So treffen wir z.B. im Text auf den String ‘Mercurium’, welcher auf das Konzept ‘Mercurius’ verweist. Dieses kann in alchemischer Literatur sowohl chemische als auch mythologische oder historische Kontexte haben. Das *Decknamen-KWIC* zeigt an einem vereinfachten Beispiel, welche Kontexte im umliegenden Text besonders vorwiegen und kann somit eine mögliche Disambiguierung anbieten. Folgend ein Beispiel einer Konkordanzansicht in zu Anschauungszwecken vereinfachtem XML:

```
<example ref="Maier.Arcana.191">
  tum, aut per <deckname>Lunam</deckname>, <deckname>argentum</deckname>,
  per <termInQuestion>Mercurium</termInQuestion>
  <deckname>hydrargyrum</deckname>, per <deckname>Saturnum</deckname>
  <deckname>plumbum</deckname>, per <deckname>Iovem</deckname>, <deckname>stannum</deckname>,
  per <deckname>Martem</deckname> <deckname>ferum</deckname>, communia intellexis
</example>
```

Hier zeigt sich an einem einfachen Beispiel, dass der in Frage stehende Term ‘Mercurius’ hier die Kontexte ‘Planeten’ (*Luna, Saturn, Iupiter, Mars*), ‘Metalle’ (*argentum, hydrargyrum, plumbum, stannum, ferrum*) und ‘Mythologie’ (*Saturn, Iupiter, Mars*) aufweist. Da aber alle umliegenden *Decknamen* Planeten oder Metalle sein können, die mit Planetennamen angesprochen werden, ist es sehr wahrscheinlich, dass mit ‘Mercurius’ hier die chemische Substanz gemeint ist, keine mythologische Figur.⁹ Wenn die Annotationen vorhanden sind, kann diese Überprüfung maschinell erfolgen. Weiterführend konnten zudem die Eigenschaften besonders wichtiger Begriffe in RDFS-Tripeln modelliert werden, wodurch Verbindungen zu anderen Konzepten automatisch erkannt werden können. Folgend ein Beispiel zu der Verbindung von Rot und Gold, die in der Alchemie den Zusammenhang zwischen dem roten Stein der Weisen und der Goldherstellung beschreibt, repräsentiert als RDFS-Tripel:

```
:PhilosophersStone :hasColor :red.
:red :hasChemicalProperty :tints.
:tints :givesPhysicalProperty :citrinitas.
:Gold :hasColor :citrinitas.
```

Die hier vorzustellende jüngst abgeschlossene Dissertation hat sich mit der Frage nach der Funktionsweise alchemischer Sprache beschäftigt. Dabei setzte sie sich zum Ziel, eine digitale Methode zur automatisierten polysemantischen Annotation und halb-automatisierten Disambiguierung des Stilmittels der sogenannten Decknamen zu entwickeln. Ein solcher Algorithmus beleuchtet das Konzept der alchemischen Sprache aus einer neuen Perspektive, indem er sie im Kontext eines digitalen Korpus mithilfe von Machine Reasoning im Sinne einer Sonderform der distributionellen Semantik analysierbar macht. Sie hatte sich zur Aufgabe gesetzt, das Korpus der Druckwerke des deutschen Iatrochymikers Michael Maier (1568–1622) mithilfe automatisierter Annotation und Disambiguierung in Bezug auf seine Decknamenverwendung zu untersuchen; einerseits mithilfe eines Semantic Web Wissensorganisationssystems unter der Verwendung von SKOS und RDFS sowie andererseits mithilfe automatisierter Annotation semantischer Ambiguität. Die Dissertation wurde im Mai 2021 abgeschlossen. Das Ergebnis wird als digitale Edition publiziert werden.

Fußnoten

1. Zur hier vorgestellten Methode ausführlicher siehe: Lang 2020, Lang 2021 und Lang 2022a, 2022b, 2022c.
2. Hier wird bewusst von ‚kryptographischen Stilmitteln‘ gesprochen, da sie verschlüsselnd wirken, allerdings nicht mit den mono- oder polyalphabetischen Substitutionsverschlüsselungen vergleichbar sind, die man sich typischerweise unter dem Begriff „Kryptographie“ vorstellt. Substitutionschiffren gibt es auch in der Alchemie. Lang / Piorko 2021 stellt ein solches Beispiel vor, das mittlerweile entschlüsselt werden konnte, vgl. <https://theconversation.com/deciphering-the-philosophers-stone-how-we-cracked-a-400-year-old-alchemical-cipher-167900> [letzter Zugriff am 1. Dezember 2021].
3. Allerdings wird betont, dass das Methodenarsenal der Experimental History of Science in seiner Anwendung aufwändig ist, weswegen es nicht auf alle Fälle alchemischer Sprache sinnvoll anwendbar ist: “Not every historical project needs or would even benefit from the inclusion of an experimental component, and not every textual process or experiment is worth the often considerable time it takes to rework it.” (Fors / Principe / Sibum 2016, 96)
4. Leibenguth stellt fest: “Charakteristisch für die Maierforschung sind mit wenigen [...] Ausnahmen eine Abhängigkeit und teils sinnentstellende Rezeption von Sekundärquellen sowie die weitgehende Unkenntnis der lateinischen Primärtexte (8).” Auch Wels pflichtet ihm in der Sache bei: “Maiers Schriften gehören zu jenen Werken, die viel betrachtet, aber wenig gelesen werden” (149). Vgl. Leibenguth 2002; Wels 2010.
5. Seine *Arcana* (1614) wurden beispielsweise im Zuge des EEBO-Projektes in TEI-XML transkribiert oder aber *Symbola* (1617) und *Examen* (1617) an der Herzog-August-Bibliothek Wolfenbüttel im Zuge des Aufbaus des Alchemie-Portals, die ebenfalls als TEI-XML mit den zugehörigen digitalen Faksimiles online zur Verfügung gestellt werden. Vgl. Maier 2009, 1617b, 1617a; Feuerstein-Herz 2017.
6. Die computerlinguistische Erforschung der automatisierten Annotation stellte dabei allerdings nicht das primäre Anliegen dar. Möglichkeiten der Automatisierung sollten evaluiert und genutzt werden, doch das Hauptziel der Arbeit bestand in der Wissensrepräsentation.
7. Realia können mithilfe vom Zipf’schen Gesetz in Bags-of-words lokalisiert werden, da sich darin Synsemantika am h-point des Graphen abspalten (vgl. Popescu / Macutek / Altmann 2009). Werden dann noch häufige Autosemantika als Stopwords von der Analyse ausgeschlossen, bleibt nur mehr eine überschaubare Anzahl an types zu kontrollieren.
8. So ergab beispielsweise eine automatisierte Annotation eines Beispielkorpus aus Maiers Werk, dass von den 99 Einträgen des HAB-Thesaurus gerade einmal 19 überhaupt in Maier angeführt wurden, da die meisten Begriffe nur zum Sprechen über Texte geeignet sind und abstrakte Forschungsthemen bezeichnen, die natürlich im Klartext der Quellen nicht so bezeichnet stehen. Es werden daher einerseits viele Begriffe des HAB-Thesaurus in der Quelle nicht angeführt, andererseits bleiben viele erklärungsbedürftige Konzepte des Maier-Textes unerklärt, da sie wohl zu speziell gewesen wären, um sie im besagten Thesaurus abzubilden.
9. Nicht alle Beispiele sind so trivial wie dieses, bei dem auch im close reading bereits auf den ersten Blick kein Zweifel besteht. Die entsprechenden mythologischen Figuren werden bei Maier allerdings tatsächlich häufig diskutiert, weswegen die Disambiguierung nicht immer so eindeutig ist.

Bibliographie

- Duncan, A. M.** (1981): "Styles of Language and Modes of Chemical Thought" in: *Ambix* 28/2: 83–107.
- Eco, Umberto** (2016): "Il Discorso Alchemico E Il Segreto Differito" in: *I Limiti Dell'interpretazione (Prima Edizione 1990)*, 97–116. La nave di Teseo.
- Feuerstein-Herz, Petra** (2017): "Alchemie Portal der Herzog August Bibliothek Wolfenbüttel." *Herzog August Bibliothek Wolfenbüttel*. <http://alchemie.hab.de>.
- Fors, Hjalmar / Principe, Lawrence / Sibum, Otto** (2016): "From the Library to the Laboratory and Back Again: Experiment as a Tool for Historians of Science" in: *Ambix* 63/2: 85–97.
- Frietsch, Ute** (2017): "Alchemie Thesaurus." *Herzog August Bibliothek Wolfenbüttel*. 2017. <http://alchemie.hab.de/thesaurus>.
- Frietsch, Ute** (2021): "Obscurum Vocabulum: Begriffe der frühneuzeitlichen Alchemie und der Alchemie-Thesaurus der Herzog August Bibliothek" in: Feuerstein-Herz, Petra / Frietsch, Ute (eds.): *Alchemie – Genealogie und Terminologie, Bilder, Techniken und Artefakte. Forschungen aus der Herzog August Bibliothek*. Harrassowitz, Wiesbaden.
- Lang, Sarah** (2022a): "A Machine Reasoning Algorithm for the Digital Analysis of Alchemical Language and its 'Decknamen'", in: *Ambix* (Special Issue).
- Lang, Sarah** (2022b): „Digital Scholarly Editions of alchemical texts as tools for interpretation“ in: Klug, Helmut / Bleier, Roman (eds.): *Digitale Edition in Österreich*. Graz.
- Lang, Sarah** (2022c): "Vom ‚Wissen in Buchform‘ zum formalen Wissensmodell. Digitale Aufbereitung wissenschaftshistorischer Drucke am Beispiel der Alchemica Michael Maiers", in: Hegel, Philipp / Krewet, Michael (eds.): *Wissen und Buchgestalt*. Wiesbaden, Harrassowitz.
- Lang, Sarah / Piorko, Megan** (2021): „An alchemical cipher in a shared notebook of John and Arthur Dee (Sloane MS 1902)“ in: *Proceedings of the 4th International Conference on Historical Cryptology HistoCrypt 2021*: <https://ecp.ep.liu.se/index.php/histocrypt/article/view/161> [letzter Zugriff am 1. Dezember 2021].
- Lang, Sarah** (2021): „Digitale Erschließungsmethoden für alchemische Texte am Beispiel der Symbola Aureae Mensae Michael Maiers“ in: Feuerstein-Herz, Petra / Frietsch, Ute (eds.): *Alchemie – Genealogie und Terminologie, Bilder, Techniken und Artefakte. Forschungen aus der Herzog August Bibliothek*. Harrassowitz, Wiesbaden.
- Lang, Sarah** (2020): „Digitale Annotation alchemischer Decknamen. Die Allegoriae werden uns nit mehr verborgen seyn“ in: Nantke, Julia / Schlupkothen, Frederik (eds.): *Annotations in Scholarly Editions and Research. Functions, Differentiation, Systematization* 201–219. De Gruyter. <https://doi.org/10.1515/9783110689112-010> [letzter Zugriff am 1. Dezember 2021].
- Lefèvre, Wolfgang** (2018): "The Méthode de Nomenclature Chimique (1787): A Document of Transition" in: *Ambix* 65/1: 9–29.
- Leibenguth, Erik** (2002): *Hermetische Philosophie des Frühbarock. Die "Cantilenae Intellectuales" Michael Maiers. Edition mit Übersetzung, Kommentar und Bio-Bibliographie*. Tübingen: Niemeyer.
- Lippmann, Edmund Oskar von** (1919): *Entstehung und Ausbreitung der Alchemie. Mit einem Anhang: Zur älteren Geschichte der Metalle. Ein Beitrag zur Kulturgeschichte*. Band 1. Berlin: Springer.
- Maier, Michael** (1617a): "Examen Fucorum Pseudo-Chymicorum". <http://diglib.hab.de/drucke/46-med-4s/start.htm>.
- Maier, Michael** (1617b): "Symbola Aureae Mensae". <http://diglib.hab.de/drucke/46-med-1s/start.htm>.
- Maier, Michael** (2009): "Arcana Arcanissima" (Digitale Edition der 1613-Ausgabe): Ann Arbor; Oxford (UK): Text Creation Partnership, 2008-09 (Eebo-Tcp Phase 1). <http://name.umdl.umich.edu/A06751.0001.001>.
- Martinón-Torres, Marcos** (2011): "Some Recent Developments in the Historiography of Alchemy" in: *Ambix* 58/3: 215–37.
- McCarty, Willard** (2004): "Modeling: A Study in Words and Meanings" in: Schreibmann, Susan / Siemens, Ray / Unsworth, John (eds.): *A Companion to Digital Humanities*. Oxford: Wiley Blackwell, 254–72.
- Moretti, Franco** (2005): *Graphs, Maps, Trees. Abstract Models for a Literary History*. London.
- Newman, William** (1996): "'Decknamen or Pseudochemical Language'? Eirenaeus Philaethes and Carl Jung" in: *Revue d'histoire Des Sciences* 49: 159–88.
- Newman, William / Principe, Lawrence** (1998): "Alchemy Vs. Chemistry: The Etymological Origins of a Historiographic Mistake" in: *Early Science and Medicine* 3/1: 32–65.
- Nummedal, Tara / Bilak, Donna** (2020): "Furnace and Fugue. A Digital Edition of Michael Maier's Atalanta Fugiens (1618) with Scholarly Commentary". <https://furnaceandfugue.org/>.
- Popescu, Ioan-Iovitz / Mačutek, Ján / Altmann, Gabriel** (2009): *Aspects of Word Frequencies*. Lüdenschied, RAM-Verlag.
- Priesner, Claus / Figala, Karin** (1998): "Vorwort der Herausgeber" in: Priesner, Claus / Figala, Karin (eds.) *Alchemie. Lexikon Einer Hermetischen Wissenschaft*. München: C.H. Beck, 7–11.
- Principe, Lawrence** (1992): "Robert Boyle's Alchemical Secrecy: Codes, Ciphers and Concealments" in: *Ambix* 39/2: 63–75.
- Principe, Lawrence** (2013): *The Secrets of Alchemy*. Chicago.
- Principe, Lawrence / Newman, William** (2001): "Some Problems with the Historiography of Alchemy" in: Newman, William / Grafton, Anthony (eds.): *Secrets of Nature: Astrology and Alchemy in Early Modern Europe*. Cambridge/Massachusetts: MIT Press, 385–432.
- Ruland, Martin** (1612): *Lexicon Alchemiae*. Frankfurt am Main.
- Schütt, Hans-Werner** (1994): "Sprachschichten der Alchemie" in: *Berichte zur Wissenschaftsgeschichte* 17: 89–99.
- Tilton, Hereward** (2003): *The Quest for the Phoenix. Spiritual Alchemy and Rosicrucianism in the Work of Count Michael Maier (1569–1622)*. Berlin/NY: De Gruyter.
- Wels, Volkhard** (2010): "Poetischer Hermetismus. Michael Maiers Atalanta Fugiens (1617/18)" in: Alt, Peter-André / Wels, Volkhard (eds.): *Konzepte des Hermetismus in Der Literatur der frühen Neuzeit. Berliner Mittelalter- und Frühneuzeitforschung*, Band 8. Göttingen: V & R Unipress, 149–94.

Multimodale KI zur Unterstützung geschichtswissenschaftlicher Quellenkritik Ein Forschungsauftritt

Muenster, Sander

sander.muenster@uni-jena.de
FSU Jena, Germany

Bruschke, Jonas

jonas.bruschke@uni-wuerzburg.de
JMU Wuerzburg, Germany

Kroeber, Cindy

cindy.kroeber@uni-jena.de
FSU Jena, Germany

Hoppe, Stephan

stephan.hoppe@kunstgeschichte.uni-muenchen.de
LMU Muenchen, Germany

Maiwald, Ferdinand

ferdinand.maiwald@uni-jena.de
FSU Jena, Germany

Niebling, Florian

florian.niebling@uni-wuerzburg.de
JMU Wuerzburg, Germany

Pattee, Aaron

aaron.pattee@lmu-muenchen.de
LMU Muenchen, Germany

Utescher, Ronja

ronja.utescher@uni-bielefeld.de
FSU Jena, Germany; U. Bielefeld, Germany

Zarriess, Sina

sina.zarriess@uni-bielefeld.de
U. Bielefeld, Germany

Einleitung

Fotografien und andere Abbilder von Architektur dienen in vielen historischen Wissenschaften als Quelle und Grundlage für fach- und theoriespezifische Untersuchungen. So werden zum Beispiel historische Fotoaufnahmen herangezogen, um den Zu-

stand eines Gebäudes zu rekonstruieren oder die Formensprache einer Epoche zu identifizieren. Ausgangspunkt dieser Szenarien aus Architektur-, Kunstgeschichte und Kulturwissenschaften ist eine durch Hilfsmittel der jeweiligen Fächer unterstützte Quellenrecherche und -kritik, auf die weitere Auswertungen und Verwendungen im wissenschaftlichen Kontext aufbauen.

Obwohl sich KI-basierte Methoden der *Computer Vision* in den letzten Jahren wesentlich weiterentwickelt haben, können diese den Prozess der Quellenrecherche und -kritik bisher allenfalls im Ansatz unterstützen, bspw. für die Exploration von Bildrepositorien oder das Retrieval von Bildern. Dies liegt zum einen daran, dass elementare diesbezügliche Vorgehensweisen zwar gut dokumentiert sind, WissenschaftlerInnen aber sehr individuell vorgehen. Zum anderen ist KI-Bildverarbeitung bisher wenig darauf ausgelegt, bildliche Inhalte multimodal zu kontextualisieren, d.h. verschiedene Quellengattungen wie Bilder und Texte zu kombinieren. Existierende Verfahren der Computer Vision extrahieren rein visuelle Merkmale und klassifizieren diese, während Texte oder Metadaten und darin enthaltenes Wissen wie bspw. Hinweise auf zeitliche Kontexte oder einzelne Motive nicht mit der Analyse verknüpft werden können.

Das BMBF-geförderte Projekt *HistKI* startete im Januar 2021 und will die Unterstützung und Modellierung von Bildquellenrecherche und -kritik als komplexe und grundlegende geschichtswissenschaftliche Arbeitstechnik durch multimodale KI-basierte Verfahren erforschen. Damit verbundene Teilfragen sind beispielsweise: Wie finden und beurteilen Historiker und andere Fachwissenschaftler Bildquellen? Welche generischen Vorgehensweisen und Teilproblemstellungen lassen sich hierfür identifizieren? Wie lässt sich dies mit KI-basierten Ansätzen befördern? Wie wirken sich KI-Techniken auf den geisteswissenschaftlichen Forschungsprozess aus?

Forschungsstand

Ausgangspunkt des Vorhabens ist der geschichtswissenschaftliche Forschungsprozess im Umgang mit Bildquellen. Für ein forschendes Handeln in den historischen Wissenschaften prägend sind einzelne Themenfelder (bspw. Kunst, Technik, Wirtschaft, Politik) sowie die darauf bezogenen Erkenntniszugänge. Leitendes Paradigma ist eine konstruktivistische Problemorientierung und damit eine quellenkritische, gegenstandsbezogene Analyse (Wengenroth, 1998, Reich, 2006). Methodologische Zugänge dazu liefern bspw. Hermeneutik, Semiotik (Holenstein, 1988, Zöllner, 2005) oder Phänomenologie (Pechtl, 2002). Grundsätzlich findet ein Zugang über Quellen statt und die damit verbundene Quellenkritik ist eine grundlegende Vorgehensweise historischer Forschung (Opgenoorth, 1997).

Modellierung forschenden Handelns: Eine Systematisierung von digitalem Forschungshandeln in den historischen Wissenschaften erfolgt bspw. aus Perspektive der eScience (Köhler et al., 2016) oder Informationswissenschaften, Ansätze zur Modellierung von Prozessabfolgen forschenden Handelns sind bspw. Forschungsprimitive (Münster and Terras, 2020, Kamposiori and Benardou, 2011). Demgegenüber stehen Ansätze zur sozialwissenschaftlichen Analyse des Informations- und Forschungsverhaltens (Münster et al., 2018, Ying et al., unpublished), bspw. der Science and Technology Studies (STS) u. a. anhand epistemischer Kulturen (Knorr-Cetina and Reichmann, 2015) oder der Wissenschaftsphilosophie anhand der Erkenntnisprozesse (Fleck, 1980, Popper, 1998). Damit gehen Ansätze zur Operationalisierung einerseits von der (1) „mechanistischen“ Idee aus, forschendes Handeln als Abfolge von operationalisierbaren Forschungsschrit-

ten (e.g. Forschungsprimitive, eScience) zu betrachten. Andererseits steht (2) die Vorstellung eines geisteswissenschaftlichen Forschungshandelns als „Black Box“ und erfahrungsbasiert, auf implizitem Wissen (Polanyi, 1966) basierend sowie der Wissensgenese als „serendipity“ (e.g. STS). Nicht zuletzt hinsichtlich der Lehrbarkeit haben sich in den bildbezogenen historischen Wissenschaften (3) semi-operationalisierte Zwischenformen etabliert – bspw. die bereits benannte Quellenkritik sowie speziell für Bildmedien die ikonologische Analyse (Panofsky, 1939). Trotz derartiger systematisierender Ansätze erfolgt eine Recherche und kritische Betrachtung von historischen Quellen in der Praxis hochgradig individuell und erfahrungsbasiert (Brieber et al., 2014, Münster et al., 2018). Vor diesem Hintergrund werden KI-basierte Ansätze derzeit vor allem zur Unterstützung von abgegrenzten Teilproblemen, wie Suche nach ähnlichen Bildern (Münster et al., 2019, Bell and Ommer, 2019), die Anreicherung von textuellen Metadaten (Lee and Münster, 2018) sowie Musteranalyse (Kohle, 2018, Klinke, 2016), eingesetzt.

Bildwissenschaftliche Zugänge: Eine Reihe von aktuellen Publikationen und Initiativen greifen das Zusammenspiel von bildbezogenen historischen Wissenschaften und KI-basierten Forschungsansätzen auf – bspw. der Band „Digital Art History“ (Kuroczynski et al., 2019) sowie das DFG-Schwerpunktprogramm „Das digitale Bild“ (Kohle, 2018). Trotz heterogener Ansätze zur Bildrecherche und Quellenkritik existiert eine Reihe allgemeiner Problemstellungen (Hoppe and Breitling, 2016). HistKI liefert mit seiner inhaltsgetriebenen und forschungsprozessorientierten Fragestellung eine wichtige und komplementäre Grundlage für diese Initiative.

Language & Vision: Neuere Ansätze der Bildverarbeitung aus dem Bereich des Deep Learning ermöglichen nicht nur eine bessere Objekterkennung, sondern erweisen sich als besonders geeignet für *transfer learning* an der Schnittstelle von Bild- und Sprachverarbeitung. Dabei werden zum Beispiel semantische Repräsentationen wie *word* oder *sentence embeddings*, die aus Texten gelernt werden, anhand von multimodalen Daten wie Bildbeschreibungen mit visuellen Repräsentationen angereichert. Damit ist das Vokabular eines Objekterkennungssystems aus der Bildverarbeitung, das typischerweise auf eine mehr oder weniger große Menge an Kategorien festgelegt ist, wesentlich erweitert und es können semantische Bezüge zwischen visuellen Objekten und einer großen Menge an Wörtern oder Phrasen erfasst werden. Beispielsweise gibt es aktuelle Untersuchungen zur automatischen Erkennung eines inhaltlichen Bezugs zwischen Bildern und Textpassagen, (Hessel et al., 2019). Diese Art von Modellierung stellt einen Schritt in Richtung der Herstellung eines gemeinsamen Kontexts von Bild und Text dar. Für die multimodale Extraktion von Informationen aus fachwissenschaftlichen Texten ist es jedoch notwendig, solche referentiellen Beziehungen zwischen Text- und Bild-Teilen auch feingliedriger zu erfassen (Utescher and Zarriß, 2021).

Segmentierung und Objekterkennung: Die Grundlagen für die Rekonstruktion aus historischen Fotografien bilden die analytischen Verfahren der Photogrammetrie, d.h. die Gewinnung zweidimensionaler Bildinformationen (Wiedemann et al., 2000). Photogrammetrische Verfahren liefern auch räumliche Relationen zwischen Fotografien und dreidimensionalen Objektgeometrien. Aus durch bildgebende Verfahren entstandenen Datensätzen lassen sich einfache Strukturen (Vosselman et al., 2004), aber auch komplexe Objekte wie Gebäude (Li et al., 2016, Agarwal et al., 2011) automatisch segmentieren sowie zuordnen (Martinovic et al., 2015, Hackel et al., 2016). Dabei können auch Rückschlüsse darauf gezogen werden, welche Bildteile welche Teile der 3D-Objektgeo-

metrien referenzieren (Xie et al., 2016, Vosselman et al., 2004). Maschinelles Lernen (ML) spielt eine zunehmend größere Rolle bei der Segmentierung von Bild und der Objekterkennung (Minaee et al., 2021, Jiao et al., 2019) sowie der Strukturerkennung in 3D-Daten (Guo et al., 2020). Insbesondere vor dem Hintergrund der Anwendung von ML Ansätzen auf historische Quellen bestehen spezifische Herausforderungen in zumeist sehr kleinen und qualitativ heterogenen Ausgangsdaten (Fiorucci et al., 2020).

Verortung und Exploration: Ein besonderes Verhältnis besteht zwischen Bildquellen und 3D-Modellen. Im BMBF-Projekt *HistStadt4D* wurden exemplarisch Methoden zur automatischen Verortung von Fotografien entwickelt (Maiwald and Maas, 2021) und diese Daten in einer virtuellen 4D-Forschungsumgebung zugänglich gemacht (Bruschke et al., 2018, Maiwald et al., 2019) sowie quantitative raumbezogene Analysemethoden evaluiert (Dewitz et al., 2019). Mit Hilfe von Verfahren des maschinellen Lernens sollen in *HistKI* zudem Objektquellen und Textquellen (z.B. Bildunterschriften) verknüpft werden, um in Zukunft eine detaillierte Kontextualisierung und Verortung der Fotografien und Texte zu erlauben und damit über bisherige Methoden des *distant viewing* (Arnold and Tilton, 2019) hinauszugehen.

Forschungsaufriß

Im Folgenden soll ein kurzer Überblick über erste Forschungsschritte im Projekt gegeben werden.

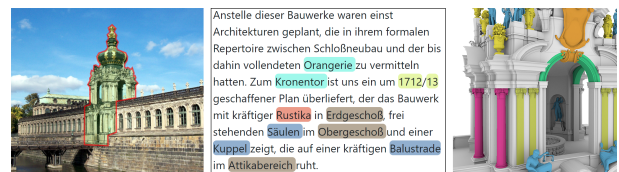


Abb. 1: Identifizierte Architekturelemente im Bild (links), im Text (mitte) und im 3D-Modell (rechts) am Beispiel des Kronentors des Dresdner Zwingers.

Identifikation von Forschungsszenarien

Der Einsatz von multimodalen KI-Techniken wird anhand von ausgewählten Szenarien untersucht, in denen Informationen aus Bildern, Texten und 3D-Modellen zur Beschreibung von Wissen über Architekturobjekte und städtebaulichen Ensembles für einen Analyseprozess miteinander verschrankt werden können. In Voruntersuchungen wurde dafür mit Hilfe qualitativer Expertenbefragungen und Workshops eine Reihe generischer Szenarien identifiziert (Kröber, 2021, Dewitz et al., 2019) und hinsichtlich einer Relevanz und Umsetzbarkeit priorisiert. Aus den ermittelten ca. 20 Szenarien wurden für eine erste Projektphase die medienübergreifende Identifikation von Objektbeschreibungen („Welche Bilder, Texte, 3D-Daten beschreiben dasselbe Objekt?“) sowie die Analyse von Beschreibungen (bspw.: „Wie kann die Datierung von historischen Bild- und Textdarstellungen von Objekten durch eine multimodale Validierung, bspw. anhand bereits datierter Medien, unterstützt werden?“) als Forschungsschwerpunkte ausgewählt.

Medienübergreifende Klassifikation

Die medienübergreifende Verarbeitung von 3D-Modellen, Bild- und Textquellen in einem möglichst universellen Format ist eine der zentralen Herausforderungen des Projekts. Eine diesbezüglich wichtige Voraussetzung ist es, medienübergreifend Elemente zu identifizieren und zu benennen. Die Entwicklung von domänenspezifischen Ontologien für architekturgeschichtliche Inhalte ist aktuell ein Schwerpunktthema und wird von einer Vielzahl von Initiativen vorangebracht – beispielhaft seien hier ICONCLASS¹, Getty Art & Architecture Thesaurus (AAT), als generische Ontologie Wikibase² sowie als übergreifende Referenzontologie CIDOC-CRM benannt. Als Grundgerüst für die strukturierende Beschreibung von Architekturelementen dient in unserem Projekt das AAT, welches trotz aktuell noch vorhandener Defizite hinsichtlich der deutschen Übersetzung durch eine umfassende und an einer kunsthistorischen Typisierung orientierte Struktur eine gute Ausgangsbasis für unser Projekt verspricht. Dabei wird von uns zunächst nur die gemessen am Gesamtvokabular kleine Untergruppe *architectural elements*³ verwendet. Die identifizierten Elemente im Text (einzelne Wörter oder Wortgruppen), Bild (polygonale Bildausschnitte) und 3D-Modell (einzelne Teilgruppenobjekte) werden den Konzepten des AAT zugeordnet (Abb. 1). Je nach Quellentyp sind dabei verschiedene Verfahren, wie z.B. semantic segmentation, named entity recognition (NER), und discourse parsing, notwendig, sowohl was die Identifizierung der Konzepte als auch die semantische Anreicherung betrifft. Als Nächstes müssen die identifizierten Konzepte zwischen den verschiedenen Quellen in einen Zusammenhang gebracht werden. Dazu werden die Identifier der AAT-Konzepte abgeglichen sowie meta-klassifiziert (z.B. Ionische Säule → Säule). Um mehrere Instanzen einer solchen Klasse zu unterscheiden (bspw. einzelne Säulen in einer Säulenreihe) werden die einzelnen Instanzen mit einer Identifikationsnummer versehen. Das AAT liefert ein detailliertes Vokabular zur Klassifizierung, reicht aber derzeit nicht aus, um alle Informationen, insbesondere Informationen aus Textquellen, vollständig abzubilden. Es müssen daher weitere Merkmale identifiziert werden, die das AAT nicht repräsentieren kann, wie z.B. Lagebeziehungen („Ostflügel“, „Westfassade“) oder Eigen-namen.

Multimodale Datenanreicherung

In einem weiteren Schritt werden Ansätze zur multimodalen Datenanreicherung und -validierung entwickelt. So werden beispielsweise im 3D-Raum in Relation zum 3D-Modell verortete Bilder verwendet, um eine im 3D-Modell vorliegende Strukturierung auf Bilddokumente zu übertragen (Niebling et al., 2018). Durch den räumlichen Zusammenhang zwischen Bild und Modell können Annotationen von bereits semantisch angereicherten 3D-Modellen bzw. ihrer einzelnen Bauteile auf die entsprechenden Bildquellen projiziert, sowie auch von bereits annotierten Bildquellen auf 3D-Modelle zurückübertragen werden. Der 3D-Raum bietet zudem erweiterte Möglichkeiten, Lagebeziehungen zwischen (Teil-)Objekten herauszufinden. Die erkannten Zusammenhänge und semantischen Beziehungen werden in einer Ontologie gespeichert.

Ausblick

Quo vadis? Während im aktuellen Projektstadium mit manuell klassifizierten Quellen gearbeitet wird, ist ein nächster Schritt ist die Untersuchung von Ansätzen zur automatisierten Erkennung und Annotation von Objektbestandteilen. Hier werden schwerpunktmäßig KI-basierte Modelle verwendet, die auf die jeweiligen Modalitäten (3D-Modelle, Bilder, (multimodale) Texte) spezialisiert sind – beispielsweise die bereits benannten computerlinguistischen Verfahren zur Texterkennung sowie ein modulares Object Retrieval für die Erkennung von architektonischen Strukturen in Bildern (Münster et al., in print) und die in Schritt 3 ausgeführte Übertragung dieser Segmentierung auf 3D-Modelle.

Darauf aufbauend erfolgen in weiteren zukünftigen Schritten einerseits die Verbesserung und multimodale Validierung von Erkennungsqualitäten sowie die Entwicklung eines Demonstrators zur Nutzererprobung mit Historiker*innen.

Dies dient auch als Grundlage einer Bewertung von KI-Ansätzen für die historische Forschung. Hier gilt es beispielsweise, die Diskrepanz zwischen dem großen Datenbedarf der KI-Modelle und der Komplexität des geschichtswissenschaftlichen Expertenwissens zu untersuchen und damit zu bewerten, wie effektiv existierende KI-Modelle mit begrenzten Datenmengen für Teilaspekte der (architektur-)geschichtlichen Quellenkritik eingesetzt werden können.

Fußnoten

1. <http://www.iconclass.org/rkd/61F/> sowie <http://www.iconclass.org/rkd/47/>, 15.07.2021.
2. <https://www.wikimedia.de/projects/wikibase/>, 15.07.2021
3. <http://vocab.getty.edu/aat/300000885>, 15.07.2021.

Bibliographie

- Agarwal, S./ Furukawa, Y./ Snaveley, N./ Simon, I./ Curless, B./ Seitz, S. M. / Szeliski, R.** (2011). Building rome in a day. *Communications of the ACM*, 54, 105.
- Arnold, T./ Tilton, L.** (2019). Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*.
- Bell, P. / Ommer, B.** (2019). Computer Vision und Kunstgeschichte – Dialog zweier Bildwissenschaften. In: KUROC-ZYNSKI, P., BELL, P. & DIECKMANN, L. (eds.) *Digital Art History*. Heidelberg.
- Brieber, D./ Nadal, M./ Leder, H. / Rosenberg, R.** (2014). Art in Time and Space: Context Modulates the Relation between Art Experience and Viewing Time. *PLoS ONE*, 9, e99019.
- Bruschke, J./ Maiwald, F./ Münster, S. / Niebling, F.** (2018). Browsing and Experiencing Repositories of Spatially Oriented Historic Photographic Images. *Studies in Digital Heritage*, 2, 138-149.
- Dewitz, L./ Kröber, C./ Messemer, H./ Maiwald, F./ Münster, S./ Bruschke, J. / Niebling, F.** (2019). HISTORICAL PHOTOS AND VISUALIZATIONS: POTENTIAL FOR RESEARCH. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W15, 405-412.
- Fiorucci, M./ Khoroshiltseva, M./ Pontil, M./ Travaglia, A./ Del Bue, A. / James, S.** (2020). Machine Learning for Cultural Heritage: A Survey. *Pattern Recognition Letters*, 133, 102-108.

- Fleck, L.** (1980). *Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv*, Frankfurt a. M., Suhrkamp.
- Guo, Y./ Wang, H./ Hu, Q./ Liu, H./ Liu, L. / Bennamoun, M.** (2020). Deep Learning for 3D Point Clouds: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1.
- Hackel, T./ Wegner, J. D. / Schindler, K.** (2016). Fast semantic segmentation of 3D point clouds with strongly varying density. *ISPRS Annals*, 3, 177–184.
- Hessel, J./ Lee, L. / Mimno, D.** (2019). Unsupervised Discovery of Multimodal Links in Multi-image, Multi-sentence Documents. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Holenstein, E.** (1988). *Linguistik, Semiotik, Hermeneutik*, Frankfurt am Main.
- Hoppe, S./ Breitling, S.** (2016). Virtual Palaces, Digital Images – an Introduction. In: S., H. & S., B. (eds.) *Virtual Palaces, Part II: Lost Palaces and their Afterlife. Virtual Reconstruction between Science and Media*. Heidelberg: arthistoricum.net.
- Jiao, L./ Zhang, F./ Liu, F./ Yang, S./ Li, L./ Feng, Z. / Qu, R.** (2019). A Survey of Deep Learning-Based Object Detection. *IEEE Access*, 7, 128837-128868.
- Kamposiori, C. / Benardou, A.** (2011). Collaboration in Art Historical Research: Looking at primitives. *Z. Kunstgeschichte* [Online]. Available at <http://www.kunstgeschichte-ejournal.net/157/> (accessed 17 June 2011).
- Klinke, H.** (2016). Big Image Data within the Big Picture of Art History. *Int. J. Digital Art History*, 2.
- Knorr-Cetina, K. / Reichmann, W.** (2015). Epistemic Cultures. *International Encyclopedia of the Social & Behavioral Sciences*. Amsterdam: Elsevier.
- Kohle, H.** (2018). *Initiative zur Einrichtung eines Schwerpunktprogramms. Das digitale Bild*.
- Köhler, T./ Günther, F./ Herbst, S./ Münster, S. / Fischer, H.** (2016). *Abschlussbericht im Projekt SUFES. Unterstützungsangebote und Strukturen im Themenfeld eScience*, Dresden.
- Kröber, C.** (2021). German Art History Students' use of Digital Repositories: an Insight *Papers Proceedings, Diversity, Divergence, Dialogue*. Cham: Springer LNCS.
- Kuroczynski, P./ Bell, P. / Dieckmann, L.** (eds.) 2019. *Digital Art History*, Heidelberg.
- Lee, E. / Münster, S.** (2018). Fishing for Knowledge in a Sea of Data (Session). *24th Annual Meeting of the European Association of Archaeologists*, 2018 Barcelona.
- Li, M./ Nan, L./ Smith, N. / Wonka, P.** (2016). Reconstructing building mass models from UAV images. *Computers & Graphics*, 54, 84-93.
- Maiwald, F./ Henze, F./ Bruschke, J. / Niebling, F.** (2019). Geo-Information Technologies for a Multimodal Access on Historical Photographs and Maps for Research and Communication in Urban History. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W11, 763-769.
- Maiwald, F. / Maas, H.-G.** (2021). An automatic workflow for orientation of historical images with large radiometric and geometric differences. *The Photogrammetric Record*, 36, 77-103.
- Martinovic, A./ Knopp, J./ Riemenschneider, H. / Van Gool, L.** (2015). 3d all the way: Semantic segmentation of urban scenes from start to end in 3d. *IEEE Computer Vision & Pattern Recognition*. 4456–4465.
- Minace, S./ Boykov, Y. Y./ Porikli, F./ Plaza, A. J./ Kehtarnavaz, N. / Terzopoulos, D.** (2021). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1.
- Münster, S./ Apollonio, F./ Bell, P./ Kuroczynski, P./ Lenardo, I. D./ Rinaudo, F. / Tamborrino, R.** (2019). Digital Heritage meets Digital Humanities. *ISPRS Archives*, XLII-2/W15, 813–820.
- Münster, S./ Bruschke, J./ Maiwald, F. / Kleiner, C.** (in print). Software and content design of a browser-based mobile 4D VR application to explore historical city architecture. *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents*.
- Münster, S./ Kamposiori, C./ Friedrichs, K. / Kröber, C.** (2018). Image Libraries and their Scholarly Use in the Field of Art and Architectural History. *Int. J. Digital Libraries*, 19, 367–383.
- Münster, S. / Terras, M.** (2020). The visual side of digital humanities. A survey on topics, researchers and epistemic cultures in visual digital humanities. *Digital Scholarship in the Humanities*, 35, 366–389.
- Opgenoorth, E.** (1997). *Einführung in das Studium der neueren Geschichte*, Paderborn.
- Panofsky, E.** (1939). *Studies in Iconology. Humanistic Themes in the Art of the Renaissance*, Oxford, Oxford University Press.
- Polanyi, M.** (1966). *The tacit dimension*, Chicago, University of Chicago Press.
- Popper, K.** (1998). *Objektive Erkenntnis. Ein evolutionärer Entwurf*, Hamburg, Hoffmann und Campe.
- Precht, P.** (2002). *Edmund Husserl zur Einführung*, Hamburg.
- Reich, K.** (2006). *Konstruktivistische Ansätze in den Sozial- und Kulturwissenschaften. Konstruktivistische Didaktik: Lehr- und Studienbuch mit Methodenpool*. Beltz.
- Utescher, R. / Zarriß, S.** (2021). What Did This Castle Look like before? Exploring Referential Relations in Naturally Occurring Multimodal Texts. In *Proceedings of the Third Workshop on Beyond Vision and LAnguage: inTEgrating Real-world kNowledge (LANTERN)*.
- Vosselman, G./ Gorte, B. G./ Sithole, G. / Rabbani, T.** (2004). Recognising structure in laser scanner point clouds. *ISPRS Archives*, 46, 33-38.
- Wengenroth, U.** (1998). *Was ist Technikgeschichte?*, o. Ort.
- Wiedemann, A./ Hemmleb, M. / Albertz, J.** (2000). Reconstruction of historical buildings based on images from the Meydenbauer archives. *ISPRS Archives*, XXXIII, 887–893.
- Xie, J./ Kiefel, M./ Sun, M. T. / Geiger, A.** (2016). Semantic instance annotation of street scenes by 3d to 2d label transfer. *IEEE Computer Vision & Pattern Recognition*. 3688-3697.
- Ying, S./ Münster, S./ Köhler, T. / Sommer, C.** (unpublished). A look at the research on design ideas generation in industrial design: Review of literature from 2007 to 2016. *Int. J. of Design Creativity & Innovation*.
- Zöllner, H.-B.** (2005). *Hermeneutischer Zirkel und hermeneutische Differenz Perspektivität und Objektivität*.

Nachhaltige Softwareentwicklung

Von der Inhouse-Lösung zur Open Source-Community am Beispiel von MerMEId

Henny-Krahmer, Ulrike

ulrike.henny-krahmer@uni-rostock.de
Universität Rostock, Germany

Stadler, Peter

pstadler@mail.uni-paderborn.de
Universität Paderborn, Germany

Einleitung

Die nachhaltige Entwicklung und Verfügbarmachung von Forschungssoftware ist eine der zentralen Herausforderungen in den Digital Humanities. Während Praktiken der Überlieferung und Sicherung von Forschungsdaten schon einen gewissen Reifegrad erreicht haben, steckt die Kultur der Entwicklung eines digitalen Gedächtnisses in Bezug auf Software noch in den Anfängen (Katerbow 2018: 5-6). Nachhaltige Software soll ermöglichen, dass Forschung transparent und nachvollziehbar bleibt, indem Tools und Dienste langfristig auffindbar, einsehbar und möglichst ausführbar bleiben und gut dokumentiert sind. Auch geht es darum, Software nachnutzbar zu machen, damit Mittel für ihre Entwicklung effizient eingesetzt werden. Die Zielstellung nachhaltiger Software ist damit eng mit den Ideen verknüpft, die auch hinter den FAIR-Prinzipien stehen (Lamprecht et al. 2020, Hasselbring et al. 2020). Wie nachhaltig Forschungssoftware ist und sein kann, wird durch viele Faktoren bedingt, u.a. durch technische, organisatorische sowie politisch-soziale Aspekte. Es stellt sich die Frage, welche Kriterien dafür genau erfüllt sein müssen, wofür es bereits verschiedene Vorschläge gibt, u.a. aus einer allgemein-theoretischen Sicht (Stürmer et al. 2017), aus technischer Sicht (Druskat 2017) oder aus einer förderpolitisch motivierten Sicht (Anzt et al. 2021).

Klar ist, dass nicht jede jemals entwickelte Forschungssoftware dauerhaft lauffähig gehalten werden kann. Für etablierte und verbreitete geisteswissenschaftliche Tools wie z.B. Stylo (Eder et al. 2016), CATMA (Gius et al. 2021), ediarum (Dumont et al. 2021) oder auch MerMEId (MerMEId Community 2021) ist dies jedoch anzustreben.¹ Die Frage der Nachhaltigkeit solcher Forschungssoftware wird immer dringender, je länger die Software existiert und auch, je häufiger sie eingesetzt wird, um Forschungsergebnisse zu produzieren (zur Problematik zahlreiche softwarebasierte DH-Projekte langfristig zu erhalten siehe z.B. Smithies et al. 2019).

Dieser Beitrag zielt darauf, bestehende Kriterien für nachhaltige Softwareentwicklung am Beispiel des musikwissenschaftlichen Metadaten-Editors MerMEId zu diskutieren. Dabei wird insbesondere die Entwicklungsgeschichte der Software in den Blick genommen, da sie zunächst als Inhouse-Lösung entwickelt wurde und kürzlich in ein Community-Projekt umgewandelt wor-

den ist. Auf der Grundlage der Erkenntnisse zu MerMEId wird die Anwendung der bisher hauptsächlich fachübergreifend formulierten Nachhaltigkeitskriterien auf geisteswissenschaftliche Forschungssoftware kritisch reflektiert.

Kriterien für nachhaltige Forschungssoftware

Die Darstellung bestehender Kriterien für nachhaltige Softwareentwicklung beschränkt sich auf zwei Beispiele: auf der einen Seite praktische Auswahlkriterien für Forschungssoftware, die langfristig gefördert werden sollte (Anzt et al. 2021) und auf der anderen Seite abstraktere, allgemeine Bedingungen für die (auch soziale und ökologische) Nachhaltigkeit digitaler Artefakte und ihrer Ökosysteme (Stürmer et al. 2017). Es gibt weitere Vorschläge für Kriterien für gute oder nachhaltige Software, auf die hier nicht umfassend eingegangen werden kann. Die beiden ausgewählten Kriteriengruppen ergänzen sich durch die unterschiedliche Schwerpunktlegung gut und eröffnen je eigene Perspektiven auf die Nachhaltigkeit von Software. Sie wurden jedoch nicht speziell für geisteswissenschaftliche Forschungssoftware entwickelt.² Im Folgenden werden die von Anzt et al. formulierten Kriterien sinngemäß wiedergegeben, da für die MerMEId im Einzelnen überprüft werden soll, ob sie erfüllt sind oder nicht:



Abb. 1: Nachhaltigkeitskriterien nach Anzt et al. 2021 (eigene Darstellung).

Den Hauptteil der Kriterien bei Anzt et al. machen die Transparenz und der Qualität der Software aus. Daneben zeigen die Bereiche "Nutzung und Impact" und "Reife", dass eine dauerhafte Förderung für Anzt et al. auch davon abhängt, wie stark die Software tatsächlich genutzt wird und etabliert ist. Den recht detaillierten, direkt anwendbaren Kriterien von Anzt et al. stehen die theoretisch hergeleiteten Kriterien von Stürmer et al. (2017) gegenüber, denen das Konzept der digitalen Artefakte (Daten oder

Code) zugrunde liegt. Diese benötigen eine technische und soziale Umgebung, um verarbeitet zu werden, und sind von einem sich veränderndem Ökosystem abhängig und davon, dass sie erstellt, verändert und genutzt werden. Aus dieser Eigenschaft leiten sich die in Abb. 2 gezeigten Grundbedingungen für Nachhaltigkeit ab.

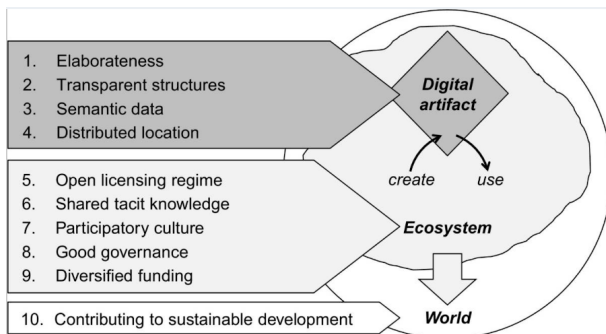


Abb. 2: Grundbedingungen für Nachhaltigkeit nach Stürmer et al. 2017.

Die Kriterien von Stürmer et al. sollen hier nicht im Einzelnen auf MerMEId angewandt werden, da sie sich in Teilen mit den Kriterien von Anzt et al. decken. Vor allem die Kriterien zum Ökosystem sind jedoch für MerMEId von Interesse, da sie den Blick verstärkt auf Aspekte der sozialen Umgebung lenken, was im Hinblick auf den Übergang der MerMEId von einer Inhouse-Lösung zu einer Community-Software besonders relevant ist. Sie werden daher in der Diskussion ergänzend berücksichtigt. Der Punkt "Shared tacit knowledge" z.B. bezieht sich darauf, dass individuelle Erfahrung und Kompetenz im Umgang mit digitalen Artefakten laufend sozial geteilt und externalisiert werden müssen, damit die Artefakte an die sich ständig ändernden Bedingungen angepasst und genutzt werden können. Dafür braucht es eine stimulierende Umgebung, was mit "Participatory culture" beschrieben wird, z.B. eine Open Source-Community, zu der gerne beigetragen wird, die aber auch durch Regeln, Normen und leitende Instanzen ("Good governance") gesteuert werden muss.

Fallbeispiel MerMEId

Die MerMEId ist ein "Metadata Editor and Repository for MEI Data", der zur Erstellung musikwissenschaftlicher (thematischer) Werkkataloge an der Königlichen Bibliothek zu Kopenhagen ab ca. 2009 entwickelt wurde. Obwohl MerMEId an erster Stelle für die eigenen Arbeiten des "Danish Centre for Music Editing" an dem "Catalogue of Carl Nielsen's Works" – dem später noch das Scheibe-Werkverzeichnis und das Hartmann-Werkverzeichnis folgen sollten – entwickelt wurde, waren doch die Entwickler Axel Teich Geertinger und Sigfrid Lundberg von Anfang an maßgeblich auch an der Gestaltung des noch jungen Datenstandards MEI beteiligt. MerMEId war dadurch frühzeitig in der MEI-Community bekannt und wurde daraufhin aktiv auf zahlreichen Workshops (z.B. bei der Edirom-Summer-School ab 2012) oder durch Konferenzbeiträge (z.B. bei der ersten MEC 2013 in Mainz) vorgestellt und verbreitet. Auch durch Axel Teich Geertingers Mitgliedschaft im MEI-Board (2015–16) und in der "Metadata and Cataloging Interest Group" gab es einen regen Austausch und wechselseitigen Einfluss zwischen dem MerMEId-Editor und dem MEI-Metadatenschema. Mehrere externe Projekte setzten daraufhin den MerMEId-Editor für die Erstellung von Werkkatalogen ein, darunter "Bruckner Online"³, die Kataloge zu Johan

Svendsen und Geirr Tveitt⁴, der "Catalogue of the Works of Frederick Delius"⁵, oder das "Bach Repertorium"⁶.

Die "Goldene Ära" von MerMEId endete dann aber 2019, als die Schließung des *Danish Centre for Music Editing* beschlossen wurde und die Entwickler keine Kapazitäten mehr für Weiterentwicklung und Pflege von MerMEId aufwenden durften.⁷ Zu diesem Zeitpunkt war der Quellcode von MerMEId bereits unter einer Apache-2.0 OpenSource-Lizenz auf GitHub veröffentlicht, es gab eine MerMEId-Sandbox (Demo-Seite), sowie eine unorganisierte (i.e. es fehlten Instrumente wie Mailinglisten, Messenger o.ä.) Community von MerMEId-Nutzer*innen.

Der *Virtuelle Forschungsverbund Edirom* 'adoptierte' daraufhin das Projekt und ziemlich schnell stießen Kolleg*innen aus der ÖAW dazu. Als weitere informelle Anfragen eingingen und Interessenten auf den Plan traten wurde klar, dass diesem breiten Interesse nur durch eine echte Community-Struktur Rechnung getragen werden konnte. Als Community-Instrumente wurden neben dem vorhandenen Code-Repository inkl. Ticketsystem und Wiki bei GitHub auch monatliche (virtuelle) Community-Meetings installiert sowie ein eigener Kanal im MEI-Slack eingerichtet. Aktuell sind die Community-Aktivitäten noch sehr auf das Refactoring des Codes konzentriert, da als erster Meilenstein ein Release einer "Community-Edition" geplant ist. Diese soll noch keine wesentlichen Feature-Neuerungen enthalten, sondern zunächst eine Umgestaltung der Softwarearchitektur zum Zwecke der besseren Wartbarkeit und des einfacheren Deployments.

Anwendung der Nachhaltigkeitskriterien auf MerMEId

Im Folgenden werden die Kriterien von Anzt et al. auf MerMEId angewandt und für jedes erfüllte Kriterium ein Punkt vergeben.

Nutzung und Impact: 3,5/5

Die meisten Kriterien zu Nutzung und Impact können für MerMEId positiv verbucht werden. Die Software wird in mehr als einer Forschungsgruppe eingesetzt (Kriterium 1) und sie ist auch die einzige Software, die das Problem eines anwenderfreundlichen Metadateneditors für MEI löst (Kriterium 4). Teilweise erfüllt ist das Kriterium 2 durch einen Aufsatz und ein Poster (Geertinger und Pugin 2011, Geertinger und Lundberg 2015). Auch das Kriterium 3 kann als teilweise erfüllt gelten, da es eine kurze Rezension von MerMEId gibt (Crandell 2015) und das Tool auf diversen Workshops vorgestellt wurde. Auch Kriterium 5 (Informations- und Lehrmaterialien) bewerten wir nur mit der halben Punktzahl, da sich im Git Repository von MerMEId zwar einige kleine Tutorials zu Spezialfällen finden, diese aber nicht als genuine Beispiele oder Tutorien für das Selbststudium gelten können.

Transparenz und Qualität der Software: 6,5/11

Für die Kriterien 6, 7 und 9 kann das Code-Repo bei GitHub als Beleg dienen, in dem u.a. auch die geforderte FLOSS-Lizenz (Apache 2.0) angegeben ist. Daneben finden sich dort auch Beispieldaten, die zusammen mit der Sandbox-Umgebung das Kriterium 11 erfüllen. Teilweise (jeweils mit 0,5 gewertet) erfüllt sind die Forderungen nach Dokumentation (zu Kriterium 8 gibt es ein User-Manual sowie ein Wiki für Entwickler*innen, in dem auch die Software-Abhängigkeiten zu eXist und Orbeon Forms beschrieben sind), sowie die Releases (Kriterium 15). Es finden

sich zwar tags in der Git-Versionsverwaltung und auch Container-Images werden bereitgestellt, allerdings gibt es keine "richtigen" Releases mit archivierbaren Binaries. Erweiterbarkeit (Kriterium 12), Interoperabilität (Kriterium 13) und Testing (Kriterium 14) müssen negativ beschieden werden, allein Kriterium 16 darf wieder positiv gewertet werden, da die Funktionalitäten von MerMEId ein Alleinstellungsmerkmal sind.

Reife: 3/5

Für MerMEId gibt es keinen Software-Management-Plan (Kriterium 17) und es kann leider auch nicht behauptet werden, dass die Software einfach zu warten wäre (Kriterium 18). Die Kriterien 19–21 wiederum werden durch das GitHub-Repositorium dokumentiert bzw. erfüllt und auch im MerMEId Slack Channel und bei den monatlichen Community-Meetings treffen sich nicht nur Entwickler*innen, sondern auch Nutzer*innen zum Austausch.

Diskussion und Fazit

Die Anwendung der Kriterien von Anzt et al. auf MerMEId funktioniert zunächst sehr gut, d.h. die Kriterien sind klar definiert und lassen sich checklistenartig beantworten. An manchen Stellen haben wir uns mit halben Punkten beholfen, da die Kriterien nur teilweise erfüllt werden konnten, uns eine komplette Negierung aber zu stark erschien. Insgesamt gibt dieser Score (13/21 -> 62%) unsere intuitive Verortung von MerMEId gut wieder: "es ist schon vieles gut, aber es gibt auch noch wesentliche Baustellen". Diese Baustellen lassen sich dank der Kriterien auch klar benennen und finden sich sowohl im Bereich der Qualität der Software als auch bei der Dokumentation.

Auffallend ist, dass sich der deutliche Bruch in der Organisation der Entwicklung (von zwei Hauptentwicklern aus derselben Institution hin zu einer internationalen Community) kaum in der Bewertung auswirkt. Obschon nicht explizit ausgeführt, würden sich die Kennzahlen kaum ändern, wenn man diese beiden Zeiträume getrennt auswerten würde. Dies mag zum einen daran liegen, dass die Community-Edition noch relativ jung ist und daher zeitlich noch keine signifikanten Änderungen bewirken konnte. Es fällt aber auf, dass die Art, wie die Entwicklung eines Softwareprojekts organisiert ist, insgesamt eine eher untergeordnete Rolle bei den Kriterien von Anzt et al. spielt. Betrachtet man die Umstellung von einer Inhouse-Lösung zur Open Source-Community mit Hilfe der Ökosystem-Kriterien von Stürmer et al., so schlägt der Systemwechsel stärker zu Buche. Eine offene Lizenzierung gab es in beiden Fällen, die übrigen vier Punkte "Shared tacit knowledge", "Participatory culture", "Good governance" und "Diversified funding" sind jedoch bei der Open Source-Community in wesentlich stärkerem Maße als bei der Inhouse-Lösung oder überhaupt erst erfüllt. Dies lässt hoffen, dass MerMEId als Community-Projekt nun gut aufgestellt ist, um nachhaltig weiterentwickelt zu werden. Allerdings werden bei Stürmer et al. als Open Source-Beispiele die Entwicklung des Linux-Kernels und Bitcoin diskutiert. Die Communities sind in diesen Fällen wesentlich größer und daher voraussichtlich stabiler als bei MerMEId, wo im jetzigen anfänglichen Stadium der Community-Entwicklung eine gute "Governance", also eine gewisse Leitung und Steuerung der Community-Prozesse, noch sehr wichtig ist.

Zur Frage der Anwendbarkeit der allgemeinen Nachhaltigkeitskriterien auf DH-Projekte wie MerMEId kann festgehalten werden, dass solche Projekte tendenziell kleiner sind, in Bezug auf Mittel, Personal und auch ihre Wirkung. Quantitative Aspekte,

wie sie bei Anzt et al. z.B. hinsichtlich Impact und Nutzung abgefragt werden, sind hier nicht unbedingt angemessen. In gleicher Weise stellt sich bei den Aspekten, die in Stürmer et al. zum Ökosystem genannt werden, die Frage, ob es bei kleineren DH-Softwareprojekten genug "kritische Masse" gibt, damit die Kriterien ihre positive Wirkung auf die Nachhaltigkeit entfalten können. Wir schließen daraus, dass es gerade bei DH-Projekten wie MerMEId essentiell ist, die weitere Entwicklung im Sinne eines Managements laufend im Blick zu behalten. So wie Anzt et al. (2021) fünfjährige Förderzyklen und Smithies et al. (2019) Managementpläne von gleicher Dauer vorschlagen, wird es auch für MerMEId erforderlich sein, "auf Sicht" zu fahren, die weitere Entwicklung der Software zu beobachten und regelmäßig zu bewerten, inwieweit Nachhaltigkeitskriterien erfüllt sind, um auf dieser Basis über die weitere aktive Entwicklung und Bewahrung der Software zu entscheiden. Neben technischen Aspekten kommt damit dem organisatorischen und sozialen Rahmen eine sehr wichtige Rolle für die Nachhaltigkeit der Softwareentwicklung zu.

Fußnoten

1. Als geisteswissenschaftliche Forschungssoftware wird hier Software verstanden, die für die Forschungsgegenstände, -daten und -methoden der Geisteswissenschaften wesentliche Funktionalitäten bereitstellt, für diese Zwecke entwickelt wurde oder entsprechend eingesetzt wird, z.B. Tools zur Textanalyse oder -annotation oder Werkzeuge zur Erfassung und Präsentation von edierten Texten und Metadaten, um nur einige Beispiele zu nennen.
2. Es gibt allerdings auch bereits spezifischere Vorschläge aus den DH, wie z.B. die "Criteria for Reviewing Tools and Environments for Digital Scholarly Editing" (Sichani und Spadini 2018) und die "Handreichung zur Rezension von Forschungssoftware in den Altertumswissenschaften" (Homburg et al. 2020), die allerdings beide die Qualität von Software im Allgemeinen adressieren und nicht primär ihre Nachhaltigkeit, auch wenn beides zusammenhängt.
3. <http://www.bruckner-online.at/> [letzter Zugriff 3. Juli 2021].
4. <https://www.musikkarven.no/english/work-catalogues/index.html> [letzter Zugriff 3. Juli 2021].
5. <https://delius.music.ox.ac.uk/catalogue/welcome.html> [letzter Zugriff 3. Juli 2021].
6. <https://www.bach-leipzig.de/de/bach-archiv/bach-repertoire> [letzter Zugriff 3. Juli 2021].
7. Die offizielle Schließung datiert vom 1. Mai 2020, die Git-Aktivitäten enden aber bereits am 13. August 2019.

Bibliographie

Anzt, Hartwig / Bach, Felix / Druskat, Stephan / Löffler, Frank / Loewe, Axel / Renard, Bernhard Y. / Seemann, Gunnar / Struck, Alexander et al. (2021): „An environment for sustainable research software in Germany and beyond: current state, open challenges, and call for action [version 2; peer review: 2 approved]“, in: *F1000Research* 9:295 10.12688/f1000research.23224.2.

Crandell, Adam (2015): „Review of MerMEId: Metadata Editor and Repository for MEI Data, by The National Library, Danish Centre for Music Publication“, in: *Notes* 71 (3): 543-544 10.1353/not.2015.0037.

Druskat, Stephan (2017): "Kriterienbasierte Evaluation und Dokumentation technischer Nachhaltigkeit von Forschungssoftware in einem Metadatenrepositorium", in: *DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts. Universität Bern, 13.-18. Februar 2017* 253-255 10.5281/zenodo.4622669.

Dumont, Stefan / Fechner, Martin / Grabsch, Sascha (2021): *ediarum*. <https://www.ediarum.org/>. GitHub: <https://github.com/ediarum> [letzter Zugriff 15. Juli 2021].

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): „Stylometry with R: A Package for Computational Text Analysis“, in: *The R Journal*. 8 (1): 107–121.

Geertinger, Axel Teich / Pugin, Laurent (2011): „MEI for bridging the gap between music cataloguing and digital critical edition“, in: *Die Tonkunst* 5 (3): 289–294.

Geertinger, Axel Teich / Lundberg, Siegfried (2015): „MerMEId: Creating Thematic Catalogues Using MEI Metadata“, in: Roland, Perry / and Kepper, Johannes (eds.): *Music Encoding Conference Proceedings 2013 and 2014*. Bavarian State Library (BSB) 122–126. URN: <http://nbn-resolving.de/urn:nbn:de:bsb:12-babs2-0000007812>.

Gius, Evelyn / Meister, Jan Christoph / Meister, Malte / Petris, Marco / Bruck, Christian / Jacke, Janina / Schumacher, Mareike / Gerstorfer, Dominik / Flüh, Marie / Horstmann, Jan (2021): *CATMA 6 (Version 6.3)*. Zenodo 10.5281/zenodo.1470118.

Hasselbring, Wilhelm / Carr, Leslie / Hettrick, Simon / Parker, Heather / Tiropanis, Thanassis (2020): „From FAIR research data toward FAIR and open research software“, in: *it - Information Technology* 62 (1): 39–47 10.1515/itit-2019-0040.

Homburg, Timo / Klammt, Anne / Mara, Hubert / Schmid, Clemens / Schmidt, Sophie Charlotte / Thier, Florian / Trognitz, Martina (2020): „Diskussionsbeitrag - Handreichung zur Rezension von Forschungssoftware in den Altertumswissenschaften / Impulse - Recommendations for the review of archaeological research software.“ *GitHub*. https://research-squirrel-engineers.github.io/Impuls_SoftwareRezensionen_DGUF/Draft.htm [letzter Zugriff 15. Juli 2021].

Katerbow, Matthias / Feulner, Georg (2018): *Handreichung zum Umgang mit Forschungssoftware*. Hrsg. von der Schwerpunktinitiative Digital Information der Allianz der deutschen Wissenschaftsorganisationen 10.5281/zenodo.1172970.

Lamprecht, Anna-Lena / Garcia, Leyla / Kuzak, Mateusz / Martinez, Carlos / Arcila, Ricardo / Martin Del Pico, Eva / Dominguez Del Angel, Victoria et al. (2020): „Towards FAIR principles for research software“, in: *Data Science* 3 (1): 37–59 10.3233/DS-190026.

MerMEId Community (2021). *MerMEId*. *GitHub.com*. <https://github.com/Edirom/MerMEId>.

Sichani, Anna-Maria / Spadini, Elena and the members of the IDE (2018): *Criteria for Reviewing Tools and Environments for Digital Scholarly Editing, version 1.0*. <https://www.i-d-e.de/publikationen/weitereschriften/criteria-tools-version-1/> [letzter Zugriff 15. Juli 2021].

Smithies, James / Westling, Carina / Sichani, Anna-Maria / Mellen, Parn / Ciula, Arianna (2019): „Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab“, in: *Digital Humanities Quarterly* 13 (1). <http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html> [letzter Zugriff 15. Juli 2021].

Stürmer, Matthias / Abu-Tayeh, Gabriel / Myrach, Thomas (2017): „Digital sustainability: basic conditions for sustainable digital artifacts and their ecosystems“, in: *Sustainability Science* 12 (2): 247–262 10.1007/s11625-016-0412-2.

Nathan nicht ihr Vater? Wissensvermittlungen im Drama annotieren

Andresen, Melanie

melanie.andresen@ims.uni-stuttgart.de
Universität Stuttgart, Germany

Krautter, Benjamin

benjamin.krautter@uni-koeln.de
Universität Stuttgart, Germany

Pagel, Janis

janis.pagel@uni-koeln.de
Universität Stuttgart, Germany

Reiter, Nils

nils.reiter@uni-koeln.de
Universität zu Köln, Germany

Die quantitative Dramenanalyse hat sich lange Zeit vor allem auf formale Merkmale der Textoberfläche konzentriert¹: Wie häufig treten Figuren auf (vgl. Marcus 1973 [1970]: 287–369), mit welchen anderen Figuren stehen sie gemeinsam auf der Bühne (vgl. etwa Yarkho 2019 [1935–1938]), wie viel sprechen sie (vgl. Moretti 2013: 2–4) und wie viel wird über sie gesprochen (vgl. Willand u.a. 2020: 177–181). All diese Informationen lassen sich, zumindest in maschinenlesbar kodierten Dramen, relativ einfach abrufen und weiterverarbeiten, etwa zu Figurennetzwerken (vgl. etwa Trilcke u.a. 2016: 255–258) oder formalen Analysen der Figurenrede (vgl. Reiter, Willand 2018: 45–75). Im Projekt Q:TRACK (QuaDrama: Tracking Character Knowledge) widmen wir uns einer stärker inhaltlich fokussierten Erschließung von Dramen, genauer Prozessen der Vermittlung von Wissen über Familienrelationen der Figuren. Das (fehlende) Wissen über Verwandtschaftsverhältnisse ist für zahlreiche deutschsprachige Dramen des 18. und 19. Jahrhunderts entscheidendes Element der Handlung, sodass sich eine systematische Untersuchung aufdrängt.

In diesem Beitrag gehen wir zunächst auf die Bedeutung von Wissen und Wissensvermittlungen für die Handlung wie auch die Wirkung von Dramen ein. Anschließend beschreiben wir, wie solche Prozesse der Wissensvermittlung in Annotationen erfasst und modelliert werden können.² Am Beispiel von Gotthold Ephraim Lessings kanonischem Drama *Nathan der Weise* (1779) zeigen wir, wie sich die Analyse eines Dramas auf diese Annotationen aufbauen lässt. Im Fazit blicken wir auf Perspektiven für die Automatisierung und die quantitative Analyse größerer Dramenbestände.

Wissensvermittlung in Dramen

Die Interferenz von innerem und äußerem Kommunikationssystem im Drama, also die Kommunikation der fiktiven Figuren auf der einen Seite und die Wahrnehmung dieser Kommunikation durch das Publikum auf der anderen Seite, gilt als eine

zentrale „Differenzqualität dramatischer Kommunikation“ (Pfister 2001: 80). Die Bühnenfiguren zeichnen sich schon mit Blick auf die Vorgeschichte des Dramas potentiell durch einen unterschiedlichen Wissensstand aus, der sich im Laufe des Stücks fortwährend verändern kann, etwa hinsichtlich ihrer Handlungsziele. Dadurch wird zugleich das Verhältnis zwischen dem Informationsstand des Publikums und demjenigen der einzelnen Dramenfiguren immer wieder neu justiert. Die Exposition reduziert etwa den zu Beginn eines Dramas vorherrschenden Wissensrückstand des Publikums gegenüber den Figuren (vgl. etwa Asmuth 2015: 122). Die Unterschiede im „Grad der Informiertheit“ – Manfred Pfister spricht hierbei in Rekurs auf den Shakespeare-Forscher Bertrand Evans von „diskrepante[r] Informiertheit“ (Pfister 2001: 80, vgl. Evans 1960: viii) – lassen sich vor allem auf zwei ursächliche Unterschiede zwischen innerem und äußerem Kommunikationssystem zurückführen: Während das Publikum in seiner Beobachterrolle jede Szene des Stücks wahrnimmt und dadurch geäußertes partielles Wissen der Figuren abgleichen und aggregieren kann, bleibt bisweilen unklar, über welches Wissen die Figuren tatsächlich verfügen. Das gilt auch für mögliche Zeitsprünge, etwa zwischen zwei Akten des Dramas. Unklar kann zudem sein, inwieweit die Äußerungen einer Figur mit den ‚Tatsachen‘ der fiktionalen Welt übereinstimmen, ob die Äußerungen also glaubwürdig sind (vgl. Jeßing 2015: 50-51). Je nach Handlungsverlauf verfügt das Publikum also zu unterschiedlichen Zeitpunkten des Dramas über einen Informationsvorsprung oder einen Informationsrückstand gegenüber den auf der Bühne agierenden Figuren. Gleiches gilt isoliert betrachtet auch für das interne Kommunikationssystem des handelnden Bühnenpersonals. Die ‚diskrepante Informiertheit‘ zweier Figuren kann so zu unterschiedlichen Bewertungen derselben Situation führen. Figuren, die etwa über das Wissen verfügen, dass zwei verlobte Figuren Geschwister sind, werden diese Verlobung anders beurteilen, als Figuren, denen dieses Wissen fehlt.

Diese Kluft zwischen dem Wissensstand der Figuren und demjenigen des Publikums ist als wichtiges Spannungselement des Dramas aufzufassen³, da sie für „anhaltende[] Aufmerksamkeit und emotionale[] Erregung“ sorgt (Anz 2007: 464). Besonders geläufig ist dahingehend das Mittel der dramatischen Ironie, das sich aus genau dieser Kluft der Informiertheit speist. Bedingung der dramatischen Ironie ist ein Informationsvorsprung auf Seiten des Publikums, die eine aus Perspektive der sprechenden Figur unverfängliche Äußerung als „gezielte[] Anspielung auf die spätere Katastrophe“ zu deuten verstehen.⁴ Elemente wie die dramatische Ironie sind folglich eng mit der Wirkung des Dramas auf das Publikum verknüpft. Schon Aristoteles bestimmt die (kathartische) Wirkung in seiner Dramenpoetik als zentrales Anliegen der Tragödie (vgl. zur Katharsis Schmitt 2008: 333–348 u. 476–510). Als wichtige Handlungsbausteine, um die von ihm für die Wirkung gewünschten Affekte hervorzurufen, betrachtet er den Handlungsumschlag (Peripetie) und die Wiedererkennung (Anagnorisis). Letztere hängt unmittelbar mit der diskrepanten Informiertheit der Figuren zusammen. Aristoteles bestimmt die Wiedererkennung als „Umschlag von Unkenntnis in Kenntnis, mit der Folge, daß Freundschaft oder Feindschaft eintritt“ (Aristoteles 1982: 35). Da solche Erkennungsszenen idealerweise mit dem Handlungsumschlag „dessen, was erreicht werden soll“ (Aristoteles 1982: 35), verknüpft sind, stellen sie zentrale Momente der Wissensvermittlung dar, die ganz entscheidend für die Interpretation des Dramas sein können.

Ziel unseres Annotations- und Modellierungsvorhabens ist es deshalb, das sich verändernde Wissen über Familienrelationen sowohl im internen als auch im externen Kommunikationssystem abzubilden. Wir wollen dabei nicht nur die zentralen Szenen

der Wiedererkennung annotieren, sondern vor allem die einzelnen Schritte nachvollziehen, die einen solchen für die dramatische Wirkung entscheidenden Wissensumschlag anleiten.

Annotation von Wissensvermittlungen

In einem ersten Schritt werden Textstellen im Drama, an denen Wissen über Familienrelationen vermittelt wird, manuell annotiert. Entscheidend für die Annotation ist, dass sich der Wissensstand einer Figur oder des Publikums tatsächlich verändert. Relevante Textstellen werden mit einem strukturiert zusammengesetzten Label versehen, das sowohl das vermittelte Wissen als auch die Quelle und das Ziel der Wissensvermittlung benennt. Optional können Attribute hinzugefügt werden, sodass die Annotationslabel nach dem folgenden Schema funktionieren:

```
transfer(QUELLE, ZIEL, WISSEN, ATTRIBUTE)
```

Quelle und Ziel sind in der Regel entweder Figuren des Dramas oder das Publikum (oder eine Liste mehrerer dieser Entitäten). Als Quelle kann aber auch ein Objekt oder Vorgang in der Welt in Betracht kommen (z. B. eine Beobachtung). Das für unsere Annotationen relevante Wissen ist auf Familienrelationen und Liebesbeziehungen zwischen den Figuren beschränkt, wobei die Annotationsrichtlinien ein festes Inventar von Relationen vorgibt. Formal können hierbei gerichtete Relationen wie `parent_of(PARENT, CHILD)` und ungerichtete Relationen wie `siblings(SIBLING-A, SIBLING-B)` unterschieden werden. Wenn beispielsweise Nathan in Lessings *Nathan der Weise* dem Tempelherrn mitteilt, dass er der Vater von Recha ist, wird diese Wissensvermittlung folgendermaßen annotiert:

```
transfer(nathan, tempelherr, parent_of(nathan, recha))
```

Durch die optionalen Attribute kann das vermittelte Wissen spezifiziert, also beispielsweise als unsicher oder als Lüge gekennzeichnet werden. Durch ein vorangestelltes Ausrufezeichen können Relationen oder Wissensbestände negiert werden. Beim Wissensstand kann es sich auf einer Metaebene auch um ein Wissen über Wissen handeln. So kann etwa annotiert werden, dass Daja dem Tempelherrn (und dadurch auch dem Publikum) anvertraut, dass Recha gar nicht bewusst ist, dass Nathan nicht ihr leiblicher Vater ist:

```
transfer(daja, [tempelherr, audience], !knowledge(recha, !parent_of(nathan, recha)))
```

Weitere Details zur Annotation lassen sich den auf unserer Webseite veröffentlichten Richtlinien entnehmen.⁵ Die Annotation wird von zwei Annotator:innen⁶ parallel mit dem CorefAnnotator (Reiter 2018) durchgeführt und im Anschluss mit der Erstautorin dieses Beitrags besprochen. Da die Annotationsrichtlinien inzwischen hinreichend konsolidiert sind, wird einer der nächsten Schritte in der Analyse des Inter-Annotator-Agreements (IAA) bestehen. Für diese Form der Annotationen – wenige, frei zu positionierende Annotationsspannen und aus zahlreichen Einzelinformationen zusammengesetzte Label – gibt es (noch) kein Standardverfahren zur IAA-Berechnung. Wir sehen das größte Potenzial in einer Variante des Gamma-Maßes (Mathet et al. 2015). Je nach Ergebnis der IAA-Analysen erscheinen künftig auch Einzelannotationen möglich.

Wissensbestände inferieren

Indem wir erfahren, dass Figur A Elternteil einer Figur B ist, lässt sich schließen, dass Figur B das Kind von Figur A ist, ohne dass dies im Text explizit gemacht werden müsste. Falls weitere Verwandte von Figur A bekannt sind, ergeben sich zudem weitere Verwandtschaftsverhältnisse für Figur B. Die annotierten Wissensvermittlungen müssen deshalb im Anschluss an die Annotation um alle weiteren, logisch inferierbaren Figurenrelationen ergänzt werden. Ziel des Projektes ist es, diese logischen Schlüsse durch ein formalisiertes Regelsystem zu ziehen, das auf die annotierten Wissensveränderungen angewendet werden kann und diese automatisch ergänzt. An einem ersten Prototyp dieses Inferenzsystems arbeiten wir derzeit.

Fallbeispiel: Nathan der Weise

Die zentrale Wiedererkennung in Lessings *Nathan der Weise* dreht sich um das Figurenpaar Recha und Tempelherr, das sich, nachdem der Tempelherr Recha aus einem brennenden Haus gerettet hat, ineinander verliebt. Die am Ende des Dramas stehende Erkenntnis, dass die beiden Geschwister sind, hängt an zahlreichen Wissensbausteinen, die sich im Laufe des Dramas ergeben. Hierzu gehört insbesondere die Tatsache, dass Recha nicht die leibliche Tochter Nathans, sondern seine Pflgetochter ist, sowie die Klärung der zunächst unbekannten Herkunft des Tempelherrn. Zusätzlich wird am Ende des Stücks die Verwandtschaft mit Sultan Saladin und dessen Schwester Sittah deutlich. Aufschlussreich kann es nun sein, nachzuvollziehen, welchen Weg die einzelnen Wissensbestände durch den Figurenbestand nehmen. Dies soll am Beispiel der Familienrelation `!parent_of(nathan, recha)` illustriert werden, also der Information, dass Nathan nicht der (leibliche) Vater Rechas ist. Abbildung 1 stellt den Weg dieser Information durch das Figurennetz grafisch dar.

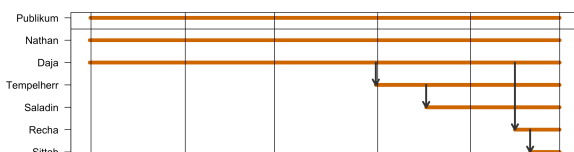


Abb. 1: Welche Figur weiß zu welchem Zeitpunkt, dass Nathan nicht Rechas leiblicher Vater ist? Dargestellt ist der Verlauf des Dramas von links nach rechts, Pfeile zeigen die Weitergabe der Information von einer Figur an die nächste an.

Für das Publikum wird dieses Wissen bereits durch die Figurentafel ersichtlich. Recha wird dort im Anschluss an Nathan als „dessen angenommene Tochter“ (Lessing 1971: 206) eingeführt. Dies verleitet zu der Annahme, dass es sich dabei um ein von allen Figuren geteiltes Wissen handelt. Direkt im 1. Auftritt spielt Daja, Rechas Gesellschafterin, auf diese Tatsache an. Sie ist demnach eingeweiht. In Bezug auf Rechas Kenntnis über ihre Herkunft bleibt das Publikum zunächst im Dunkeln. Ihr Ausruf, „Da kommen die Kamele meines Vaters“ (Lessing 1971: 209), ist auch für eine Pflgetochter, die sich dieses Umstands bewusst ist, denkbar. Dem Tempelherrn gegenüber stellt sich Nathan im 5. Auftritt des 2. Aufzuges als Rechas Vater vor. Dass Nathan tatsächlich Rechas Pflgevat er ist, erfährt der Tempelherr zum Ende des 3. Aufzuges von Daja, die auf eine christliche Heirat von Recha hofft.

Auf der Metaebene („Wissen über Wissen“) wird dem Tempelherrn zudem offenbar, dass Recha sich ihrer tatsächlichen familiären Relation zu Nathan nicht bewusst ist, und klärt diese Frage damit ebenfalls für das möglicherweise noch zweifelnde Publikum. Der Tempelherr gibt dieses Wissen, empört über die zurückhaltende Reaktion Nathans auf seinen Heiratsantrag, an Saladin weiter (4. Aufzug, 4. Auftritt). In der folgenden Aussprache mit Nathan (5. Aufzug, 5. Auftritt) gibt der Tempelherr ihm gegenüber zu, von Daja bereits die wahren Verwandtschaftsverhältnisse erfahren zu haben. Abseits der Bühne hat Daja inzwischen auch Recha über ihren Status als Pflgetochter informiert. Dies erfährt das Publikum, indem Recha diesen Umstand auch Saladins Schwester Sittah berichtet (5. Aufzug, 6. Auftritt), sodass das Wissen nun alle Figuren im Kern des Dramas erreicht hat.

Für gleich mehrere Figuren lässt sich aus den Annotationen jedoch nicht direkt ableiten, zu welchem Zeitpunkt sie erstmals über das relevante Wissen verfügen. Nathan und Daja wissen bereits vor Beginn der Dramenhandlung, dass Nathan nicht Rechas leiblicher Vater ist. Direkt aus der Figurentafel lässt sich dieser Umstand indes nicht ableiten. Dass ein Vater darüber informiert ist, wer (nicht) seine leiblichen Kinder sind, ist auch in Dramen wahrscheinlich (principle of minimal departure, vgl. etwa Ryan 1980), aber nicht alternativlos. Im ersten Auftritt erfährt das Publikum also zunächst expositorisch, dass Nathan und Daja über dieses Wissen schon vor Handlungsbeginn verfügen. Ähnlich dazu wird auch der Moment, in dem Recha erfährt, nicht Nathans leibliche Tochter zu sein, nicht auf der Bühne dargestellt. Erst durch ihren Dialog mit Sittah wird offenbar, dass sie es zwischenzeitlich abseits der Bühne erfahren haben muss.

Abbildung 2 stellt den Wissensverlauf für die Information, dass Recha und der Tempelherr Geschwister sind, dar. Nathan (und damit das Publikum) hegen einen entsprechenden Verdacht (in der Abbildung hell dargestellt), seit der Tempelherr im zweiten Akt seinen Familiennamen genannt hat. Erst nachdem sich dieser Verdacht im Gespräch mit dem Klosterbruder bestätigt, eröffnet Nathan allen anderen anwesenden Figuren am Ende des Dramas, dass Recha und der Tempelherr Geschwister sind.



Abb. 2: Welche Figur weiß zu welchem Zeitpunkt, dass Recha und der Tempelherr Geschwister sind? Die hellere Markierung bei Nathan und dem Publikum zeigt an, dass bereits der Verdacht besteht, der sich erst später bestätigt.

Perspektiven für die quantitative Analyse

Liegt eine größere annotierte Stichprobe vor, können die annotierten Daten auf Muster untersucht werden, die im Hinblick auf zeitgenössische Dramenpoetiken und deren Normvorstellungen zu interpretieren sind. Anhand unserer bislang annotierten Dramen wollen wir dazu abschließend eine erste statistische Auswertung skizzieren. Das dazugehörige Analysekorpus umfasst zum gegenwärtigen Zeitpunkt elf Dramen.⁷ Tabelle 1 zeigt die Ge-

samtzahl der sechs am häufigsten annotierten Relationen innerhalb dieses Dramenkorpus. Wir führen dabei die Anzahlen beider Annotationen getrennt voneinander auf. Auffällig ist hierbei, dass der Status von Liebesrelationen besonders häufig Gegenstand der Weitergabe von Wissen zu sein scheint.

Tab. 1: Anzahl der sechs dramenübergreifend häufigsten Relationen im annotierten Korpus, aufgeschlüsselt für jeweils beide Annotationen.

Relation	Anzahl für Annotation 1	Anzahl für Annotation 2
in_love_with	66	58
child_of	40	41
parent_of	32	27
in_love_with	17	16
engaged	17	15
spouses	14	15

Darüber hinaus lässt sich feststellen, dass die Textstellen, an denen Wissensvermittlungen annotiert werden, ungleich über den Verlauf der Dramen verteilt sind. So treten zu Beginn und gegen Ende eines Dramas gehäuft Annotationen auf (jeweils 13% aller Annotationen), während die übrigen Annotationen relativ homogen über den Handlungsverlauf verteilt sind.

Ausgehend von diesen ersten Auswertungen ergeben sich für künftige quantitative Analysen vielversprechende Perspektiven. Neben der bloßen Anzahl an Relationen, die im Verlauf der Stücke als neues Wissen an andere Figuren weitergegeben werden, und der Frage nach dem Zeitpunkt der Wissensweitergabe im Verlauf des Dramas, ergeben sich auch literaturwissenschaftlich avanciertere Fragestellungen. Unterscheiden sich die Muster der Wissensweitergabe für verschiedene Gattungen, also etwa die dramatischen Großgattungen Tragödie und Komödie? Welche Figuren geben das Wissen über familiäre Figurenrelationen weiter, an welche Figuren wird es weitergegeben? Lassen sich hierbei Muster identifizieren, etwa hinsichtlich des Geschlechts der Figuren? Ist es darüberhinaus möglich, die Szenen der Wissensweitergabe näher zu charakterisieren: Wie viele Figuren stehen in diesen Szenen auf der Bühne? Wie viele sind davon an der Wissensweitergabe aktiv beteiligt?

Fazit

Eine systematische Annotation von Prozessen der Wissensvermittlung im Drama ermöglicht eine Analyse, die über formale Merkmale der Textoberfläche hinausgeht. Liegen die Wissensbestände der Figuren und ihre Entwicklung im Verlauf des Dramas in maschinenlesbarer Form vor, lassen sich Zusammenhänge zwischen verschiedenen Textstellen identifizieren, an denen Widersprüche im Wissen der Figuren deutlich werden oder konflikthafte Relationen auftreten, wenn etwa zwei Figuren zugleich Geschwister und Liebespaar sind. Diese Widersprüche sollen über ein formalisiertes Regelsystem automatisch aus der Annotation der Familienrelationen inferiert werden.

Die Erweiterung der quantitativen Analyse auf Phänomene jenseits der Textoberfläche ist naturgemäß mit größeren Herausforderungen für die Automatisierung verbunden. Vielfach zeigen sich aber sprachliche Muster, etwa Wiederholungen und Rückfragen, die einen als überraschend markierten Wissenszuwachs verdeutlichen (siehe Abbildung 3) und Hoffnung für die automatische Identifikation derartiger Textstellen machen.

TEMPELHERR. Nicht mehr! Ich bitt' Euch! – Aber Rechas Bruder? Rechas Bruder ...

NATHAN. Seid Ihr!

TEMPELHERR. Ich? ich ihr Bruder?

RECHA. Er mein Bruder?

SITTAH. Geschwister!

SALADIN. Sie Geschwister!

Abb. 3: Auszug aus *Nathan der Weise*, der die Reaktionen der Figuren auf die Information darstellt, dass Recha und der Tempelherr Geschwister sind.

Fußnoten

1. Natürlich gibt es hiervon auch Ausnahmen. Dabei handelt es sich zumeist um „Bag-of-Words“-Ansätze, etwa mittels Topic Modeling oder stilometrischer Analysen. Vgl. exemplarisch Estill, Meneses 2018.
2. Wir verstehen die Annotation als Methode, die Texte oder Textstellen um bestimmte Angaben anreichert (etwa Wortarten, Entitäten, Erzählebenen). Die Annotationsdaten können dabei verschiedene Funktionen einnehmen. Sie können als Trainings- oder Testdaten für maschinelle Lernverfahren dienen, andererseits aber auch die Interpretation eines Textes oder einer Textstelle unterstützen und die annotierten theoretischen Begriffe im Annotationsprozess iterativ schärfen (vgl. Pagel u.a. 2020: 125–141).
3. Das gilt sowohl für die *Was*-Spannung, also der Frage, wie ein Stück ausgeht, als auch für die *Wie*-Spannung, wenn das Ende des Stücks bereits zu erraten ist, beispielsweise bei kanonischen Stoffen oder einem durch die Gattung fest vorgegebenen Schema (vgl. Anz 2007: 465).
4. Anders als diese Formulierung suggeriert, ist die dramatische Ironie nicht auf Tragödien beschränkt, sondern ist ebenso in Komödien zu finden.
5. <https://doi.org/10.5281/zenodo.5729706>.
6. Wir danken Jonas Hirner und Christian Lantzing herzlich für ihre Unterstützung bei der Annotation!
7. Es handelt sich um Goethe: *Stella*, Grillparzer: *Die Ahnfrau*, Hebbel: *Maria Magdalena*, Hofmannsthal: *Der Rosenkavalier*, Hofmannsthal: *Elektra*, Kleist: *Familie Schrockenstein*, Lenz: *Der Hofmeister*, Pfeil: *Lucy Woodvil*, Schiller: *Die Räuber*, Schiller: *Die Braut von Messina*, Schnitzler: *Komtesse Mizzi oder Der Familientag*.

Bibliographie

- Anz, Thomas (2007): „[Art.] Spannung“, in: Müller, Jan-Dirk / Braungart, Georg / Fricke, Harald / Grubmüller, Klaus / Vollhardt, Friedrich / Weimar, Klaus (eds.): *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. Bd. III: P–Z. Berlin / New York: De Gruyter 464–467.
- Aristoteles (1982): *Poetik*. Stuttgart: Reclam.
- Asmuth, Bernhard (2016): *Einführung in die Dramenanalyse*. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05472-2_9.
- Estill, Laura / Meneses, Luis (2018): „Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare's Plays“, in: *Digital Studies / Le champ numérique* 8.1: 1–22. DOI: doi.org/10.16995/dscn.295.
- Evans, Bertrand (1960): *Shakespeare's Comedies*. Oxford: Clarendon Press.
- Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hecht, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019). „Programmable Corpora – Die digitale Literaturwissen-

schaft zwischen Forschung und Infrastruktur am Beispiel von DraCor“, in: *DHd 2019 Conference Abstracts*. Frankfurt a.M. / Mainz: 194–197. <https://doi.org/10.5281/zenodo.2596095>.

Jeßing, Benedikt (2015): *Dramenanalyse. Eine Einführung*. Berlin: Erich Schmidt Verlag.

Lessing, Gotthold Ephraim (1791): „Nathan der Weise. Ein dramatisches Gedicht in fünf Aufzügen (1779)“, in: Göpfert, Herbert G. (eds.): *Gotthold Ephraim Lessing: Werke*. Bd. 2. München: Carl Hanser Verlag 205–347.

Mathet, Yann / Widlöcher, Antoine / Métivier, Jean-Philippe (2015): „The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment“, in: *Computational Linguistics* 41.3: 437–479.

Marcus, Solomon (1973 [1970]): *Mathematische Poetik*. Frankfurt a.M.: Athenäum.

Moretti, Franco (2013): „Operationalizing: or, the function of measurement in modern literary theory“, in: *Literary Lab* 6: S. 1–13. <https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf> [letzter Zugriff 15. Juli 2021].

Pagel, Janis / Reiter, Nils / Rösiger, Ina / Schulz, Sarah (2020): „Annotation als flexibel einsetzbare Methode“, in: Reiter, Nils / Pichler, Axel / Kuhn, Jonas (eds.): *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Berlin / Boston: De Gruyter 125–141.

Pfister, Manfred (2001): *Das Drama. Theorie und Analyse*. München: W. Fink.

Reiter, Nils (2018): „CorefAnnotator – A New Annotation Tool for Entity References“, in: *Proceedings of EADH 2018*. Galway, Ireland. <https://eadh2018.exordo.com/programme/presentation/118> [letzter Zugriff 15. Juli 2021].

Reiter, Nils / Willand, Marcus (2018): „Poetologischer Anspruch und dramatische Wirklichkeit: Indirekte Operationalisierung in der digitalen Dramenanalyse. Shakespeares natürliche Figuren im deutschen Drama des 18. Jahrhunderts“, in: Bernhart, Toni / Willand, Marcus / Richter, Sandra / Albrecht, Andrea (eds.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Berlin / Boston: De Gruyter 45–75.

Ryan, Marie-Laure (1980): „Fiction, Non-Factuals, and the Principle of Minimal Departure“, in: *Poetics* 9: 403–422.

Schmitt, Arbogast (2018): „Kommentar“, in: Flashar, Hellmut (eds.): *Aristoteles. Werke in deutscher Übersetzung*. Bd. 5: *Poetik*. Darmstadt: Wissenschaftliche Buchgesellschaft 193–742.

Trilcke, Peer / Fischer, Frank / Göbel, Matias / Kampkasper, Dario (2016): „Dramen als small worlds? Netzwerkdaten zur Geschichte und Typologie deutschsprachiger Dramen 1730–1930“, in: *DHd 2016 Conference Abstracts*. Leipzig: 255–258. DOI: 10.5281/zenodo.3679331.

Wiedmer, Nathalie / Pagel, Janis / Reiter, Nils (2020): „Romeo, Freund des Mercutio: Semi-Automatische Extraktion von Beziehungen zwischen dramatischen Figuren“, in: *DHd 2020 Conference Abstracts*. Paderborn: 194–200. <https://doi.org/10.5281/zenodo.4621777>.

Willand, Marcus / Krautter, Benjamin / Pagel, Janis / Reiter, Nils (2020): „Passive Präsenz tragischer Hauptfiguren im Drama“, in: *DHd 2020 Conference Abstracts*. Paderborn: 177–181. DOI: 10.5281/zenodo.3666690.

Yarkho, Boris I. (2019 [1935–1938]): „Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism)“, in: *Journal of Literary Theory* 13.1: 13–76.

Poesie als Fehler

Ein ‘Tool Misuse’-Experiment zur Prozessierung von Lyrik

Sluyter-Gäthje, Henny

sluytergaeth@uni-potsdam.de
Universität Potsdam, Germany

Trilcke, Peer

trilcke@uni-potsdam.de
Universität Potsdam, Germany

Problemhorizont und Fragestellung

Analysen der Computational Literary Studies (CLS) vorverarbeiten ihre Untersuchungsgegenstände typischerweise mit Tools des Natural Language Processing (NLP). Dabei weichen literarische Texte aufgrund ihrer historischen und/oder ästhetischen Eigenart teils eklatant von den Daten ab, auf deren Grundlage die *Models* der NLP-Tools erstellt wurden. Entsprechend sinkt die *Accuracy* der Tools etwa bei der Tokenisierung, der Lemmatisierung oder dem POS-Tagging von literarischen Texten (Scheible et al. 2011; für POS-Tagger: Rayson et al. 2007; Herrmann 2018; Bamman 2020), wobei die besondere ‘Schwierigkeit’ literarischer Texte alle gängigen Tools zu betreffen scheint (für Lemmatisierung vgl. Ortman, Roussel, Dipper 2019: 219).

Der *Performance Drop* der NLP-Tools bei Literatur ist ein computerlinguistisches Problem der Domänenadaption. Für die CLS könnte die ‘Fehlerhaftigkeit’ der Tools im Sinne devianzpoetischer Positionen (Fricke 1981) zudem die Möglichkeit bieten, ein computationelles Verständnis vom spezifischen Abweichungscharakter literarischer Texte auszubilden: Wenn die Tools für standardsprachliche Texte entwickelt wurden, dann könnten die Fehler, die diese Tools auf nicht-standardsprachlichen Texten wie der Literatur produzieren, etwas über deren Charakteristika aus computationeller Sicht verraten.

Das folgende Experiment geht von dieser basalen Überlegung aus. Gegenstand des Experiments ist die Lyrik, der regelmäßig eine “Tendenz zu erhöhter Devianz” (Müller-Zettlmann 2000: 100) attestiert wird, die sich in Form gattungsspezifischer “Störungen” (Zymner 2019: 29f.) ausdrückt. Wir entwickeln eine Pipeline, die gezielt ‘Fehler’ von NLP-Tools provoziert und diese ‘Fehler’ regelbasiert typologisiert. Damit möchten wir für die CLS auch exemplarisch den Ansatz des *Tool Misuse* profilieren, bei dem die Erzeugung von ‘fehlerhaftem’ Output computationaler Tools Grundlage für Erkenntnisse über Literatur wird.

Operationalisierung und Korpora

Fehler von Tools wären idealerweise über Daten mit Gold-Standard-Annotationen zu ermitteln. Solche Daten liegen für unser Szenario nicht vor. Deshalb implementieren wir als Workaround eine Pipeline, die folgende Idee umsetzt: Die Verarbeitung einer Zeichenkette durch ein NLP-Tool (Tokenisierung, Lemmatisierung, POS-Tagging) sollte eine Zeichenkette ergeben, die in

Tab. 2: Typen potenzieller Fehler und Identifikationsregeln

Bezeichnung	Beschreibung	Regel
PUNC	Satzzeichen sind als ADJ, NOUN, VERB pos-getaggt oder Satzzeichen sind als Teil eines Wortes tokenisiert.	Wenn im Type ein Satzzeichen (Ausnahme: Apostrophe und Bindestriche, bei denen vor und nach dem Bindestrich mindestens ein Buchstabe steht) vorkommt, typisiere.
SHORT	Einzelne Buchstaben oder Ziffern sind als ADJ, NOUN, VERB pos-getaggt.	Wenn ein Type nur zwei Zeichen oder weniger aufweist, typisiere.
ORTH_SZ	Wörter verwenden die historische Schreibung mit "ß".	Ersetze "ß" im Lemma durch "ss" und prüfe im Wörterbuch; wenn im Wörterbuch zu finden, typisiere.
ORTH_UPPER	ADJ oder VERB wurden am Versanfang großgeschrieben.	Wenn ein Type am Anfang einer Zeile steht, ersetze initiale Großschreibung bei ADJ und VERB mit Kleinschreibung und prüfe im Wörterbuch; wenn im Wörterbuch zu finden, typisiere.
ELISION_APO	Vokale innerhalb eines Wortes wurden getilgt und durch Apostroph ersetzt.	Wenn innerhalb eines Type ein Apostroph vorkommt, typisiere.
ELISION_SIMPLE	Vokale wurden in der vorletzten oder letzten Silbe eines Wortes ohne Markierung durch Apostroph getilgt.	Wenn ein Type auf ["nen", "ner", "ne", "n"] endet und davor kein Vokal und kein "l" oder "r" steht, typisiere.
ELISION_END	Auslautende Vokale in NOUN wurden getilgt.	Ergänze am Ende eines NOUN ein "e" und prüfe im Wörterbuch; wenn im Wörterbuch zu finden, typisiere.
EPITHESES	An NOUN wurde ein auslautendes "e" angehängt.	Tilge bei NOUN, die auf "e" enden, das "e" und prüfe im Wörterbuch; wenn im Wörterbuch zu finden, typisiere.
CONTRACT	Ein nachfolgendes Pronomen "es" wurde in Form von "'s" an das voranstehende Wort kontrahiert.	Wenn ein Type auf "'s" endet, typisiere.
COMP_DASH	Mittels Bindestrich wurden mehrere Wörter zu einem Wort zusammengefügt, das nicht im Wörterbuch steht.	Wenn innerhalb eines Type ein Bindestrich vorkommt, typisiere.
COMP	Mehrere NOUN wurden zu einem Wort zusammengefügt, das nicht im Wörterbuch steht.	Wenn ein NOUN gleich viele oder mehr Zeichen aufweist als der Mittelwert der Zeichenzahl aller Token im "all"-Set (8,8, gerundet 9), typisiere.
PART_ADJ	VERB wurde zu einem partizipialen ADJ abgeleitet, das nicht im Wörterbuch steht.	Wenn ein ADJ auf "end" endet, typisiere.
PREFIXED	Durch ein Präfix wurde ein Wort abgeleitet, das nicht im Wörterbuch steht.	Wenn ein Type mit einem Präfix aus einer vorgegebenen Liste beginnt, entferne Präfix aus Lemma und prüfe im Wörterbuch; wenn im Wörterbuch zu finden, typisiere.

Fehlerkommentierung

Tab. 3: Relative Häufigkeit für die Typen potenzieller Fehler für die beiden pFail-Sets

	Lyrik	Lyrik	Lyrik	Lyrik	Lyrik	Prosa	Prosa	Prosa	Prosa	Prosa
	merged_Types	spacy-Token	stanza-Token	RN-N-Token	Tree-Token	merged_Types	spacy-Token	stanza-Token	RNN-Token	Tree-Token
PUNC	0,454	0,271	0,151	0,565	0,451	0,576	1,543	0,527	1,758	0,818
SHORT	0,223	0,865	1,087	0,623	6,116	0,124	0,355	0,524	0,251	0,776
ORTH_SZ	0,120	0,186	0,180	0,194	0,183	0,133	0,195	0,191	0,192	0,194
ORTH_UPPER	0,627	1,152	1,093	0,905	0,846	0,020	0,102	0,059	0,061	0,054
ELISION_APO	2,083	2,424	2,693	2,614	2,349	0,314	0,167	0,191	0,207	0,274
ELISION_SIMPLE	2,574	5,185	5,256	4,998	4,854	0,978	1,276	1,338	1,178	1,170
ELISION_END	1,081	2,129	1,279	2,093	1,233	0,246	0,555	0,401	0,582	0,335
EPITHESES	0,285	0,380	0,359	0,368	0,354	0,128	0,139	0,132	0,133	0,124
CONTRACT	0,639	0,350	0,406	0,144	0,892	0,271	0,127	0,197	0,068	0,864
COMP_DASH	1,926	1,280	0,217	1,252	1,269	4,957	2,354	0,323	2,309	2,497
COMP	37,783	29,506	30,400	29,608	28,160	46,432	33,805	35,196	33,173	32,839
PART_ADJ	3,234	5,190	5,294	5,242	4,852	2,060	4,898	4,844	4,580	4,392
PREFIXED	2,302	2,052	2,079	2,071	1,939	3,643	3,196	3,106	2,951	2,773

53,33 % der Types im pFail-Set für Lyrik und 59,88 % der Types im pFail-Set für Prosa werden identifiziert (vgl. Tab. 3). Die

identifizierten Typen können zu Gruppen zusammengefasst werden: PUNC und SHORT sind überwiegend unterhalb der Wortebene anzusiedelnde Zeichen, meist Rauschen, das bei Lyrik und Prosa in vergleichbarem Umfang auftaucht. ORTH_SZ dokumentiert den ebenfalls bei Lyrik und Prosa vergleichbar ausgeprägten Effekt der *Historischen Orthographie*, die in unseren Korpora durch Modernisierungen bereits weitgehend abgefangen ist. Ein weiterer Normalisierungsschritt etwa mit dem DTA::CAB-Web-services⁷ könnte hier Abhilfe schaffen.

Die 10 weiteren Typen lassen sich zu drei Gruppen zusammenführen. COMP_DASH, COMP, PART_ADJ, PREFIXED versammeln *Kreative Lexik*, d.i. Wortbildungsmechanismen (Komposition, Derivation); hier handelt es sich häufig um *Out-of-Vocabulary*-Wörter, also um Pipelinefehler, nicht um Toolfehler. Bei der Lyrik lassen sich 45,25 % des "pFail"-Sets dieser Gruppe zuweisen, bei der Prosa 57,09 %. Eine erwartungsgemäß höhere Fehlerrate für Lyrik (0,62 %) als für Prosa (0,02 %) produziert die Pipeline bei ORTH_UPPER, mit dem – in Form der versinitialem Großschreibung – eine Eigenart *Lyrischer Typographie* identifiziert wird. Ebenfalls höher ist die Gruppe *Prosodische Deformation*, bestehend aus ELISION_APO, ELISION_SIMPLE, ELISION_END, EPITHESES, CONTRACT, die bei der Lyrik 6,62 % des pFail-Sets, bei der Prosa 1,93 % des pFail-Sets beschreibt. Da im Prosakorpus umfangreich auch direkte Rede enthalten ist, liegt die Annahme nahe, dass die Deformationen hier tatsächlich auf die metrisch-bedingte Hinzufügung bzw. Tilgung von Vokalen zurückzuführen ist.

Methodenkritik und Ausblick

Zu resümieren, dass die spezifisch lyrische 'Störung' für die NLP-Tools insbesondere aus der *Prosodischen Deformation* des Wortmaterials und den Eigenarten der *Lyrischen Typographie* resultiert, wohingegen die *Kreative Lexik* (für die zudem präzisere Regeln notwendig wären) auch bei der Prozessierung literarischer Prosa erhebliche Schwierigkeiten bereitet, erweist sich nicht nur angesichts fehlender Signifikanztests als zu einfach. Denn darüber hinaus ist erstens unsere Pipeline noch zu grob gebaut: zu viele potenzielle Fehler sind, wie etwa bei der kreativen Lexik, faktisch keine Tool-Fehler, sondern Pipelinefehler. Zweitens vermag unsere regelbasierte Typologisierung mit 53,33 % nur etwa die Hälfte des pFail-Sets zu beschreiben.

Darin zeigen sich zwei Felder für Anschlussforschungen: Erstens wäre zu erproben, ob sich bessere Pipelines für die automatisierte NLP-Tool-Fehleridentifikation ohne Annotationsdaten konzipieren lassen, dafür wäre es hilfreich, die Pipeline auf einem kleinen Set an Gold-Standard-Annotationen zu evaluieren; zweitens könnte auf der Grundlage unserer Pipeline gegen die Baseline von 53,33 % das regelbasierte Typologisierungsverfahren optimiert werden. Die manuelle Annotation einer kleinen Sammlung von Gedichten mit Informationen zum Abweichungscharakter jedes einzelnen Wortes würde es ermöglichen, unsere Annahme, dass unser Verständnis der Devianz durch die Nutzbarmachung des Problems der Domänenadaption von NLP-Tools operationalisiert werden kann, zu prüfen. So könnte sichergestellt werden, dass wir durch die Fehlertypisierung der NLP-Tools tatsächlich etwas über die Spezifik des Literarischen erfahren.

In jedem Fall haben wir mit dem vorliegenden *Tool Misuse*-Experiment noch nicht gut genug gelernt, die NLP-Tools 'falsch' zu verwenden.

Fußnoten

1. <https://textgridrep.org/>
2. Die Korpora sowie der Pipeline-Code sind hier zu finden: https://gitup.uni-potsdam.de/sluytergaeth/poetry_as_error
3. <https://www.projekt-gutenberg.org/>
4. <https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/departments-of-linguistics/chairs/general-and-computational-linguistics/ressources/lexica/germanet/>
5. <https://www.dwds.de/>
6. Im Prosakorpus ist die Quote der Types, die von "all" nach "pFail" übergeben werden, auffällig größer als im Lyrikkorpus, was u.a. an als NOUN getaggtten Eigennamen liegt. Eine abschließende Erklärung muss einer genaueren Analyse des Prosa-korpus vorbehalten bleiben.
7. <https://www.deutschestextarchiv.de/public/cab/>

Bibliographie

- Bamman, David** (2020): "LitBank: Born-Literary Natural Language Processing" [Preprint]. https://people.ischool.berkeley.edu/~dbamman/pubs/pdf/Bamman_DH_Debates_CompHum.pdf [letzter Zugriff 14. Juli 2021].
- Bers, Anna** (2020): "Nachwort" in: Anna Bers (eds.): *Frauen / Lyrik. Gedichte in deutscher Sprache*. Stuttgart: Reclam 793-851.
- Beutin, Wolfgang et al.** (2019): *Deutsche Literaturgeschichte*. Von den Anfängen bis zur Gegenwart. Berlin: 9. Aufl., Metzler.
- Braam, Hans** (2019): *Die berühmtesten deutschen Gedichte*. Auf der Grundlage von 300 Gedichtsammlungen. Stuttgart: 2. Aufl., Kröner."
- Fricke, Harald** (1981): *Norm und Abweichung*. Eine Philosophie der Literatur. München: Beck.
- Hamp, Birgit / Feldweg, Helmut** (1997): "GermaNet - a Lexical-Semantic Net for German", in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid: 9-15. <https://aclanthology.org/W97-0802.pdf> [letzter Zugriff 14. Juli 2021].
- Henrich, Verena / Hinrichs, Erhard** (2010): "GernEdiT - The GermaNet Editing Tool", in: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta: 2228-2235. http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf [letzter Zugriff 14. Juli 2021].
- Herrmann, J. Berenike** (2018). "Praktische Tagger-Kritik. Zur Evaluation des PoS-Tagging des Deutschen Textarchivs", in: *DHd2018: Kritik der digitalen Vernunft*. Book of Abstracts. Köln: 287-290. https://zenodo.org/record/3684897#.YO_x1W5CTOQ [letzter Zugriff 14. Juli 2021].
- Honnibal, Matthew / Montani, Ines / Van Landeghem, Sofie / Boyd Adriane** (2020): *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. <https://doi.org/10.5281/zenodo.1212303> [letzter Zugriff 14. Juli 2021].
- Klein, Wolfgang / Geyken, Alexander** (2010): "Das 'Digitale Wörterbuch der Deutschen Sprache DWDS'", in: *Lexicographica* 26: 79-96.
- Müller-Zettlmann, Eva** (2000): *Lyrik und Metalyrik*. Theorie einer Gattung und ihrer Selbstbespiegelung anhand von Beispielen aus der englisch- und deutschsprachigen Dichtkunst. Heidelberg: Winter.

Ortmann, Katrin / Roussel, Adam / Dipper, Stefanie (2019): "Evaluating Off-the-Shelf NLP Tools for German", in: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*: 212-222. https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_55.pdf [letzter Zugriff 14. Juli 2021].

Qi, Peng / Dozat, Timothy / Zhang, Yuhao / Manning, Christopher D. (2018): "Universal dependency parsing from scratch", in: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brüssel: 160-170.

Rayson, Paul / Archer, Dawn / Baron, Alistair / Culpeper, Jonathan / Smith, Nicholas (2007): "Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora", in: *Proceedings of Corpus Linguistics (CL2007)*. https://eprints.lancs.ac.uk/id/eprint/13011/1/192_Paper.pdf [letzter Zugriff 14. Juli 2021].

Schmid, Helmut (1994): "Probabilistic part-of speech Tagging using decision trees", in: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK: 154-162.

Schmid, Helmut (1995): "Improvements in Part-of-Speech Tagging with an Application to German", in: *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland: 13-25. <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf> [letzter Zugriff 14. Juli 2021].

Schmid, Helmut (2019): "Deep learning-based morphological taggers and lemmatizers for annotating historical texts", in: *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, Brüssel: 133-137.

Scheible, Silke / Whitt, Richard J. / Durrell, Martin / Bennett, Paul (2011): "A gold standard corpus of Early Modern German" in: *Proceedings of the 5th Linguistic Annotation Workshop*: 124-128. <https://dl.acm.org/doi/abs/10.5555/2018966.2018981> [letzter Zugriff 14. Juli 2021].

Zymner, Rüdiger (2019): "Begriffe der Lyrikologie" in: Hildebrandt, Claudia et al. (eds.) *Lyrisches Ich, Textsubjekt, Sprecher?* (= Grundfragen der Lyrikologie, Bd. 1). Berlin: De Gruyter 25-50.

Pragmatisches Forschungsdatenmanagement Qualitative und quantitative Analyse der Bedarfslandschaft in den Computational Literary Studies

Helling, Patrick

patrick.helling@uni-koeln.de
Institut für Deutsche Philologie, Lehrstuhl für
Computerphilologie und neuere deutsche Literaturgeschichte,
Universität Würzburg

Jung, Kerstin

kerstin.jung@ims.uni-stuttgart.de
Institut für Deutsche Philologie, Lehrstuhl für
Computerphilologie und neuere deutsche Literaturgeschichte,
Universität Würzburg

Pielström, Steffen

pielstroem@biozentrum.uni-wuerzburg.de
 Institut für Deutsche Philologie, Lehrstuhl für
 Computerphilologie und neuere deutsche Literaturgeschichte,
 Universität Würzburg

Einleitung

Die *Computational Literary Studies* (CLS) sind ein aufstrebendes, interdisziplinäres Forschungsfeld, in dem Gegenstände und Fragestellungen aus der Literaturwissenschaft mit computergetriebenen, teilweise quantitativen Methoden bearbeitet werden. Damit verorten sich die CLS am Schnittpunkt von Literaturwissenschaft, Computerlinguistik und Informatik.

Bedingt durch diese digitalen Methoden spielen Forschungsdaten unterschiedlichster Art eine zentrale Rolle für die CLS: Die Basis eines jeden Projektes stellt ein Korpus digitalisierter literarischer Texte dar. Dazu kommen weitere Arten von Forschungsdaten, unter anderem Textannotationen, zusammenfassende Statistiken und Visualisierungen, Metadaten und, bedingt durch den aktuellen Deep-Learning-Trend in der Computerlinguistik, zunehmend auch komplexe statistische Sprachmodelle. Entsprechend zeichnet sich die Forschungsdatenmanagement-Bedarfslandschaft der CLS durch eine starke Heterogenität aus, die auch in vielen anderen Teildisziplinen der Geisteswissenschaften festzustellen ist (Pempe 2012). Der Umgang mit diesen Forschungsdaten über den gesamten Forschungsdatenlebenszyklus hinweg stellt dabei eine Grundbedingung wissenschaftlichen Fortschritts dar (Bryant, Lavoie & Maipas 2017) und ist nicht erst seit Bestrebungen hin zu einer fachlich getriebenen Nationalen Forschungsdateninfrastruktur (RfII 2016, 2017) wesentlicher Bestandteil guter wissenschaftlicher Praxis (DFG 2019).¹ Für die Transparenz der wissenschaftlichen Methode und die Reproduzierbarkeit der Ergebnisse ist ein fachspezifisches Management der zu Grunde liegenden Forschungsdaten im Sinne der FAIR-Prinzipien (Wilkinson et al. 2016) bis hin zur nachhaltigen Publikation und Archivierung auch in den CLS von zentraler Bedeutung. Einige spezifische Aspekte, wie die Verwendung nicht exakt reproduzierbarer, stochastischer Verfahren oder die oft komplexe rechtliche Situation der teilweise urheberrechtlich geschützten Primärdaten stellen hierbei besondere Herausforderungen dar (vgl. Schöch et al. 2020; Kleinkopf et al. 2021).

Das DFG Schwerpunktprogramm SPP 2207 „Computational Literary Studies“ (SPP CLS) setzt sich aus insgesamt 11 an verschiedenen Universitäten in Deutschland und der Schweiz angesiedelten Forschungsprojekten und einem Datenkoordinationsteam zusammen.²

Das Datenkoordinationsteam ist mit zwei halben Stellen sowie einer Koordinierungsstelle ausgestattet und zentral an der Gesamtkoordinationsstelle des SPP CLS angesiedelt. Seine Mitglieder verfügen sowohl über langjährige Erfahrungen und Kompetenzen im methodischen Bereich der Computational Literary Studies als auch im fachspezifischen, geisteswissenschaftlichen Forschungsdatenmanagement (FDM). Zusätzlich zur Entwicklung und Umsetzung einer gemeinsamen Strategie für das Management von Forschungsdaten für das gesamte SPP unterstützt das Team auch die Koordination des gesamten Schwerpunktprogramms.

Das SPP CLS bietet durch seine Bündelung von verschiedenen CLS-Forschungsvorhaben einen hervorragenden Rahmen, um zu untersuchen, welche Art von Forschungsdaten in CLS-Projekten

wie genutzt werden. Ziel einer solchen Landschaftsvermessung ist es, die Forschungspraxis im SPP 2207 zu erfassen und zu beobachten, um daraus *Best Practices* im Umgang mit Forschungsdaten zur Schaffung eines Mehrwerts für das gesamte Feld zu identifizieren und zu aggregieren sowie methodisch verwandte Fachbereiche wie bspw. textbasiert arbeitende Digital Humanities oder die Computerlinguistik mit zu adressieren.

Darüber hinaus können die methodischen Ansätze zur Entwicklung einer Strategie für das Forschungsdatenmanagement sowie die strukturelle und organisatorische Einbindung des Datenkoordinationsteams in den Gesamtkontext des Schwerpunktprogramms als ein konkret erprobtes Praxisbeispiel für die Bedienung von FDM-Bedarfen und das Management von Forschungsdaten innerhalb von Forschungs- und Infrastrukturverbünden wie bspw. Sonderforschungsbereichen, Exzellenzclustern oder grundsätzlich übergreifenden Informationsinfrastrukturprojekten verstanden werden.

Vorgehen zur Landschaftsvermessung

Zur Analyse der Bedarfe zum Forschungsdatenmanagement in den CLS wurde ein Vorgehen mit Interviewgesprächen sowie mehreren Analyseschritten und Reviewphasen entwickelt. Durch diesen Doppelschritt konnte sich bei der Datenerfassung besonders nah am tatsächlichen Forschungsalltag sowie den aktuellen Bedingungen und Bedarfen jedes einzelnen Projekts orientiert werden. Dies kann bei bspw. quantitativen, ggf. sogar anonym durchgeführten, Onlineumfragen zu FDM-Bedarfen, die durchaus fehleranfällig sein können, nicht zwangsläufig gewährleistet werden, da hier i.d.R. keine Möglichkeit besteht auf die Antworten der Befragten konkreter einzugehen.

Zunächst wurde mit jedem Projekt ein Interview auf Basis eines Leitfadens aus 47 offenen, nach Projektphasen gruppierten Fragen durchgeführt³: (i) zum Umgang mit Daten und lebenden Systemen im laufenden Projekt, sowie (ii) zu Publikations- und Archivierungsstrategien am Ende des Projekts. Dabei dienten die qualitativen Interviews neben der Landschaftsvermessung auch dem Kennenlernen der jeweiligen Projekte sowie dem Aufbau der Kommunikation zwischen den Projektbeteiligten und dem Datenkoordinationsteam.

Im Interview wurden allgemeine Beispiele zur Erläuterung der Fragen angegeben. Aus den gegebenen Antworten wurde ein Antworteninventar erstellt, das den Projekten zusammen mit den eigenen Antworten zum Review zur Verfügung gestellt wurde. Dieser Schritt stellte sicher, dass die Antworten der Projekte korrekt zugeordnet wurden und, dass Aspekte, die auf mehrere Projekte zutreffen, aber nicht von allen erwähnt wurden, am Ende dennoch für die folgenden Analysen umfassend erfasst werden konnten.

Dabei wurden Review und Analyse zunächst auf Aspekte zum laufenden Projekt konzentriert, da Fragen zum Ende der Projektphase eher tentativ beantwortet wurden. Zum Zeitpunkt dieser Einreichung befanden sich die Fragen zum Ende der Projektphase im Reviewprozess.

Für die Landschaftsvermessung der CLS in Bezug auf (1) wissenschaftliches Arbeiten, (2) Management von Forschungsdaten sowie entsprechende (3) Trends, (4) *Best Practices* und (5) community-getriebene Standards sowie die Entwicklung einer gemeinsamen Datenstrategie wurden die Interviews zunächst quantitativ ausgewertet (siehe Abschnitt 3). Zur Umsetzung pragmatischer Lösungsstrategien für das gesamte SPP CLS durch die Datenkoordination im Sinne der Identifikation und Umsetzung

von FDM-Lösungsstrategien orientiert an der (a) Gesamtheit der Bedarfe im Schwerpunktprogramm und auf der (b) Basis existierender Werkzeuge und Angebotsstrukturen in der gesamten FDM-Landschaft, war neben dieser quantitativen Analyse auch eine qualitative Auswertung der Inhalte relevant (siehe Abschnitt 4).

Quantitative Auswertung: Erste Ergebnisse der Landschaftsvermessung

Zur Beschreibung der Datenlandschaft und Entwicklung einer passgenauen Datenstrategie für das gesamte SPP CLS ist zentral, welche Datentypen und -formate genutzt und produziert werden. Im Kontext der digitalen Literaturwissenschaften vermeintlich wenig überraschend arbeiten alle Projekte des Schwerpunktprogramms mit Textdaten und beinahe genauso viele mit Softwarecode. Aber auch numerische und bibliographische Daten, sogar Bilddaten spielen bei einigen Projekten eine wichtige Rolle (Tabelle 1) und müssen bei der Archivierung und Nachnutzbarmachung von Projektergebnissen mitberücksichtigt werden.

Tabelle 1: Genutzte Datentypen im SPP CLS.

Datentypen	Projekte
Text	10
Softwarecode	9
Numerische Daten	6
Bilddaten	5
Bibliographische Daten	4
Wörterbücher/Listen	2
Interviewdaten	2
Netzwerkdaten	1

Tab. 1: Genutzte Datentypen im SPP CLS.

Mit XML, PlainText-Formaten sowie PDF nutzen viele Projekte textbasierte Datenformate, die sich bereits vergleichsweise gut für eine nachhaltige Archivierung und Nachnutzung eignen. Ähnliches gilt für die Nutzung von CSV-Dateien. Dennoch wird deutlich, dass in der Bandbreite genutzter Formate einige Projekte auch proprietäre Lösungen verwenden (Abb. 1), wodurch aus FDM-Perspektive gegen Projektende eine Formatmigration nötig werden könnte.

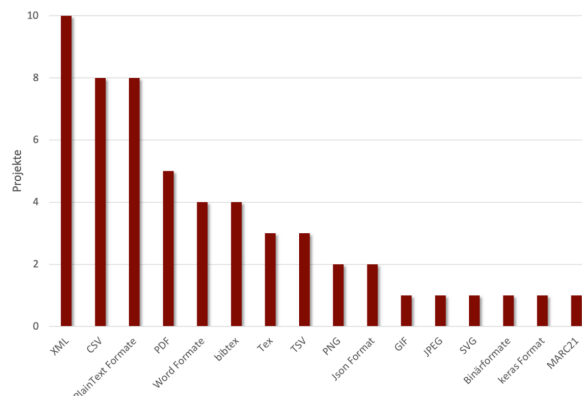


Abb. 1: Genutzte Datenformate im SPP CLS.

Vor dem Hintergrund der hohen Relevanz von Softwarecode (vgl. Tabelle 1) spielt auch die Nutzung von Programmier- und Skriptsprachen in einem zentralen Datenmanagement eine wichtige Rolle (Tabelle 2).

Programmier-/Skriptsprache	Projekte
Python	9
R	4
Shell-Skripte	4
Java	3
X-Technologien	2
JavaScript	2
HTML	2
CSS	2
SQL	1

Tab. 2: Genutzte Programmier- und Skriptsprachen im SPP CLS.

In diesem Zusammenhang sind gleichzeitig der Umgang mit lebenden Systemen (Tabelle 3) und hier verwendeter Technologie-Stacks (Abb. 2) am Ende der Projektlaufzeit eine zentrale Herausforderung. Insbesondere die durch die Projekte teilweise selbst motivierte Nutzung von statischen Systemen wie bspw. Jekyll wird die langfristige Verfügbarkeit von lebenden Systemen dabei deutlich erleichtert.

Lebende Systeme	Projekte
Website	7
Tools/Anwendungen	3
Bibliotheken	1
Dashboard	1

Tab. 3: Geplante lebende Systeme im SPP CLS.

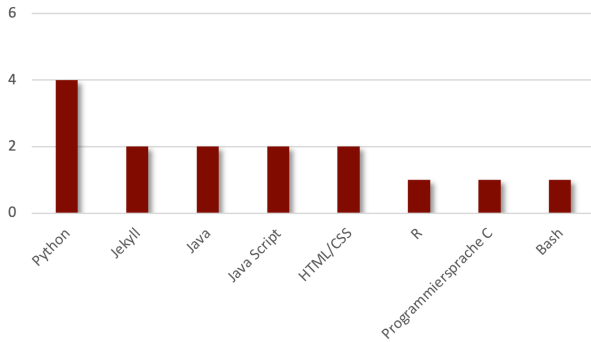


Abb. 2: Geplante Technologie-Stacks zur Entwicklung lebender Systeme im SPP CLS.

Neben der Erfassung von Informationen, die für das FDM relevant sind, war es bei der Landschaftsvermessung im SPP CLS auch ein Ziel Aussagen über methodische und organisatorische Best Practices innerhalb des Forschungsfelds zu treffen.

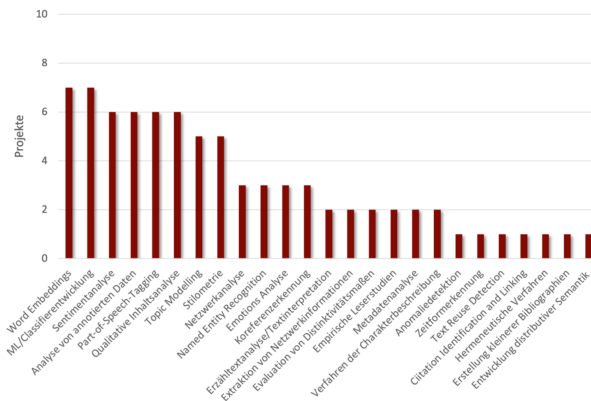


Abb. 3: Analyseverfahren und methodische Werkzeuge im SPP CLS.

Während die Abfrage von angewandten Analyseverfahren und methodischer Werkzeuge innerhalb der einzelnen Projekte erste Trends ablesen lassen (Abb. 3), können Informationen über genutzte Tools zum Projektmanagement sowie zur Annotation von Daten dabei helfen, infrastrukturelle Bedarfe der Community zu identifizieren (Abb. 4 und Tabelle 4).

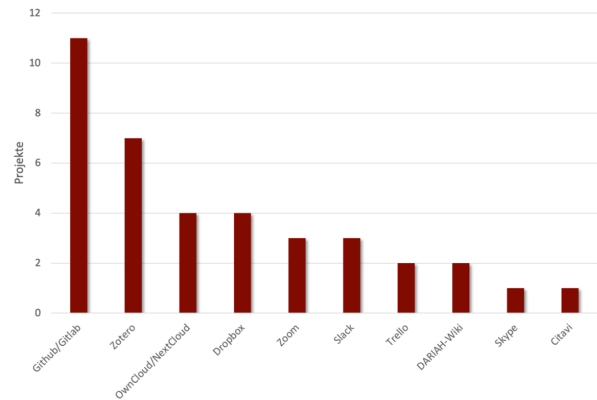


Abb. 4: Genutzte Tools zur Projektverwaltung im SPP CLS.

Annotationstool	Projekte
Catma	5
TreeTagger	1
CorefAnnotator	1
Sentiment Analyzer	1

Tab. 4: Genutzte Annotationstools im SPP CLS.

Darüber hinaus können die Angaben zu Archivierungs- und Publikationsstrategien, trotz ihres noch tentativen Charakters, bereits für die Ableitung von Best Practices innerhalb der Fachdisziplin genutzt werden (Abb. 5 und Abb. 6).

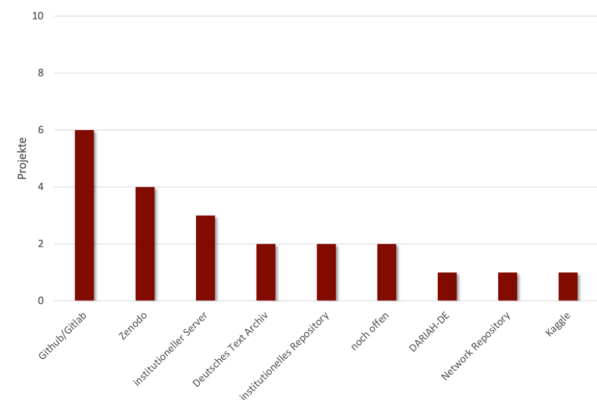


Abb. 5: Tentative Angaben zu genutzten Archivierungsinfrastrukturen im SPP CLS.

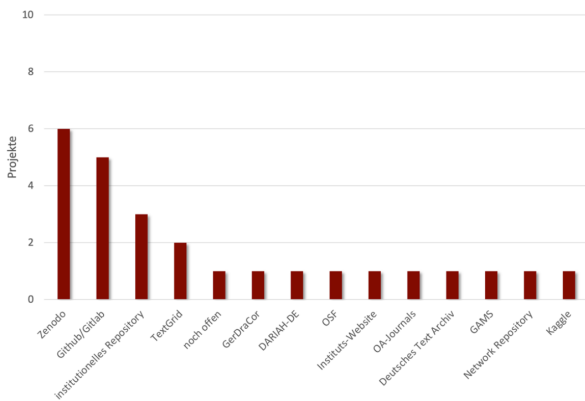


Abb. 6: Tentative Angaben zu genutzten Publikationsinfrastrukturen im SPP CLS.

Qualitative Auswertung: Abgeleitete FDM-Maßnahmen

Beispielhaft für die pragmatische Identifikation und Umsetzung von Maßnahmen zur Handhabung der Bedarfe im Forschungsdatenmanagement in den CLS gehen wir hier auf zwei Aspekte ein.

Zum Arbeiten in und zum Austausch zwischen den Projekten sowie der projektübergreifenden Arbeit einzelner Arbeitsgruppen wurde nach einer geeigneten Plattform gesucht. Dabei waren folgende Kriterien ausschlaggebend:

- Die Zusammenarbeit muss über Fach-, Universitäts-, und Ländergrenzen vollumfänglich möglich sein.
- Die Ablage und der Austausch von Daten muss möglich sein, idealerweise auch das gemeinsame Arbeiten auf der Plattform, die daher zumindest eine Versionierung zur Verfügung stellen muss.
- Textdateien und Annotationen müssen ebenso verwaltet werden können wie formale Metadaten und Softwarecode.
- Die Gesamtkapazität muss im hohen GB-Bereich liegen, da Datensätze sowie Modelle aus dem Maschinellen Lernen im zweistelligen GB-Bereich als Einzeldateien zu erwarten sind.
- Bereiche für Projekte, Projektgruppen sowie Arbeitsgruppen müssen leicht angelegt und verwaltet werden können.
- Der Speicherort muss bekannt und sicher sein, damit rechtliche Belange, bspw. urheberrechtlicher Natur, gewahrt und dem Missbrauch der Daten vorgebeugt werden kann.

Cloud-Lösungen sind für den Datenaustausch und das kollaborative Arbeiten oft die erste Wahl. Jedoch sind sie meist auf bestimmte, ggf. lokale Nutzergruppen ausgerichtet (z.B. Dienste für Hochschulen eines Bundeslandes), bezüglich des Speicherorts intransparent oder mit kommerziell tätigen Unternehmen verbunden, was durch unterschiedliche Richtlinien der Universitäten ebenfalls nicht alle beteiligten Projekte einschließen kann.

Letzteres trifft auch auf zentral zugreifbare Entwicklungsplattformen wie Github oder Gitlab zu, die allerdings den zusätzlichen Vorteil haben, dass sie auch für gemeinsame Codeentwicklung zur Verfügung stehen und Features zum Projektmanagement, wie z.B. Ticketsysteme anbieten.

Wenngleich es einen wachsenden, disziplinübergreifenden Bedarf an der Nutzung von kollaborativen Versionierungssystemen gibt, der sich u.a. aus den mittlerweile breit aufgestellten Schu-

lungs- und Workshop-Angeboten ableiten lässt,⁴ gibt es gleichzeitig einen Mangel an standortübergreifend nutzbaren, zentralen Angeboten solcher Systeme. Entsprechend wurde, trotz eines hohen Betreuungs- und Verwaltungsaufwandes, eine eigene Gitlab-Instanz für das SPP CLS auf universitären Servern aufgesetzt. Dabei war neben der Erfüllung der Kriterien ausschlaggebend, dass im Schwerpunktprogramm bereits Erfahrungen im Umgang mit git-basierten Lösungen vorhanden waren. Für große Einzeldateien wurde das sogenannte Large File Storage zur Verfügung gestellt. Backups erfolgen durch die Infrastruktur des universitären Rechenzentrums. Die Instanz wird durch das Datenkoordinationsteam des SPP CLS betrieben und verwaltet. Da durch die große Funktionalität von Gitlab eine gezielte Verwendung mit Einstiegshürden verbunden sein kann, wurde eine spezifische Dokumentation zusammengestellt und ein Einstiegsworkshop für die Mitglieder des Schwerpunktprogramms organisiert.

Ein zweiter Aspekt ist der Bedarf einer zentralen Publikationsplattform, z.B. für Materialien, die nicht in fachspezifischen Repositorien oder Publikationsorganen unterkommen (Posterpräsentationen, Folien, Handreichungen, aber perspektivisch auch Datendumps oder Snapshots lebender Systeme). Wichtige Anforderungen sind hierbei eine langfristige Auffindbarkeit und Zitierbarkeit sowie die Möglichkeit Ergebnisse der verschiedenen Projekte im Projektverbund gemeinsam sichtbar zu machen:

- Die abgelegten Daten sollen bei einer Speicherinstitution liegen, bei der klar ist, wo die entsprechende Infrastruktur unterhalten wird und wer darauf Zugriff hat.
- Die Dauerhaftigkeit der Speicherinstitution sollte gegeben sein.
- Die Ablage von Daten sollte, auch in größeren Mengen, für Forschende ohne zusätzliche Kosten möglich sein.
- Eine maximale Daten-/Dateiobergrenze sollte es nicht geben. Der verfügbare Speicherplatz sollte mindestens im zweistelligen GB-Bereich liegen.
- Die Speicherinstitution sollte die Vergabe von persistenten Identifiern ermöglichen.
- Mit Hilfe von Versionierung und möglichst auch versionierbarer, persistenter Identifier sollten verschiedene Zustände von Publikationen, Daten und sonstigen Materialien veröffentlicht werden können.
- Die Vergabe von Lizenzen und Möglichkeit eines abgestuften Zugriffs sollte unterstützt werden, um ggf. auch rechtlich geschützte Materialien gesammelt abzulegen.
- Technische Komponenten des Systems sollten transparent sein.
- Zur Steigerung der Auffindbarkeit und Nachnutzung von Publikationen sollte die Speicherinstitution Schnittstellen zu anderen Portalen anbieten und Metadaten an weitere Onlinekataloge weitergeben.

Unterschiedliche projektfinanzierte Repositorien und Publikationssysteme stellen zwar mögliche Lösungen für die skizzierten Anforderungen dar, allerdings können Förderstrategien, insbesondere innerhalb der deutschsprachigen Wissenschaftslandschaft, deren langfristige Weiterfinanzierung und somit die Dauerhaftigkeit eines Services i.d.R. nicht gewährleisten. Bereits institutionalisierte Angebotsstrukturen verfügen hingegen häufig entweder über eine Begrenzung des Adressatenkreises, oder haben sich fachlich oder formatspezifisch stark spezialisiert. Die Nutzung von wirtschaftlich-kommerziellen Angeboten kann wiederum, abgesehen von wissenschaftsethischen Einwänden und beschränkenden universitären Richtlinien, von den nutzenden Projekten und Wissenschaftler*innen selbst auf Dauer nicht getragen und finanziert werden.

Zur Bedienung der skizzierten Bedarfe wurden zwei Lösungen identifiziert und in die Datenstrategie des SPP CLS integriert: Zunächst wurde eine eigene Community für das Schwerpunktprogramm auf dem Online-Speicherdienst Zenodo eingerichtet, welcher nahezu allen Anforderungen entspricht.⁵ Das generische Repositorium ist mittlerweile in vielen Fachdisziplinen als Dienst etabliert. Es ermöglicht die Verwendung reichhaltiger Metadaten zur Beschreibung von Publikationen, unterstützt die Vergabe von versionierten Digital Object Identifiern (DOI), gibt Metadaten an aggregierende Portale wie bspw. OpenAIRE weiter und wird vom CERN in der Schweiz dauerhaft betrieben.⁶

Für die langfristige Veröffentlichung von lebenden Systemen wie bspw. Websites, Tools und einfacher Anwendungen stellt in Ergänzung Github eine Lösung dar. Auch der netzbasierte Dienst zur Versionsverwaltung verfügt in diversen wissenschaftlichen Communities, obwohl er mittlerweile von Microsoft betrieben wird, über einen großen Nutzendenkreis und kann für den abgegrenzten Gegenstandsbereich einiger lebender Systeme in Frage kommen: Github ermöglicht neben der kollaborativen Entwicklung auch die dokumentierte und quelloffene Bereitstellung von Software und verfügt über eine Schnittstelle zu Zenodo, wodurch Github-Repositorien in einem bestimmten Zustand auf Zenodo publiziert und persistent referenzierbar gemacht werden können.

Weitere Perspektiven auf eine gemeinsamen Datenstrategie

Trotz erster pragmatischer und bedarfsorientierter Lösungsstrategien im Rahmen der Entwicklung einer gemeinsamen Datenstrategie innerhalb des SPP CLS hat die erste Review- und Analysephase der Landschaftsvermessung deutlich gemacht, dass innerhalb des Schwerpunktprogramms sehr heterogene Bedingungen und Bedarfe in Bezug auf das Forschungsdatenmanagement vorherrschen. Unterschiedliche Methoden angewandt auf verschiedene Korpora erzeugen teilweise individuelle Forschungsdaten und -ergebnisse, die es sowohl projektintern als auch für das gesamte SPP CLS langfristig zu sichern sowie verfügbar zu machen gilt.

Eine zentrale Herausforderung der zweiten Review- und Analysephase, mit Fokus auf der Konkretisierung individueller Archivierungs- und Publikationsstrategien innerhalb der einzelnen Projekte, ist die Integration etablierter Vorgehensweisen in die gemeinsame Datenstrategie des SPP CLS. Darüber hinaus wird es eine Hauptaufgabe sein die Entwicklung von lebenden Systemen innerhalb der einzelnen Projekte so weit zu begleiten und zu betreuen, dass möglichst alle individuellen Websites, Tools und kleineren Anwendungen auch über die Projektphasen hinaus in einer statischen Form mit geringem Kurationsaufwand, bspw. via Github und Zenodo, auffindbar, zugänglich, interoperabel und nachnutzbar bleiben.

In unserem Vortrag werden wir das Schwerpunktprogramm als Blaupause für das Forschungsdatenmanagement innerhalb der CLS dezidiert beschreiben und die hier beschriebenen Ergebnisse mit weiteren Erkenntnissen, auch zu möglichen fachspezifischeren Lösungen, die wir aus der bis dahin abgeschlossenen zweiten Review- und Analysephase gewinnen werden, kompletieren.

Fußnoten

1. Vgl. <https://www.nfdi.de/> und <https://www.gwk-bonn.de/themen/weitere-arbeitsgebiete/informationsinfrastrukturen-nfdi/> (letzter Zugriff: 14. Juli 2021).
2. DFG Schwerpunktprogramm „Computational Literary Studies“, Online: <https://dfg-spp-cls.github.io/> (letzter Zugriff: 09. Juli 2021).
3. Interviewleitfaden zur FDM-Bestandsaufnahme im Schwerpunktprogramm „Computational Literary Studies“, Online: <http://doi.org/10.5281/zenodo.4269639>.
4. Vgl. u.a. diverse Workshops zum Forschungsdatenmanagement mit Gitlab, durchgeführt durch die Landesinitiative für Forschungsdatenmanagement (fdm.nrw), Online: <https://www.fdm.nrw/index.php/fdm-nrw/versionierung-gitlab/> (letzter Zugriff: 09.11.2021); „Workshop: Git und Gitlab für Anfänger*innen“, durchgeführt durch die Landesinitiative FDM Thüringen, 21. Juli 2021, Online: <https://forschungsdaten-thueringen.de/veranstaltung/workshop-git-gitlab-de.html> (letzter Zugriff: 09.11.2021); „git and GitLab basics workshop“, durchgeführt durch NFDI4Ing, 21.09.2021, Online: <https://nfdi4ing.de/git-and-gitlab-basics-workshop-3/> (letzter Zugriff: 09.11.2021); Workshop „Datenversionierung mit Git – Advanced Track“, durchgeführt durch Carolin Odebrecht auf der RDA Deutschland Tagung 2020, Potsdam, 25.02.2020, Online: <https://www.rda-deutschland.de/presentationen-2020/gitlabworkshop2020.pdf> (Folien) (letzter Zugriff: 09.11.2021).
5. Online: <https://zenodo.org/>
6. Online: <https://www.openaire.eu/>

Bibliographie

- Bryant, Rebecca / Lavoie, Brian / Malpas, Constance** (2017): *A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1*. Dublin, Ohio: OCLC Research. DOI: <https://doi.org/10.25333/C3PG8J>.
- DFG - Deutsche Forschungsgemeinschaft** (2019): *Guidelines for Safeguarding Good Research Practice. Code of Conduct*. Zenodo: <http://doi.org/10.5281/zenodo.3923602>.
- Kleinkopf, Felicitas / Jacke, Janina / Gärtner, Markus** (2021): „Text- und Data-Mining: urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora und ihrer Bedeutung für die Digital Humanities“ in: *MMR: Zeitschrift für IT-Recht und Recht der Digitalisierung*, Jahrgang 2021, Heft 3. München: C.H.BECK oHG 196 ff. Online: <http://dx.doi.org/10.18419/opus-11445>.
- Pempe, Wolfgang** (2012): „Geisteswissenschaften“ in: Neuth, Heike / Strathmann, Stefan / Oßwald, Achim / Scheffel, Regine / Klump, Jens / Ludwig, Jens (eds.): *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Boizenburg: Verlag Werner Hülsbusch 137-160.
- RfII 2016, RfII - Rat für Informationsinfrastrukturen** (2016): *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen. Online: <https://d-nb.info/1104292440/34> (letzter Zugriff: 14. Juli 2021).
- RfII - Rat für Informationsinfrastrukturen** (2017): *Schritt für Schritt - oder: Was bringt wer mit? Ein Diskussionsimpuls für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI)*. Göttingen, Online: <https://d-nb.info/1131083113/34> (letzter Zugriff: 14. Juli 2021).

Schöch, Christof / Döhl, Frédéric / Rettinger, Achim / Gius, Evelyn / Trilcke, Peer / Leinen, Peter / Jannidis, Fotis / Hinzmänn, Maria / Röpke, Jörg (2020): „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen“ in: *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel, Online: https://doi.org/10.17175/2020_006.

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand Jan / Appleton, Gabrielle / Axton, Myles / Baak, Arie / Blomberg, Niklas / Boiten, Jan-Willem / da Silva Santos, Luiz Bonino / Bourne, Philip E. / Bouwman, Jildau / Brookes, Antony J. / Clark, Tim / Crosas, Mercè / Dillo, Ingrid / Dumon, Oliver / Edmunds, Scott / Evelo, Chris T. / Finkers, Richard / Gonzalez-Beltran, Alejandra / Gray, Alasdair J.G. / Groth, Paul / Goble, Carole / Grethe, Jeffrey S. / Heringa, Jaap / A.C't Hoen, Peter / Hooft, Rob / Kuhn, Tobias / Kok, Ruben / Kok, Joost / Lusher, Scott J. / Martone, Maryann E. / Mons, Albert / Packer, Abel L. / Persson, Bengt / Rocca-Serra, Philippe / Roos, Marco / van Schaik, Rene / Sansone, Susanna-Assunta / Schultes, Erik / Sengstag, Thierry / Slater, Ted / Strawn, George / Swertz, Morris A. / Thompson, Mark / van der Lei, Johan / van Mulligen, Erik / Velterop, Jan / Waagmeester, Andrea / Wittenburg, Peter / Wolstencroft, Katherine / Zhao, Jun / Mons Barend (2016): „The FAIR Guiding Principles for scientific data management and stewardship“ in: *Scientific Data* 3, Article number: 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>.

Praktiken der digitalen Erinnerung an den 2. Weltkrieg Netzwerkmodellierungen des „Axis History Forum“

Glawion, Anastasia

anastasia.glawion@tu-darmstadt.de
TU Darmstadt, Germany

Die digitale Erinnerung (*digital memory*) ist ein zentraler Begriff in der Erinnerungsforschung, welcher sich vor allem durch die Arbeiten von Andrew Hoskins etabliert hat (z.B. Hoskins 2018a). Hoskins beobachtet die Veränderungen der Erinnerungskultur, die durch digitale Medien initiiert werden, und beschreibt Prozesse, die als Abkehr von dem traditionellen Begriff der kollektiven Erinnerung verstanden werden können (Hoskins 2018b: 85–86). Im digitalen Rahmen, so Hoskins, würde eine andere Art der sozialen Formationen zustande kommen, die er *multitudes* genannt hat (Hoskins 2018b: 86). Im Kontrast zu der klassischen Assmann'schen Dichotomie des kulturellen und des kommunikativen Gedächtnisses (Assmann 2005), die ein abstraktes Kollektiv als Träger des Gedächtnisses voraussetzt, bilden *multitudes* Strukturen der digitalen Erinnerung ab, die aus Formationen von Nutzenden bestehen und auf einer Ebene zwischen Individuum und Kollektiv angesiedelt sind.

Ähnliche Ansätze waren in vielen Studien der Internet Studies präsent, wo durch die Anwendung der Netzwerkanalyse homophile Cluster auf verschiedensten Online-Plattformen entdeckt wurden (z. B. Wojcieszak/Mutz 2009; Himelboim et al. 2016;

Barnett/Benefield 2017; Bond/Sweitzer 2018). Es fehlten allerdings Anwendungsbeispiele aus dem Bereich der Memory Studies, die Daten von gedächtnisrelevanten Online-Communities theoriegeleitet analysieren würden, und somit eine Brücke zwischen dem theoretischen Konzept der digitalen Erinnerung und der netzwerkasierten Online-Forschung schlagen würden.

Inspiziert von dieser Leerstelle und darüber hinaus von Jeffrey Olicks und Joyce Robbins' Idee, den Fokus der empirischen Erinnerungsforschung auf Praktiken der Erinnerung zu legen (Olick/Robbins 1998), interpretierte ich in meiner im Sommer 2021 verteidigten Dissertation „Practices of transnational Memory – A Mixed Methods Approach to the Study of a historical online Forum“ die Interaktionen innerhalb geschichtlicher Foren als Praktiken transnationaler Erinnerung. Der Begriff „transnational“ betonte dabei den Übergang der Gedächtnisproduktion von nationalen Instituten zu anderen „Trägern der Transnationalität“, zu denen man unter anderem digitale Medien zählt (Assmann/Conrad 2010: 2-4). Die Kulturen des digitalen Gedächtnisses wurden somit über unterschiedliche Praktiken auf dem militärhistorischen Axis History Forum (AHF) operationalisiert.

Das Forum umfasst ein großes Datenmassiv,¹ das sich organisch über einen Zeitraum von fast 20 Jahren angesammelt hat. Das *too big to read*-Argument, welches die Verwendung von computer-gestützten Methoden in der Literaturwissenschaft motivierte, gilt in dem Fall auch für nicht-literarische Texte. Im Rahmen meiner Dissertation entwarf ich einen methodischen Zugang zu diesem Datenmassiv, in dessen Kern ein Dreischritt aus Netzwerkmodellierungen liegt. Der aktuelle Beitrag fasst das Vorgehen der Dissertation zusammen, und zeigt an diesem Beispiel, wie der netzwerkanalytische Dreischritt an entscheidenden Stellen gut begründete Informationsreduktion ermöglicht.

Vorgehen

Im Rahmen der Dissertation wurden drei Forschungsfragen beantwortet:

1. Wie positioniert sich AHF zu anderen Online-Ressourcen, die innerhalb der Militärgeschichtsgemeinschaft populär sind?
2. Welche Gruppen bilden Nutzer:innen durch ihre Interaktionen auf dem Forum und welche Themen sind in den Diskussionen dieser Gruppen repräsentiert?
3. Welche Arten von Diskussionen führen Nutzer:innen dieser Gruppen?

Um die erste Forschungsfrage zu beantworten, wurden die Entstehungsgeschichte des Forums, die Forums- und Moderationsregeln, die Links auf statischen Elementen der Webseite und die Veränderung der von den Moderator:innen eingeführten Unterforumstruktur seit 2002 ausgewertet. Dieser Zugang zu AHF als einem Internet-Artefakt bot viele interessante Einblicke in die Selbstinszenierung des Forums als Gemeinschaft „seriöser“ Forscher mit einem Fokus auf die Geschichte der Achsenmächte. Es stellte sich heraus, dass AHF unter anderem mit Hilfe der visuellen Elemente eine breite Nutzerschaft anspricht, die von professionellen Militärhistoriker:innen bis hin zu Nationalsozialismussympathisant:innen reicht. In den Forumsregeln hingegen sind strikte Bedingungen für die Diskussionen festgehalten (bspw. Verbot der Glorifizierung von Nationalsozialismus, Verbot der Holocaustleugnung usw.), die durch die Löschoptionen und Bannmöglichkeiten für Moderator:innen verstärkt werden. Diese Strategie führt dazu, dass stetig Interaktionen auf dem Forum stattfinden, aber auch, dass einige Fragen repetitiv behandelt werden.

Um die anderen beiden Forschungsfragen zu beantworten, wurde ein formalisierter Ansatz gewählt, der die Anwendung der Netzwerkanalyse voraussetzt. Dafür mussten Postinhalte, Postmetadaten und Nutzer:inneninformationen von der Webseite extrahiert werden, wozu das rvest-Package von R Studio verwendet wurde (Wickham 2016). Anschließend wurde eine Adjazenzmatrix der Nutzer:innenbeziehungen erstellt, in der eine Verbindung zwischen zwei Knoten dann eingezeichnet wurde, wenn die durch die Knoten repräsentierten Nutzer:innen Kommentare in derselben Diskussion hinterlassen hatten.

Dieses Netzwerk beinhaltete über 25.000 Knoten und über 2 Millionen Kanten. Um besonders dichte Untergruppen zu finden, wurden Kanten mit dem Wert 1 rausgefiltert, und anschließend ein Modularitätsclustering angewendet, mit dem Clusterstrukturen innerhalb großer Netzwerke besonders gut erkannt werden (Clauaset/Newman/Moore 2004). Dieser Algorithmus teilte die Knoten in 10 vergleichbar große Cluster, innerhalb derer die Nutzer:innen besonders viel miteinander diskutiert hatten.

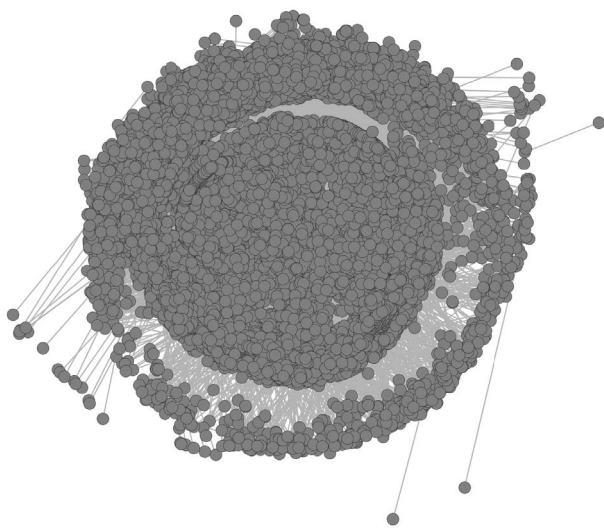


Abb. 1: Netzwerkmodell der Nutzer:innenbeziehungen von AHF.

Themenstruktur von AHF

Das Modularitätsclustering führte zu einer Aufteilung in dichte Untergruppen, wobei die Gründe für das Clustering vorerst unklar waren. Zunächst wurde davon ausgegangen, dass die Cluster um unterschiedliche thematische Schwerpunkte entstanden sind. Um das zu überprüfen, eignete sich die Methode des Topic Modeling sehr gut. Die Clusterkorpora wurden mit Hilfe von LDA-basiertem Topic Modeling in Mallet untersucht (McCallum 2002), nachdem das Korpus lemmatisiert und die englischen Stoppwörter entfernt wurden. Anstatt die Topics eines Modells nur zu kategorisieren, wurden die Begriffsüberschneidungen zwischen Topics als ein Netzwerkmodell dargestellt (s. Abbildung 2 unten).

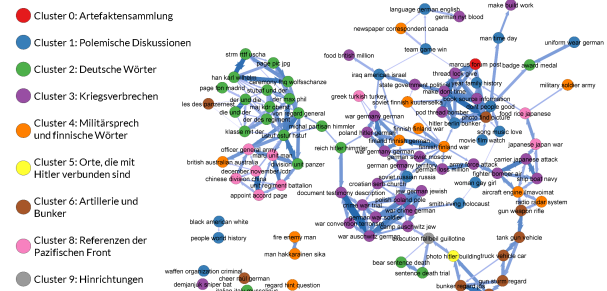


Abb. 2: Begriffsüberschneidungsnetzwerk der Topic-Modeling-Daten mit Simmelian-Backbone-Transformation (Nick et al. 2013). Knoten stellen Topics dar; eine Verbindung zwischen zwei Knoten besteht, wenn Begriffsüberschneidungen innerhalb der ersten 100 Wörter vorhanden sind. Die Farbe der Knoten weist auf das Clusterkorpus, in dem das Topic mit der höchsten Wahrscheinlichkeit zu finden ist. Die Knoten sind benannt nach den ersten drei Wörtern des repräsentierten Topics. Innerhalb des Forumkorpus waren, trotz der ausdrücklichen Vorgabe auf Englisch zu diskutieren, auch zahlreiche deutsche Texte dabei, die vor allem von Nutzern des Cluster 2 (grün) verwendet wurden. Im Preprocessing-Stadium wurden nur englische Stopwords entfernt.

Das Modell in Abbildung 2 wurde mit einem Simmelian-Backbone-Algorithmus gefiltert, welcher nur die Kanten behält, die Teil von einer besonders hohen Anzahl an Dreiecken sind. Mit Hilfe dieser Darstellung sieht man, dass einige Cluster eine höhere thematische Kohärenz haben: Diskussionen der Nutzer:innen von dem grünen Cluster 2 beinhalten viele deutsche Wörter, das daneben liegende Cluster 8 (rosa) zeigt einige Verweise auf die Pazifische Front. In der Mitte sieht man eine dichte Untergruppe von Diskussionen über Kriegsverbrechen (Cluster 3, lila). Darüber befinden sich viele Topics, die Wörter beinhalten, die ich „Militärsprech“ bezeichnet habe: man merkt, dass die Diskussionen Kriegsepisoden referenzieren, aber es gibt keine konkreten Verweise zu Orten oder Schlachten. Cluster 4 ist eine Ausnahme: neben Militärsprech schließt es geografische Referenzen zu Finnland mit ein. In der rechten unteren Ecke befinden sich Topics, die die materielle Ausstattung der Armeen thematisieren: große und kleine Artillerie, Waffen und Bunkerbauten (Cluster 6 und 5).

Netzwerkvisualisierungen von Clusterdiskussionen

Die Topic-Modeling-Ergebnisse lieferten einen ersten Einblick in den Inhalt der Clusterkorpora, doch eine genauere Betrachtung der Forumdiskussionen war notwendig. Im letzten Schritt wurde ein Sample aus 50 Diskussionen aus jedem Cluster gewählt, gelesen und in Kategorien unterteilt. Zusätzlich zum Lesen der insgesamt 500 Diskussionen wurde für jedes der 10 Cluster eine bimodale Netzwerkmodell der Nutzer:innen und der Diskussionen erstellt, was eine bessere Einschätzung der Position der 50 gelesenen Diskussionen ermöglichte. Darüber hinaus bekam man mit Hilfe des Netzwerkmodells Zugang zu strukturell äquivalenten Diskussionsgruppen: besonders große strukturell äquivalente Diskussionsgruppen wurden gesichtet, um anschließend die Kategorienunterteilung des Samples zu bewerten.

Mit Hilfe der bimodalen Netzwerkdarstellung konnte bspw. herausgefunden werden, ob es zu einer Diskussion strukturell äquivalente Diskussionen gibt, und somit – ob eine Diskussion typisch für ein Cluster ist. Bei mehreren Clustern konnte festgestellt werden, dass große Teile des Diskussionssamples ähnlich sind: beispielsweise waren 25 von 50 Diskussionen im Sample von Clus-

ter 0 Anfragen zur Preiseinschätzung von Artefakten aus dem Zweiten Weltkrieg. Durch den Einsatz des bimodalen Netzwerks konnte diese Feststellung durch weitere Belege von strukturell äquivalenten Diskussionsgruppen der gleichen Art verfestigt werden.

Neben zusätzlicher Evidenz erleichterte diese Netzwerkdarstellung Urteile über ein Cluster zu fällen. Cluster 5 in Abbildung 3 ist hierfür ein Beispiel. Anhand des Diskussionssamples war es uneindeutig, ob das Cluster sich mit der Biografie Hitlers auseinandersetzt, oder einen besonderen Wert auf Orte legt, die im Kontext der Geschichte des Dritten Reiches wichtig waren. Die Hervorhebung von den zehn Diskussionen mit der höchsten Zentralität zeigt, wie stark der Fokus auf den Erinnerungsorten liegt. Die zusätzliche Betrachtung von strukturell äquivalenten Diskussionen deutete ebenfalls darauf hin.

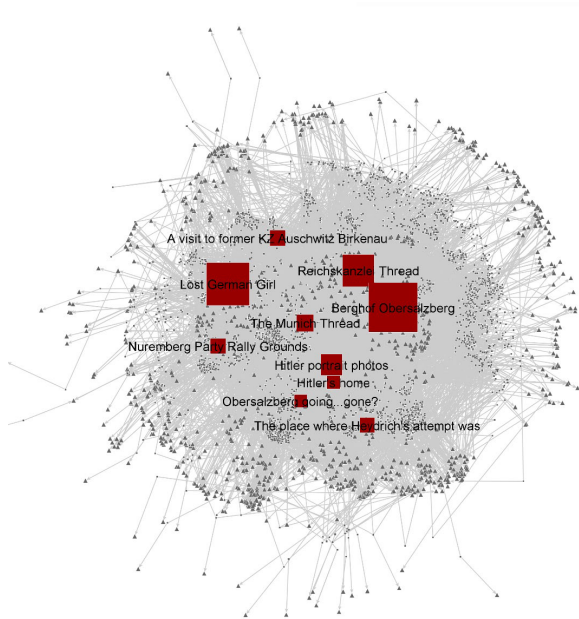


Abb. 3: Bimodales Netzwerk von Cluster 5.

Klassifizierung der Erinnerungspraktiken

Das Topic Model zeigte bereits, dass einige Cluster eine höhere thematische Kohärenz aufgezeigt hatten als andere. Aufbauend auf der Klassifizierung der Diskussionen kam ich zu der folgenden Klassifizierung der Erinnerungspraktiken auf AHF (Abb. 4).



Abb. 4: Schematische Darstellung der Klassifikation von Praktiken digitaler Erinnerung auf AHF und der Quellen, die die Nutzer:innen in den Diskussionen erwähnten.

Empirische Praktiken der Erinnerung

Vier Cluster wurden in die Gruppe der empirischen Erinnerungspraktiken aufgenommen, davon zwei, die sich mit Artefakten aus dem Zweiten Weltkrieg auseinandersetzen, und zwei die eine besondere Aufmerksamkeit Plätzen der Erinnerung geschenkt haben. Interessierte an Bunkern (Cluster 6) behandelten diese wie austauschbare Erinnerungsräume, während Nutzer:innen von Cluster 5 Erinnerungsorte im geschichtlichen Kontext betrachteten und ihnen eine spezifische Bedeutung zuschrieben. Nutzer:innen, die an diesen Erinnerungspraktiken interessiert waren, nannten oft persönliche Erfahrungen mit Artefakten oder Orten als Ursprung ihres Interesses am Zweiten Weltkrieg. Erlebnisse und ihre körperliche Ebene spielten eine große Rolle in diesen Diskussionen: das galt sowohl für Nutzer:innen, die Orte besucht haben, als auch für die Verhandlungen in Cluster 7 und Cluster 0, wo unterschiedliche Arten des Artefaktentausches kommuniziert wurden.

Konversationelle Praktiken der Erinnerung

Die zweite Gruppe der Erinnerungspraktiken umfasste ebenfalls 4 Cluster, von den sich das größte, Cluster 3, akademisch geprägten Diskussionen über den Holocaust und Kriegsepisoden widmete. Cluster 9 beschäftigte sich mit den Details von Hinrichtungen in Deutschland der Nachkriegszeit. Das dritte Cluster beinhaltete polemische Diskussionen und Themen, die den Zweiten Weltkrieg viel seltener referenzierten, als es in anderen Clustern der Fall war. Cluster 1 demonstrierte darüber hinaus einen klaren Dominanz von den USA als Nutzer:innenlocation. Es wurde daher als multidirektionell (Rothberg 2009, 2014) interpretiert, weil die US-Referenzen und die Postingmetadaten deutlich im Kontext der 9-11 Attacken platziert werden konnten. Michael Rothbergs Theorie der multidirektionellen Erinnerung geht auf Verbindungen zwischen unterschiedlichen Erinnerungsnarrativen und Gedenkkulturen ein, die er am Beispiel von Kolonialgeschichte und Holocaust schildert. Studien zeigen, dass die Bush-Administration in dem Framing des Afghanistankrieges diskursive Verbindungen und Assoziationen zum „good war“-Narrativ herstellte (z.B. Bond 2014). Cluster 1 könnte ein Beweis für die Wirksamkeit dieser Strategie sein.

Schließlich konnte Cluster 4, das eine hohe Anzahl an finnischen Nutzer:innen als besonders dem Winter- und Fortsetzungskrieg zugewandt interpretiert werden. Arbeiten der finnischen Historikerin Tiina Kinnunen demonstrieren, dass das Interesse in

dieser Ausprägung im Kontext der neo-patriotischen Bewegung in der finnischen Erinnerungskultur betrachtet werden muss (z. B. Kinnunen/Jokisipilä 2012). Die Anordnung der Cluster innerhalb dieser Gruppe ist nicht zufällig, sondern bildet die steigende Valenz der Diskussionen ab. Diskussionen von Cluster 3 und 9 waren weniger emotional, innerhalb des polemischen Cluster 1 kam es öfter zu hitzigen Debatten, während die konfliktreiche Interpretation des Winterkrieges innerhalb von Cluster 4 oft auch von Nutzer:innen als besonders emotional wahrgenommen wurde. Diese Cluster beschäftigten sich häufig mit Details der Militärgeschichte, mit Stereotypen und Sekundärquellen unterschiedlicher Qualität.

Konservierungspraktiken

Die Interaktionen der letzten beiden Cluster wurden als konservierende, aufbewahrende Praktiken interpretiert. Diskussionen waren hierbei selten, vielmehr bestanden Interaktionen aus Anfragen und Antworten. Dabei übernahmen einige Nutzer:innen Brokerrollen in der Wissensvermittlung: bei Cluster 8 handelte es sich um Anfragen zu Übersetzungen von Archivmaterial aus dem japanischen oder chinesischen. Cluster 2 beinhaltete eine große Anzahl von Anfragen von biographischen Informationen über SS-Funktionäre, während andere Nutzer:innen diese Infos aus früheren Recherchen bereitstellten. Über die Gründe solcher Recherchen sollte noch weiter geforscht werden.

Fazit

Abbildung 5 fasst die Netzwerkanwendungen der Dissertation noch ein Mal zusammen: 1. Netzwerkmodell der Nutzer:innenpraktiken, das mit einem modularitätsbasierten Clustering die Nutzer:innen in Gruppen unterteilt; 2. die Darstellung der Begriffsüberschneidungen in einem Topic Model der Clusterkorpora; 3. die Kontextualisierung der Diskussionssamples mit Hilfe von einem bimodalen Netzwerkmodell unter spezieller Beachtung der strukturellen Äquivalenz und der Zentralitätsmaße von Diskussionen. Somit konnte ein Zugang entworfen werden, der dem *too big to read*-Argument entgegenwirkt und den Gegenstand greifbar und untersuchbar macht. Dieser netzwerkanalytische Dreischritt kann auf Nutzer:innennetzwerke jeder Art angewendet werden – im Rahmen des nächsten Papers wird eine Anwendung auf Fanfictionnetzwerke vorbereitet.



Abb. 5: Dreischritt der Netzwerkanwendungen in der Analyse des Axis History Forums.

Fußnoten

1. Auf AHF sind über 80,000 Nutzer:innen registriert, von denen weniger als die Hälfte etwas auf dem Forum geschrieben hat.

Seit März 2002 haben die Nutzer:innen an über 200,000 Diskussionen mit über 2 Mio. Posts teilgenommen. Das dazugehörige Korpus umfasst über 150 Mio. Tokens. Das untersuchte Korpus beinhaltet Kommentare aus dem Zeitraum März 2002-Dezember 2018. Die Hauptsprache des Forums ist Englisch, worauf in den Forumsregeln hingewiesen wird.

Bibliographie

Assmann, Jan (2005): *Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. München: Beck.

Assmann, Aleida / Conrad, Sebastian (2010): „Introduction“ in: Assmann, Aleida / Conrad, Sebastian (eds.): *Memory in a Global Age: Discourses, Practices and Trajectories*. Basingstoke: Palgrave Macmillan 1–16.

Barnett, George A. / Benefield, Grace A. (2017): „Predicting international Facebook ties through cultural homophily and other factors“ in: *New Media & Society* 19: 217–239.

Bond, Lucy (2014): „Types of Transculturality: Narrative Frameworks and the Commemoration of 9/11“ in: Bond, Lucy / Rapson, Jessica (eds.): *The transcultural turn: interrogating memory between and beyond borders*. Berlin, München, Boston: De Gruyter 61–80.

Bond, Robert M. / Sweitzer, Matthew D. (2018): „Political Homophily in a Large-Scale Online Communication Network“ in: *Communication Research* 1–23.

Brandes, Ulrik / Wagner, Dorothea (2013): „Visone: – Analysis and Visualization of Social Networks“ in: Jünger, Michael / Mutzel, Petra (eds.): *Graph drawing software*. Berlin: Springer 321–340.

Clauset, Aaron / Newman, M. E. J. / Moore, Cristopher (2004): „Finding community structure in very large networks“ in: *Physical Review E* 70, 066111.

Himelboim, Itai / Sweetser, Kaye / Tinkham, Spencer F. / Cameron, Kristen / Danelo, Matthew / West, Kate (2016): „Valence-based homophily on Twitter: Network Analysis of Emotional and Political Talk in the 2012 Presidential Election“ in: *New Media & Society* 18, 1382–1400.

Hoskins, Andrew (ed.) (2018a): *Digital memory studies: media pasts in transition*. New York: Routledge, Taylor & Francis Group.

Hoskins, Andrew (2018b): „Memory of the multitude: the end of collective memory“ in: Hoskins, Andrew (ed.): *Digital memory studies: media pasts in transition*. New York; London: Routledge 85–109.

Kinnunen, Tiina / Kivimäki, Ville (ed.) (2012): *Finland in World War II: history, memory, interpretations*. Leiden; Boston: Brill. (= History of warfare volume 69).

McCallum, Andrew Kachites (2002): *Mallet: A Machine Learning for Language Toolkit*.

Nick, Bobo / Lee, Conrad / Cunningham, Pádraig / Brandes, Ulrik (2013): „Simmelian Backbones: Amplifying Hidden Homophily in Facebook Networks“ in: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. The Association for Computing Machinery.

Olick, Jeffrey K. / Robbins, Joyce (1998) „Social Memory Studies: From ‚Collective Memory‘ to the Historical Sociology of Mnemonic Practices“ in: *Annual Review of Sociology* 24, 105–140.

Rothberg, Michael (2009): *Multidirectional Memory. Remembering Holocaust in the Age of Decolonization*. Stanford: Stanford University Press.

Rothberg, Michael (2014): „Multidirectional Memory in Migratory Settings: The Case of Post-Holocaust Germany“ in: De Cesari, Chiara/Rigney, Ann (eds.): *Transnational Memory: Circulation, Articulation, Scales*. Berlin, München, Boston: De Gruyter 123–146.

Wickham, Hadley (2016): *rvest: Easily Harvest (Scrape) Web Pages*. <https://CRAN.R-project.org/package=rvest>.

Wojcieszak, Magdalena E. / Mutz, Diana C. (2009): "Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement?" In: *Journal of Communication* 59, S. 40–56.

Softwarezitation als Technik der Wissenschaftskultur

Vom Umgang mit Forschungssoftware in den Digital Humanities

Henny-Krahmer, Ulrike

ulrike.henny-krahmer@uni-rostock.de
Universität Rostock, Germany

Jettka, Daniel

daniel.jettka@uni-paderborn.de
Universität Paderborn

Einleitung

Software spielt in der gegenwärtigen geisteswissenschaftlichen Forschung eine zentrale Rolle bei der Gewinnung, Anreicherung, Auswertung und Veröffentlichung von digitalen Daten und hat damit einen wesentlichen Anteil an der Schaffung eines digitalen Gedächtnisses. Doch wie steht es um die Erinnerung an die Software selbst im digitalen Kontext?

Mit diesem Beitrag wird anhand einer Analyse der DHd-Konferenzabstracts der Jahrgänge 2016 bis 2020 untersucht, wie Forschungssoftware in den Digital Humanities zitiert wird. Zunächst wird erläutert, welche Relevanz Softwarezitation für die Anerkennung und Nachhaltigkeit von Forschungssoftware in den Digital Humanities hat. Im Anschluss werden bestehende Empfehlungen für Softwarezitation ausgewertet, um Bestandteile von Zitationen zu identifizieren, zu denen Informationen erhoben werden können. Ausgehend von den DHd-Abstracts wird eine Liste erwähnter Software generiert. Für eine Auswahl der so gefundenen Software wird geprüft, auf welche Weise sie zitiert wird und welche Informationen mit den Zitaten gegeben oder weggelassen werden.

Das Ziel der Analyse ist eine Bestandsaufnahme der Praxis der Softwarezitation in den Digital Humanities, in Anlehnung an eine Studie von Howison/Bullard (2016) zu Softwarezitationen in biologischen Forschungspublikationen. Damit soll das Thema der Zitation von Forschungssoftware in den Digital Humanities in den

Fokus gerückt werden, um eine Verbesserung der bestehenden Praxis zu fördern.¹

Relevanz von Softwarezitation

Das Thema Softwarezitation wird im wissenschaftlichen Bereich bisher vor allem aus der Perspektive des Research Software Engineerings (RSE) diskutiert. Es existieren bereits entsprechende Empfehlungen (u. a. Jackson o. D., Smith et al. 2016, Chue Hong et al. 2019a, 2019b, Druskat 2021a, 2021b).

Forschungssoftware zu zitieren hat wichtige Funktionen. Wie bei der Zitation anderer Forschungsergebnisse auch geht es darum, Anerkennung für die wissenschaftliche Leistung auszudrücken, verwendete Quellen offenzulegen und auf sie zu verweisen. In den Digital Humanities gibt es immer wieder eine Diskussion darüber, ob die mit den DH verbundenen Tätigkeiten als Forschung oder als Dienstleistung für Forschung anzusehen sind (Eckhart 2020). Beide Ausprägungen sind möglich,² wodurch es umso wichtiger wird, die Entwicklung von Forschungssoftware in den Digital Humanities als wissenschaftlichen Beitrag zu kennzeichnen und entsprechend zu zitieren.

Dafür bedarf es allerdings einer Definition von Forschungssoftware. Wir verstehen hierunter Software, die für Forschungsfragen, -gegenstände, -daten und -methoden und damit für die Forschungsergebnisse wesentliche Funktionalitäten bereitstellt. Dabei ist unerheblich, ob die Software für Forschungszwecke entwickelt wurde, für diese eingesetzt wird oder z.B. selbst Gegenstand der Forschung ist. Ein Textverarbeitungsprogramm oder ein generischer XML-Editor z. B. fallen in der Regel nicht unter diese Definition, eine für bestimmte Forschungsdaten konfigurierte Datenbank hingegen schon.³

Neben der Anerkennung als wissenschaftliche Leistung sollte Forschungssoftware generell auffindbar sein. Selbst wenn nicht alle Software langfristig lauffähig bleibt, sollten Code oder Beschreibungen verfügbar sein, auf die mit einer Zitation verwiesen werden kann (Smith et al. 2016). So wie Forschungsdaten sollte Software unter Beachtung der FAIR-Prinzipien publiziert werden (Lamprecht et al. 2020). Erst dadurch werden sinnvolle Softwarezitationen möglich und mit Software erzielte Forschungsergebnisse transparenter und nachvollziehbarer.

Empfehlungen und Kriterien für Softwarezitation

In vorhandenen Empfehlungen für Softwarezitation werden Vorschläge gemacht, welche Bestandteile Erwähnungen von Software in wissenschaftlichen Texten, insbesondere in Form von bibliographischen Angaben, haben sollten.

Smith et al. (2016) nennen sechs Prinzipien für die Zitation von Forschungssoftware: *Importance*, *Credit and attribution*, *Unique identification*, *Persistence*, *Accessibility* und *Specificity*. *Importance* meint, dass Software genau wie andere wissenschaftliche Ergebnisse auch in den Metadaten (also der Bibliographie) eines Beitrags aufgeführt werden sollte, wenn sie zitiert wird. *Anerkennung* (*Credit and attribution*) sollte denjenigen zukommen, die tatsächlich zur Entwicklung der Software beigetragen haben. Dies kann bedeuten, dass man die Entwickler:innen als Autor:innen der Software nennt und z. B. nicht Autor:innen einer Publikation über eine Software, da sich beide Gruppen nicht zwingend entsprechen. Software sollte darüber hinaus einen Identifikator haben, der glo-

bal eindeutig, interoperabel und sowohl menschen- als auch maschinenlesbar ist. Dieser Identifikator sollte genau wie Metadaten zur Software persistent sein. Zitationen sollten den Zugang zur Software ermöglichen, indem sie auf die Software selbst (binär oder als Code), auf Metadaten zur Software oder auf Dokumentationen der Software verweisen. Specificity meint, dass Zitationen die Identifikation und den Zugang zu bestimmten Versionen der Software erlauben sollten (Smith et al. 2016).

Abgesehen von den o. g. Prinzipien und Empfehlungen stammen Vorschläge für Softwarezitationen vor allem aus allgemeinen Zitierstilen (u. a. MLA oder APA) oder von Entwickler:innen selbst. Bestandteile dieser Empfehlungen lassen sich in der Regel auf die oben genannten Prinzipien zurückführen.

Auf der Grundlage der genannten Empfehlungen für Softwarezitationen formulieren wir die nachfolgenden Kriterien um die Erwähnung von Forschungssoftware in den DHd-Abstracts zu charakterisieren. Die Kriterien sind als TEI-Taxonomie modelliert und verfügbar in Henny-Krahmer/Jettka (2021) sowie auf GitHub⁴.

- **Bibliographieeintrag für Software (Bib.Soft):** Die Bibliographie enthält einen Eintrag für die Software selbst. Dieser kann den Namen der Software selbst enthalten, Namen von Verantwortlichen, eine URL, einen PID, Versionsangaben, usw.
- **Bibliographieeintrag für Referenzpublikation (Bib.Ref):** Die Bibliographie enthält einen Eintrag mit einer Publikation über die Software.
- **Nur namentliche Nennung der Software (Name.Only):** Die Software ist nur namentlich genannt.
- **Namentliche Nennung der Verantwortlichen (Agent):** Personen, Gruppen oder Institutionen, die für die Entwicklung der Software verantwortlich sind, werden namentlich genannt.
- **URL:** Die Zitation enthält eine URL, die auf die Software selbst verweist (z. B. zu einer Webseite über die Software, einem Code-Repository, einem Metadatensatz oder einer ausführbaren Version).
- **Persistenter Identifikator (PID):** Die Zitation enthält einen persistenten Identifikator (PID), z. B. eine DOI, der auf die Software selbst verweist (z. B. zu einer Webseite über die Software, einem Code-Repository, einem Metadatensatz oder einer ausführbaren Version).
- **Version (Ver):** Die Zitation enthält die Angabe einer bestimmten Softwareversion oder -revision und ggf. anderweitig notwendige Spezifikationen (z. B. eine Version für ein spezifisches Betriebssystem, ein bestimmtes Softwarepaket oder ein Datum).

Bei Empfehlungen wird z. T. zwischen der Perspektive von Software-Anbieter:innen, die Zitiervorschläge machen, und der Perspektive von Nutzer:innen, die Software zitieren, unterschieden. Der Fokus liegt hier auf der Perspektive der Verfasser:innen von wissenschaftlichen Publikationen, in denen Software zitiert wird. Bei der Analyse solcher Zitationen ist zu beachten, dass spezifische Zitiervorschläge von den Anbietenden einen Einfluss darauf haben können, wie die entsprechende Software zitiert wird. Dass solche Vorschläge gemacht werden, ist wichtig und trägt wesentlich dazu bei, dass alle essentiellen Informationen über eine Software verfügbar sind. Insofern stellt eine Analyse von Zitervorschlägen durch Entwickler:innen eine sinnvolle Folgeuntersuchung dar.

Daten und Methoden

Die Datengrundlage für die Analyse zur Praxis der Zitation von Forschungssoftware in den Digital Humanities bilden die Bände der DHd-Konferenzabstracts aus den Jahren 2016 bis 2020, die vom DHd-Verband auf GitHub in PDF- und TEI-Format zur Verfügung gestellt werden.⁵ Die Jahrgänge 2014 und 2015 wurden von der Untersuchung ausgeschlossen, da die Abstracts für diese Jahre nur im PDF-Format verfügbar sind und mit den anderen Jahrgängen bereits eine breite Datenbasis bestehend aus insgesamt 686 Abstracts für Panels, Workshops, Poster und Vorträge mit insgesamt ca. 55.000 Sätzen und 1,2 Mio. Tokens verfügbar ist.⁶

Erschließung von Softwareentitäten

Für die Erhebung der Softwarezitation in den Konferenzabstracts wurde zunächst eine Liste von in den Digital Humanities häufig verwendeter Software erstellt. Diese wurde zum Auffinden konkreter Nennungen in den DHd-Abstracts genutzt. Neben der Auflistung uns bereits bekannter einschlägiger Software⁷ wurden auch automatische Methoden evaluiert, um weitere Benennungen von Software aus den DHd-Abstracts zu erschließen, ohne dass dieser Beitrag damit einen Schwerpunkt auf die Entwicklung automatisierter Verfahren zur Erkennung von Softwareentitäten legt. Vielmehr ging es darum, praktikable Ansätze zur Gewinnung einer Datenbasis zu entwickeln, mit der Softwarezitationen untersucht werden können.

Da existierende Ansätze zu Software Entity Recognition aus der Bioinformatik (Duck et al. 2015) und Biomedizin (Wei et al. 2020) nicht ohne Weiteres auf die Domäne Digital Humanities anwendbar sind, wurde ein Ansatz evaluiert, der allgemeine Named Entity Recognition (NER)⁸ nutzt, um Kandidaten von Software-Benennungen zu ermitteln und im Nachgang auszuwerten. Hierzu wurden die Abstracts mit Hilfe von WebLicht (CLARIN-D/SfS-Uni. Tübingen 2012; Hinrichs et al. 2010) bzw. WebLicht as a Service⁹, automatisch mit Named Entities annotiert. Da der NER-Service¹⁰ nicht auf die Erkennung von Software als Named Entity trainiert ist, bietet sein Einsatz zwar eine Möglichkeit der Annäherung an weitere Kandidaten, allerdings nur in sehr eingeschränktem Maß. Durch manuelle Nachbearbeitung der Liste aller ermittelten Named Entities (insgesamt 29.028), bei der nur Einträge mit mindestens 10 Vorkommen in den DHd-Abstracts betrachtet wurden (910 Named Entities), wurden lediglich 10 Namen von Software ermittelt.

Obwohl der Einsatz allgemeiner NER für die Ermittlung von in den DHd-Abstracts genannter Forschungssoftware keinen großen Mehrwert bieten konnte, steht durch die Kombination der händisch erstellten Auswahl mit den automatisch erzielten Ergebnissen schließlich eine Liste von 138 Softwarenamen zur Verfügung, die für die Ermittlung von Nennungen und Zitationen in den DHd-Abstracts genutzt werden kann.

Erfassung von Softwarezitationen

Auf Basis der ermittelten Softwarenamen können nun Zitationen (und Nicht-Zitationen) von Forschungssoftware in den DHd-Konferenzabstracts erschlossen und klassifiziert werden. Im Weiteren wird der Begriff Zitation auch für reine Namensnennungen (also im engeren Sinn Nicht-Zitationen) verwendet, da diese ebenfalls ausgewertet werden sollen. Die Softwareliste wurde zunächst

anhand der Anzahl vorkommender Instanzen der Software in den Abstracts sortiert, und somit die Analyse häufig genannter Software höher priorisiert, da einerseits möglichst viele Varianten von Zitationen abgedeckt werden sollten und andererseits aufgrund des großen Aufwands nur eine Auswahl von Software erfasst werden konnte.

Mit Hilfe der o. g. TEI-Taxonomie wurden 995 Vorkommen von 32 Softwarenamen in Kombination mit evtl. vorhandenen Zitationen manuell in den TEI-Dateien der DHd-Abstracts annotiert.¹¹ Die verwendete Softwareliste, die TEI-Taxonomie, die annotierten TEI-Dokumente und die Daten zur Auswertung sind verfügbar in Henny-Krahmer/Jetka (2021)¹².

Ergebnisse und Diskussion

Um einen Eindruck von der Zitationspraxis für Forschungssoftware in den DHd-Konferenzabstracts der Jahrgänge 2016 bis 2020 zu erlangen, wurden die manuellen Annotationen der TEI-Dokumente ausgewertet. Da Software häufig mehrfach in einem Abstract genannt, aber sinnvollerweise nicht bei jeder Nennung vollständig zitiert wird, wird das Vorhandensein von Zitationsbestandteilen für jede Software einmal pro Beitrag gezählt und nicht pro Nennung (wenn z. B. einmal im Beitrag eine URL genannt wird, zählt dieses Kriterium als erfüllt). Hierbei ist zu beachten, dass die Zitationsarten (bis auf Name.Only) nicht exklusiv sind, und nicht selten mehrere Zitationsarten mit einer Software verbunden sind. So kann eine Software in einem Beitrag beispielsweise sowohl mit Bib.Soft als auch Bib.Ref zitiert worden sein. Die Verteilung (vgl. Abbildung 1) basiert auf einer Gesamtzahl von $n=218$, welche die einfach gezählten Nennungen einer bestimmten Software in einem bestimmten Beitrag repräsentiert. Die einzelnen absoluten Werte zeigen die Häufigkeit einer Zitationsart einer bestimmten Software, jeweils einfach gezählt pro Abstract, an.

Zitationstyp	Abs. Häufigkeit (n = 218)	Rel. Häufigkeit (%)
Bib.Soft	46	21
Bib.Ref	84	39
Name.Only	46	21
Agent	45	21
URL	114	52
PID	1	0,5
Ver	16	7

Abb. 1: Häufigkeitsverteilung von Softwarezitationen

Die Verteilung der verschiedenen Zitationsarten und -bestandteile zeigt, dass in ca. der Hälfte der erfassten Fälle eine URL für eine Software angegeben wurde, während in nur einem von 218 Fällen ein persistenter Identifikator bereitgestellt wurde. Wenn eine Softwarezitation über einen bibliographischen Eintrag erfolgt, dann zumeist über eine Referenzpublikation (in 39% der Fälle), seltener über einen Bibliographieeintrag für die Software selbst (21%). Ebenfalls für ca. ein Fünftel der betrachteten Softwarenennungen erfolgte eine Nennung von verantwortlichen In-

stitutionen oder Personen, allerdings in ähnlichem Maß auch gar keine Zitation. In 7% der Fälle wurde eine Version der Software genannt.

Die Ergebnisse der Erhebung zeigen, dass Forschungssoftware einerseits zwar in der Regel in Verbindung mit einer der Zitationsarten genannt wird, andererseits jedoch relativ selten eine direkte, langfristige Zitation über bibliographische Einträge für die Software selbst oder unter Verwendung eines persistenten Identifikators erfolgt. Zu prüfen wäre nun, ob sich das Bewusstsein für die Notwendigkeit der nachhaltigen Zitation von Forschungssoftware seit dem Jahr 2016 gewandelt haben könnte und die Berücksichtigung des gesamten Fünf-Jahres-Zeitraums möglicherweise kein adäquates Bild des aktuellen Stands zeichnet. Betrachtet man allerdings die Verteilung der relativen Häufigkeiten der Zitationsarten pro Jahr, ließe sich sogar eine abnehmende Tendenz der Häufigkeit direkter Softwarezitationen in Bibliographien vermuten (vgl. Abbildung 2). Signifikante Aussagen bedürfen jedoch einer größer angelegten Studie.

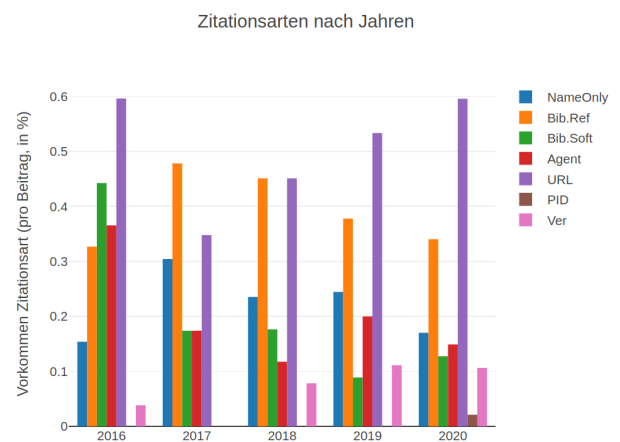


Abb. 2: Relative Häufigkeiten von Zitationsarten pro Jahrgang

Fazit

Ausgehend von der Darstellung der Rolle von Forschungssoftware für den wissenschaftlichen Erkenntnisprozess und den Aufbau eines digitalen Gedächtnisses in den Digital Humanities wurden im vorliegenden Beitrag Empfehlungen für Softwarezitation vorgestellt, die darauf abzielen, Software als wissenschaftliches Werkzeug und Ergebnis adäquat zu identifizieren, anzuerkennen und nachzuhalten.

Um sich ein erstes Bild vom aktuellen Status von Forschungssoftware in den Digital Humanities zu machen, wurden Kriterien für ihre Erwähnung in den DHd-Abstracts der Jahrgänge 2016 bis 2020 formuliert und in Form einer TEI-Taxonomie repräsentiert. Anhand der Kriterien sowie einer Liste von Software, die in den Digital Humanities verwendet wird, erfolgte eine manuelle Annotation ausgewählter Software und ihrer Zitationen in den DHd-Abstracts.

Die vorliegende Bestandsaufnahme der Praxis der Softwarezitation in den Digital Humanities weist deutlich auf Verbesserungsbedarf hin, sowohl im Hinblick auf die Verwendung bibliographischer Einträge für Forschungssoftware selbst und den Einsatz von persistenten Identifikatoren als auch mit Blick auf die Nennung von verantwortlichen Personen und Institutionen, deren Leistun-

gen entsprechend anerkannt werden sollten. Sowohl auf Seiten der Zitationspraxis, also bei Anwender:innen, als auch auf Seiten der Zitationsempfehlungen von Softwareprojekten, also bei Entwickler:innen und Betreiber:innen, besteht in diesem Zusammenhang noch Handlungsbedarf.

Fußnoten

1. Der Hintergrund dieser Einreichung ist das Projekt NF-DI4Culture, in dessen Arbeitsbereich zu Forschungstools und Datendiensten ("Research Tools and Data Services") Überlegungen, Empfehlungen, Beratung und Angebote für die nachhaltige Entwicklung von Forschungssoftware im Kulturerbe fallen (s. a. <https://nfdi4culture.de/de/aufgaben/aufgabenbereiche/aufgabenbereich-3.html>).
2. Siehe z. B. die Unterscheidung zwischen Forschung und Dienstleistung bei den DH an der Uni Bern: <https://www.dh.unibe.ch>, das Service Center Digital Humanities in Münster (<https://www.uni-muenster.de/EScience/schwerpunkte/dh.html>) oder den Lehrstuhl für Digital Humanities in Trier (<https://www.uni-trier.de/universitaet/fachbereiche-faecher/fachbereich-ii/faecher/computerlinguistik-und-digital-humanities/digital-humanities>).
3. Für andere Definitionsvorschläge siehe Hettrick et al. 2014 und Homburg et al. 2020.
4. <https://github.com/daniel-jettka/software-citation-dhd/blob/main/conf/citation-taxonomy.xml>.
5. <https://github.com/DHd-Verband>.
6. Die Satz- und Tokenanzahl wurde ermittelt mit Hilfe des BBAW Tokenizer and Sentence Splitters (WebLicht Service Handle PID: <https://hdl.handle.net/21.11120/0000-0008-3183-C>), der in WebLicht (CLARIN-D/SfS-Uni. Tübingen, 2012) als Webservice zur Verfügung steht.
7. Für eine erste Annäherung wurden neben einer selbst erstellten Liste und Ergebnissen aus einer ersten Sichtung der DHd-Abstracts folgende Quellen herangezogen: die Webseite des Projekts forTEXT (<https://fortext.net/>), in der Zeitschrift RIDE rezensierte Tools (<https://ride.i-d-e.de/issues/issue-11/>) sowie Software, die im Anhang es Projektantrags von NF-DI4Culture genannt ist (öffentliche Fassung: <https://riojournal.com/article/57036/>).
8. Sticker Named Entity Recognizer, <https://github.com/stickeritis/sticker/>.
9. Die Definition der Webservice-Chain und ein Beispieldokument für deren Aufruf sind verfügbar in Henny-Krahmer/Jettka (2021).
10. WebLicht Service Handle PID: <http://hdl.handle.net/211022/0000-0007-DA29-6>.
11. Direkte Anwendung finden alle Kategorien aus der TEI-Taxonomie außer Name.Only, welches sich implizit aus dem Fehlen der anderen Kategorien ergibt.
12. Verschiedene Entwicklungsstände der Datensätze sind ebenfalls abrufbar über <https://github.com/daniel-jettka/software-citation-dhd>.

Bibliographie

CLARIN-D/SfS-Uni. Tübingen (2012): "WebLicht: Web-Based Linguistic Chaining Tool" Online. <https://weblicht.sfs.uni-tuebingen.de/> [Letzter Zugriff: 13.07.2021].

Chue Hong, Neil (ed.) (2019a): "Software Citation Checklist for Authors" (Version 0.9.0). *Zenodo*. <http://doi.org/10.5281/zenodo.3479199>.

Chue Hong, Neil (ed.) (2019b): "Software Citation Checklist for Developers" (Version 0.9.0). *Zenodo*. <http://doi.org/10.5281/zenodo.3482769>.

Druskat, Stephan (2021a): "Research software citation for researchers" *Research Software Citation. Cite and Make Citable!* (Version 1.1). <https://cite.research-software.org/researchers/> [Letzter Zugriff: 13.07.2021].

Druskat, Stephan (2021b): "Research software citation for developers" *Research Software Citation. Cite and Make Citable!* (Version 1.1). <https://cite.research-software.org/developers/> [Letzter Zugriff: 13.07.2021].

Duck, Geraint / Kovacevic, Aleksandar / Robertson, David L. / Stevens, Robert / Nenadic, Goran (2015): "Ambiguity and variability of database and software names in bioinformatics", in: *Journal of Biomedical Semantics* 6 (29). <https://doi.org/10.1186/s13326-015-0026-0>.

Eckhart, Arnold (2020): "Digital Humanities: Is it Research or is it Service?" *dhmuc. Digital Humanities München*. Blog post. 26.7.2020. <https://dhmuc.hypotheses.org/2834> [Letzter Zugriff: 13.07.2021].

Henny-Krahmer, Ulrike / Jettka, Daniel (2021): "Software-citation in den Digital Humanities" (Version 0.1). *Zenodo*. <http://doi.org/10.5281/zenodo.5106391>.

Hettrick, Simon / Antonioletti, Mario / Carr, Les / Chue Hong, Neil / Crouch, Stephen / De Roure, David / Emsley, Iain et al. (2014, dec): "UK Research Software Survey 2014" *Zenodo*. <https://doi.org/10.5281/zenodo.14809>.

Hinrichs, Erhard W. / Hinrichs, Marie / Zastrow, Thomas (2010): "WebLicht: Web-Based LRT Services for German", in: *Proceedings of the ACL 2010 System Demonstrations* 25–29.

Homburg, Timo / Klammt, Anne / Mara, Hubert / Schmid, Clemens / Schmidt, Sophie C. / Thiery, Florian / Trognitz, Martina (2020): "Diskussionsbeitrag - Handreichung zur Rezension von Forschungssoftware in den Altertumswissenschaften / Impulse - Recommendations for the review of archaeological research software" *GitHub*. https://research-squirrel-engineers.github.io/Impuls_SoftwareRezensionen_DGUF/Draft.htm.

Howison, James / Bullard, Julia (2016): "Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature", in: *Journal of the Association for Information Science and Technology (JASIST)* 67 (9): 2137–2155. <https://doi.org/10.1002/asi.23538>.

Jackson, Mike (o. D.): „How to cite and describe software.“ *Software and research: The Software Sustainability Institute's Blog*. Blog post. <https://www.software.ac.uk/how-cite-and-describe-software> [Letzter Zugriff: 13.07.2021].

Lamprecht, Anna-Lena / Garcia, Leyla / Kuzak, Mateusz / Martinez, Carlos / Arcila, Ricardo / Del Pico, Eva M. / Dominguez Del Angel, Victoria et al. (2020): "Towards FAIR principles for research software", in: *Data Science* 3 (1): 37–59. <https://doi.org/10.3233/DS-190026>.

Smith, Arfon M. / Katz, Daniel S. / Niemeyer, Kyle E. / FORCE11 Software Citation Working Group (2016): "Software citation principles", in: *PeerJ Computer Science* 2:e86. <http://dx.doi.org/10.7717/peerj-cs.86>.

Wei, Qiang / Zhang, Yaoyun / Amith, Muhammad / Lin, Rebecca / Laypeyrolerie, Jenay / Tao, Cui / Xu, Hua (2020): "Recognizing software names in biomedical literature using machine learning", in: *Health Informatics Journal* 26 (1): 21–33. <https://doi.org/10.1177/1460458219869490>.

The Digital Archive and the Politics of Digitization

Zaagsma, Gerben

gerben.zaagsma@uni.lu

Luxembourg Centre for Contemporary and Digital History (C²DH)

Much has been, and is, made of the transformative potential of digital resources and historical data for humanities and historical research in recent years. Historians in the global North are flooded with retro-digitised and born-digital materials and tend to take them for granted, grateful for the opportunities they afford. As the late Roy Rosenzweig predicted already in 2003, historians “may be facing a fundamental paradigm shift from a culture of scarcity to a culture of abundance” (Rosenzweig 2003: 739). Yet, if we accept that we do indeed live in a culture of abundance, that abundance is still rarely questioned and qualified, let alone contextualized in time and space. To put it simply: the question of why, where and how we can access what we can access, and how this affects ‘memory’ is rarely posed.

Few historians would deny that archives or libraries are repositories of carefully selected and curated collections and thus far from neutral: “No archive is innocent”, as Elizabeth Yale wrote (Yale 2015: 332). By the same token, the digitisation of historical sources, is far from neutral. In a research environment that increasingly privileges what is available online, where traditional archives are sometimes even referred to as ‘hidden’, and ‘old-fashioned’ browsing is replaced by surgical discovery, we would do well to start imagining what a world of historical scholarship based upon digital resources looks like. Just as the differences between ‘analogue’ sources and their digital, yet equally material, representations are easily overlooked, so too changing modes of access to digital sources are rarely scrutinised for their consequences for historical research. In sum, there is a marked discrepancy between the use of digital resources by many historians and their lack of interest in, or understanding of, how these are created and constituted.

Archives are neither repositories nor *purveyors* of ‘memory’, as so much contemporary discourse would have it: more accurately, they provide (part of) the raw material that feeds into its construction. The ‘archive equals memory’ equation obscures the role of *mediation* in the process of turning archival materials into reconstructions of the past, and the manifold ways in which this influences ‘memory’, be it individual or social/cultural.

Increased access to retrodigitised sources does not imply completeness, even when mass digitisation is concerned. Many materials are not, and will never be, digitised. Indeed, digitisation first and foremost means selection. GLAMs select materials to be digitised on the basis of a variety of criteria. These include the preservation of fragile materials, easy access to collection highlights and/or often-used material, the research value of certain collections and academic research agendas. Memory politics, public discourses on the past, and the articulation of a country’s imagined ‘national’ identity are of similar importance while legal, ethical and copyright frame and constraint digitisation strategies. Given the costs involved, the availability of funding, public or private, plays a key role in enabling digitisation projects in the first place (Zaagsma 2013).

As digitisation entails a selection of already selected analog materials, historians find themselves facing old questions pertaining to new and unfamiliar digital environments. How do digital resources shape the historical themes, topics and debates that can be researched and how might they influence research agendas more broadly? In what ways do they enable us to address new research questions and venture into new research avenues that challenge existing master narratives? Can they facilitate research into transnational histories when most digitisation projects are, in one way or another, so often nationally framed? In sum: what are the histories that we can and cannot tell with digitised cultural heritage, and how could we as historians best navigate the challenges that are involved in using them? What, then, are the politics of digitisation and what are its implications for historical research?

There are many aspects of digitisation that can be considered “political”, from selection for digitisation to modes of access to broader questions about ‘infrapolitics’. None of these is specific to our digital age and historical context is crucially important. Digitisation is only the most recent technological option for heritage preservation and reproduction, which has a history that dates back to the invention of the microfilm in the late 19th century, and the first uses of photography for research purposes in the early 20th. Similarly, the politics of heritage and the political dimensions of heritage preservation, as well as the relation between archives, social memory, knowledge and power have long been discussed by historians, philosophers, archival scientists and heritage scholars. And as long as archives have existed, the question of access has been key in determining *who* writes history.

In this paper I will discuss key parameters of the politics of digitisation within a broader historical and global context with the aim to encourage further debate on its implications for historical research.

In the first part, I will outline the global dimensions of the politics of digital cultural heritage with a particular focus on developments within and between Europe and Africa, framed within the broader context of the politics of heritage and its preservation and recent debates about ‘postcolonial digital humanities’ (Risam 2019). In the second part, I will discuss the history and current state of digitisation in Europe and Africa. Here I will partly draw upon the the IFLA/UNESCO Survey on Digitisation and Preservation that was conducted in 1998, at the dawn of the era of (mass) digitisation, and the web archive of the accompanying IFLA/Unesco Directory of Digitised Library Collections (2002-2006), as well as recent global and European digitisation surveys.

In the European Union area, cultural heritage digitisation is inextricably linked to strengthening a sense of European identity and embedded in a digital agenda that “seeks to optimise the benefits of information technologies for economic growth, job creation and the quality of life of European citizens” (Commission Recommendation 2011). Supranational projects such as Europeana and Time Machine both frame themselves as contributing to a European common identity and history. In the latter case, a video created as part of a marketing campaign explicitly suggested Europe was at a turning point in its history and the Time Machine project would act as savior of a mythical occidental European enlightened past and enabler of a common history (Time Machine Trailer 2019). In Western Europe, where digital resources are comparatively plenty, debates about the effects of digitisation on historical scholarship are relatively muted. In Eastern Europe and Russia, however, the politics of digital heritage are of greater scholarly concern within a context where historians face increasing political pressures, if not active censorship and obstruction (Golubev 2021).

In Africa, digitisation should be seen within a postcolonial context where the geographical overlap between 'nation' and 'state' that many assume in Europe, is absent. In this respect, Kahn and Tanner have pointed to the complex interplay between digitisation and (post-colonial) nation-building and national identity in post-colonial (South)Africa and plead for "build[ing] digital collections that reflect an indigenous African identity, not an imagined Westernised one" (Tanner and Kahn 2014: 125). They follow Premesh Lalu, who earlier argued forcefully for a "politics of digitisation that will expand what can be said about the history of liberation struggles in Southern Africa" (Lalu 2007: 42). The latter points to the much broader context in which digitisation in and within Africa should be situated: North/South relations, the involvement of public and private parties, questions of access, privilege, ownership mix in complex ways which have created distinct concerns that have variously been described as 'digital imperialism' (2000s), the 'complex of the digital savior' (2010s) and appropriation of the discourse on 'endangered archives' (Chamelot, Hiribarren and Rodet 2020).

As will be clear from this very short outline, heritage is highly political in nature, and this is no different in the digital realm, where the struggle for 'memory' and the past increasingly takes place. This plays out in both the global North and South, a division that has some explanatory value when assessing the availability of resources for digitisation and the effects of colonialism yet should not obscure significant internal variations. While (mass) heritage digitisation is most advanced in Western Europe, in terms of scale, even there not everything is, or will ever be, digitized. What is digitised, however, shapes the stories we can tell about the past. This is of course similar to the general question of what heritage is preserved and how that affects historical research and engagements with the past in general, yet 'digital' enhances and amplifies these impacts in various ways, which will be discussed in the second part of the paper.

In order to perform a more structured analysis of the *process* of digitisation and its political dimensions, I will expand upon a scheme proposed by the sociologist Richard Harvey Brown and the librarian Beth Davis-Brown in their seminal 1998 article 'The Making of Memory'. In their analysis, the Browns explored four ideological and political functions of archival and curatorial work "as these are understood by professional librarians and archivists in the United States" and argued how and why these also constituted "deployments of power" (Brown and Brown 1998). These functions are easily transposed to the digital realm:

Tab. 1

Political dimensions of archival and curatorial work (Brown and Brown 1998)	Digital equivalents
Collections are allocated to different depositories, libraries, or archives in the name of efficiency in avoiding redundancy = allocation of control.	Which institutions digitise and control digital collections? What infrastructures and data frameworks are used?
Collection development refers to decisions concerning what is and what is not collected, what is merely stored but not catalogued (and hence made intellectually accessible), and what is thrown.	What is digitised and why? What is metadated?
Cataloging and classification refer to the organizational and intellectual description of what is held. Whose schema will be used?	How is it classified and how is it metadated?
Circulation and access refer to decisions about who gets to see what, and this is shaped in part by the classification system or categorical order.	How is access provided and mediated?

The paper will conclude by highlighting the paradoxical situation we currently face with regard to digitisation and the state of 'memory' in both the global North and South. It might be increasingly common to describe non-digitised heritage as 'hidden', but that label suggests digitisation as a miracle cure which can solve the issue. The real problem, however, is that much of our cultural

heritage can not even be discovered digitally through institutional collection databases. In Africa, this problem has even more dire consequences as Chamelot, Hiribarren, and Rodet recently pointed out: "There is now a greater risk that archives which have not been previously classified and inventoried will be lost because the slow work of digitization projects monopolizes the time of many archivists" Chamelot, Hiribarren and Rodet 2020). More attention should therefore be paid to (online) cataloguing before digitisation, in the case of materials where neither is done, as well as to linking online archives to catalogues/ descriptive information about offline resources. Cataloguing is a fundamental precondition for enabling access to heritage and without the ability to even find out about important archival holdings online, the question of whether they are digitised or not becomes moot. This is especially true in historical research where knowledge about the materials that exist and could be part of one's evidentiary basis is a key aspect in framing research designs, and where justifying the choice of materials that are to be used in a given research project, whether these can be found online or offline, an essential step before the actual research even begins.

Bibliography

- Brown, Richard Harvey, and Beth Davis-Brown** (1998): "The Making of Memory: The Politics of Archives, Libraries and Museums in the Construction of National Consciousness", in: *History of the Human Sciences* 11/4: 17-32. <https://doi.org/10.1177/095269519801100402>.
- Chamelot, Fabienne, Vincent Hiribarren, and Marie Rodet** (2020): "Archives, the Digital Turn, and Governance in Africa", in: *History in Africa* 47/1: 101-118. <https://muse.jhu.edu/article/761254>.
- Ebdon, Richard and Sara Gould** (1999): *IFLA/UNESCO Survey on Digitisation and Preservation*. IFLA Programme on Preservation and IFLA Programme for UAP.
- Golubev, Alexey** (2021): "Digitizing Archives in Russia: Epistemic Sovereignty and Its Challenges in the Digital Age" in: Gritsenko, Daria, Mariëlle Wijermars and Mikhail Kopotev (eds.): *The Palgrave Handbook of Digital Russia Studies*. Cham: Springer 353-369.
- McCauley, Denis** (2016): *A new age of culture. The digitisation of arts and heritage*. Economist Intelligence Unit.
- Risam, Roopika** (2019): *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston: Northwestern University Press.
- Rosenzweig, Roy** (2003): "Scarcity or Abundance? Preserving the Past in a Digital Era", in: *The American Historical Review* 108/3: 735-762, 739.
- Tanner, Simon, and Kahn, Rebecca** (2014): "Building Futures: The Role of Digital Collections in Shaping National Identity in Africa" in T. Barringer, M. Wallace, and J. Damen (eds.): *African Studies in the Digital Age: DisConnects?* Leiden: Brill 111-127.
- Yale, Elizabeth** (2015): "The History of Archives: The State of the Discipline", in: *Book History* 18/1: 332-59. .
- Zaagsma, Gerben** (2013): "On Digital History", in: *BMGN - Low Countries Historical Review* 128/4: 3-29.
- European Commission** (2011): *COMMISSION RECOMMENDATION of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation*.
- Time Machine - A Common History for the Continent** (2019). See: <https://www.timemachine.eu/trailer/>.

Digicol - Directory of Digitised Library Collections (2002-2006). See: <https://web.archive.org/web/20060612205929/http://www.unesco.org/webworld/digicol/>.

eNumerate (2017). See: [https://en.wikipedia.org/wiki/Enumerate_\(project\)](https://en.wikipedia.org/wiki/Enumerate_(project)). The former eNumerate project website now redirects to EGMUS (European Group on Museum Statistics) which focuses only on museums.

Verwendung von Wissensgraphen zur inhaltlichen Ergänzung kleinerer Textkorpora

Hagen, Thora

thora.hagen@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg

Problemstellung

Die Verwendung von statischen oder dynamischen Word Embeddings wie FastText (Bojanowski et al. 2016) oder BERT (Devlin et al. 2019) hat die Verarbeitung natürlicher Sprache auch im Bereich der digitalen Geisteswissenschaften wesentlich verbessert. Allerdings setzen diese Verfahren voraus, dass man zu ihrem Training über ein sehr großes Textkorpus verfügt, zum Beispiel Wikipedia oder das OSCAR Korpus (Suárez et al. 2020), mit mehreren Gigabyte Umfang. Viele DH-Projekte können aber nur auf sehr viel kleinere Textmengen zurückgreifen. Andererseits beschäftigen sich viele DH-Projekte mit kultureller Überlieferung, die schon seit längerer Zeit erforscht wird, so dass etwa Wörterbücher oder andere strukturierte Nachschlagewerke vorliegen. Dieses Paper diskutiert, wie man ein Word Embedding wie FastText auf sehr kleinen Textmengen trainieren kann und durch die Hinzufügung von Wörterbüchern als Wissensgraphen eine deutlich verbesserte abstrakte semantische Repräsentation des Korpus erreichen kann.

Wissensgraphen oder Knowledge Graphen sind eine Form der Informationsrepräsentation, bei der systematisch Aussagen in der Form Subjekt-Prädikat-Objekt (Tripel) dargestellt werden. Die Informationen im Graph können einer spezifischen Domäne angehören oder auch Allgemeinwissen insgesamt abbilden. Das automatische Umwandeln einer lexikalisch-semantischen Ressource (LSR) – das können beispielsweise Wörterbücher oder Enzyklopädien sein – in einen Wissensgraphen ist nicht zuletzt durch die eher offen gehaltene Definition eines Wissensgraphen unproblematisch, da die Einträge in LSR häufig bereits in einer Art Tripel-Struktur organisiert sind. Die Erstellung des Graphen aus einer LSR reicht über einfache regelbasierte Verfahren (Chodorow 1985) über Clustering Methoden (Oliveira und Gomes 2011), bis hin zu der Verwendung von Sprachmodellen, wobei hier hauptsächlich der Anspruch besteht, Tripel aus Fließtext zu extrahieren (siehe z.B. Yang et al. 2020).

Anreicherung von Word Embeddings durch Wissensgraphen

Die Forschung im Bereich der natürlichen Sprachverarbeitung und Knowledge Graphen hat gezeigt, dass Word Embeddings sowie auch Sprachmodelle von strukturiertem Wissen profitieren können. Der Ansatz ist getrieben von der Intuition, dass einige semantische Relationen in Form von Fließtext selten ausgedrückt werden, da sie für Menschen offensichtlich sind, z.B. "er aß die gelbe Banane" oder "Friedrich Schiller war eine Person" (anstatt die Berufsbezeichnungen zu nennen). Lediglich durch distributionelle Semantik würden sich solche Beziehungen nicht unbedingt in darauf basierenden Modellen niederschlagen.

Um statische Embeddings mit Informationen aus einem Wissensgraphen anzureichern, gibt es hauptsächlich zwei Ansätze:

1. nachträgliches Angleichen der Embeddings an den Wissensgraph („Retrofitting“; Faruqui et al. 2015) oder
2. Konkatenation von Knowledge Graph Embeddings und Word Embeddings sowie anschließende Dimensionsreduktion („Fusion“; Thoma et al. 2017).

Ebenfalls möglich ist das parallele Trainieren der Embeddings auf Fließtext und Wissensgraph (Xu et al. 2014), welches sich allerdings vor allem gegenüber des erstgenannten Ansatzes aufgrund der höheren benötigten Rechenleistung nicht durchgesetzt hat.

Beim Retrofitting werden bereits vortrainierte Word Embeddings im Nachhinein durch einen Wissensgraphen angepasst. Dabei werden die unmittelbaren Nachbarschaften im Graphen ausgenutzt: iterativ werden die Wortvektoren so angepasst, dass die Distanzen zu den direkten Nachbarn im Graphen und gleichzeitig die Distanz zum jeweiligen ursprünglichen Wortvektor minimiert werden.

Die Fusion-Methode verfolgt einen anderen Ansatz: zuerst werden Embeddings auf Basis des Wissensgraphen berechnet. Ähnlich wie bei Word Embeddings auch wird dabei jeder Entität im Graph ein Vektor zugeordnet, welcher durch die Nähen zu den anderen Vektoren aus dem Graph die Bedeutung der Entität abbildet. Populäre Ansätze sind zum Beispiel TransE (Bordes et al. 2013) oder RotatE (Sun et al. 2019). Bei TransE werden die Vektoren so trainiert, dass die Summe aus Subjekt- und Prädikatvektor möglichst nah an dem Vektor des Objekts liegt. Viele andere Verfahren bauen auf der Idee auf, so auch RotatE – hier wird die Beziehung zwischen Objekt und Subjekt durch eine Rotation im Vektorraum über das Prädikat (anstelle der Summe) abgebildet. Für die Fusion werden dann ebenfalls vortrainierte Word Embeddings mit den Embeddings der Entitäten aus dem Graph konkateniert. In einem zweiten Schritt werden dann die konkatenierten Embeddings auf die gewünschte Dimension reduziert, zum Beispiel durch eine Principle Component Analysis (PCA). Dabei können die verschiedenen Embeddings je nach Anwendungsfall unterschiedlich gewichtet werden. Für Knowledge Base Completion zum Beispiel, also die automatische Vorhersage neuer Relationen in einem Graph, eignen sich Fusionsembeddings, bei welchen der Wissensgraph stärker gewichtet wurde, besser (Thoma et al. 2017).

Das hier dargestellte Konzept kann ebenso für das Anreichern von Sprachmodellen wie BERT verwendet werden, denn gerade Sprachmodelle benötigen so wie das Trainieren von Word Embeddings auch viele Textdaten, um eine Sprache angemessen abbilden zu können. Auch hierbei gibt es verschiedene Möglichkeiten; darunter beispielsweise das Einhängen der Tripel-Informationen

in den Fließtext einhergehend mit dem Anpassen des Attention Mechanismus für das Pre-training (Liu et al. 2019) oder das Erstellen eines gänzlich neuen Textes mittels zufälliger Pfade aus dem Graphen, welcher via eigener Adapter in das Pre-training des Sprachmodells integriert wird (Lauscher et al. 2020).

Methodik

Im Folgenden sollen exemplarisch Ergebnisse für das Anpassen von statischen Word Embeddings mithilfe eines Wissensgraphen auf Basis einer kleinen Textmenge dargestellt werden. Um eine kleine Domäne zu simulieren wurde aus dem Deutschen OSCAR Korpus eine Menge an zufälligen Sätzen so ausgewählt, dass etwa 20MB (ca. 3.6M Tokens, 47.000 Types) an Text daraus entstanden sind. Mit diesen Daten wurde dann ein 300-dimensionales Word Embedding Modell mit FastText trainiert. Für den Wissensgraph wurde GermaNet (Hamp und Feldweg 1997, Henrich und Hinrichs 2010) herangezogen. Ähnlich zu dem englischen WordNet werden in GermaNet semantische Beziehungen zwischen Wörtern verzeichnet (Synonyme, Hyponyme etc.). Es ist deshalb zu erwarten, dass die angepassten Word Embeddings vor allem Wortähnlichkeiten besser abbilden können. Obwohl sowohl statische als auch kontextualisierte Embeddings für das Experiment verwendet werden können, wurden hier die statischen Embeddings gewählt, da diese für das Abbilden von semantischen Beziehungen immer noch genauso gut geeignet sind (Ehrmanntraut et al. 2021).

Speziell für diese Evaluation wurden deshalb mehrere Datensätze ausgewählt, welche Wortähnlichkeiten und Wortverwandtschaften prüfen: Schm280 (Köper et al. 2015), SimLex-999 (Leviant und Reichart 2015), ZG222 (Zesch und Gurevych 2006) und Gur65 sowie Gur350 (Gurevych 2005). Bei allen Datensätzen besteht jede Testinstanz aus einem Wortpaar und einer manuell annotierten Wertung der Wortähnlichkeit. Da nicht immer alle Wörter einer Testinstanz in den Word Embeddings gefunden werden, werden nicht alle Instanzen bei der Evaluation berücksichtigt. Die Anzahl der tatsächlich verwendeten Testinstanzen je Testset können in Tabelle 1 eingesehen werden.

Tab. 1: Performanz der angepassten Word Embeddings auf den ausgewählten Datensätzen (Spearman Korrelationen zwischen den Kosinus Ähnlichkeiten der Wortvektorpaaire und der menschlichen Bewertungen). Für das FastText Modell sind zusätzlich die Standardabweichungen von jeweils 15 Durchläufen gegeben.

	SimLex-999	Schm280	ZG222	Gur65	Gur350
# Instanzen	825	242	120	49	237
FastText	0,224 (0,004)	0,495 (0,01)	0,299 (0,01)	0,320 (0,03)	0,653 (0,01)
Retro_all	0,267	0,512	0,291	0,490	0,607
Retro_syms	0,253	0,487	0,273	0,374	0,652
Fusion	0,250	0,484	0,347	0,426	0,666
Retro+Fusion	0,278	0,537	0,337	0,497	0,660

Sowohl Retrofitting als auch der Fusions-Ansatz wurden hier getestet. Faruqui et al. (2013) verwenden beim Retrofitting nur ein Subset der Relationen: nur Synonyme oder Synonyme zusammen mit Hyponymen und Hyperonymen. Für dieses Experiment wurden ebenfalls zwei verschiedene Relationssets ausgewählt: 1) alle Relationen von GermaNet (*Retro_all*) und 2) nur Synonym-Tripel (*Retro_syms*). Für beide Fälle wurden nur jene Tripel auch verwendet, bei welchen Subjekt sowie Objekt in den vortrainierten Word Embeddings enthalten waren. Somit wurden für 1) etwa 80.000 Tripel für das Retrofitting verwendet während bei 2) nur etwa 10.000 verwendet wurden. Um die Tripel zu erstellen wurden nur die Lemmas und nicht die Synset-Struktur aus GermaNet verwendet; also Wörter aus mehreren Synsets werden demselben

Vektor in den Word Embeddings zugeordnet, ohne dass eine Disambiguierung stattfindet. Für die Implementierung wurde eine optimierte Version des Retrofitting Algorithmus von Lengerich et al. (2017) herangezogen.

Für die Fusions-Methode wurden RotatE Embeddings (Implementierung von Zhu et al. (2019)) mit einer Dimension von 128 trainiert. Nachteil dieser Methode ist, dass die Mehrheit der Wörter aus dem Vokabular der Word Embeddings kein Gegenstück in den Entitäten von GermaNet haben (etwa 62%). Fehlende GermaNet Vektoren wurden in diesem Experiment deshalb durch zufällig bestimmte Vektoren innerhalb der Grenzen des GermaNet RotatE Vektorraumes erstellt. Damit ist sichergestellt, dass alle Modelle auf Basis des gleichen Vokabulars beurteilt werden. Die Word und Knowledge Graph Embeddings wurden nicht weiter gewichtet; die Dimensionsreduktion wurde mit einer PCA vorgenommen.

Berechnet wurden die Spearman Korrelationen zwischen den menschlichen Bewertungen und der Kosinus-Ähnlichkeiten der Vektoren der Wortpaare (siehe Ergebnisse in Tabelle 1). Für das vortrainierte FastText sind die Mittelwerte sowie die Standardabweichungen der Spearman Korrelationen aus 15 identisch trainierten Modellen angegeben, um etwaige Schwankungen aufgrund der nicht-deterministischen Modellerstellung anzuzeigen.

Für die Auswertung wurden außerdem *Retro_all* und das Fusions-Modell miteinander kombiniert um ein Ensemble-Modell zu präsentieren und einen Konsens zwischen den Modellen zu bilden. Im Ensemble-Modell werden deshalb die Kosinus Ähnlichkeiten beider Modelle gemittelt und für die Auswertung herangezogen.

Auswertung und Diskussion

Anhand der Ergebnisse zeigt sich, dass auf allen Datensätzen das Anpassen der Word Embeddings mit GermaNet zu einer Verbesserung der Performanz führt. Vor allem das Ensemble-Modell *Retro_all+Fusion* erzielt dabei konsistent bessere Resultate. Insbesondere für die Repräsentation von Wortähnlichkeiten (bzw. Synonymie in SimLex-999) erscheint es lohnenswert, die Anpassung der Word Embeddings durch GermaNet vorzunehmen. Speziell beim Retrofitting fallen die Ergebnisse des auf dem gesamten Relationsbestand von GermaNet optimierten Modells besser aus als nur bei den Synonymen. Bemerkenswert ist trotzdem, dass *Retro_syms*, welches auf einem vergleichsweise kleinem Set aus Tripeln abgestimmt wurde, ebenfalls schon in manchen Fällen Fortschritte erzielen kann.

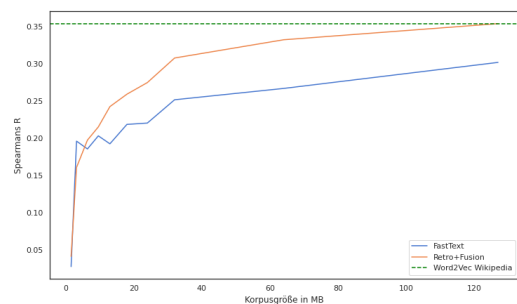


Abb. 1: Performanz des Retro_all+Fusion Modells im Vergleich zu den nicht angepassten FastText Äquivalenten auf dem SimLex-999 Datensatz mit zunehmender Korpusgröße. Als Baseline ist die Performanz des auf Wikipedia trainierten Word2Vec Modells von Leviant und Reichart (2015) gegeben.

Um deutlich zu machen, wie sich die Korpusgröße insgesamt auf die Ergebnisse auswirkt, wurde noch ein weiteres Experiment umgesetzt. Dafür wurden zunächst verschieden große Textsamples aus dem OSCAR Korpus generiert, um dann jeweils ein normales FastText sowie ein durch GermaNet angepasstes Modell auf verschiedenen großen Korpora zu vergleichen. Für das Fitting wurde *Retro_all+ Fusion* gewählt; die Evaluation wurde auf SimLex-999 durchgeführt. Die Ergebnisse sind in Abbildung 1 dargestellt.

Hauptsächlich drei Beobachtungen können aus diesem Experiment abgeleitet werden. Erstens gibt es trotzdem ein unteres Limit für die Korpusgröße, ab der das Fitting der Vektoren zu keiner Verbesserung führt, vermutlich da das FastText Modell allein schon zu wenig Informationen enthält. Hier sind es etwa 10MB für das OSCAR Korpus; allerdings ist es möglich, dass für tatsächlich domänenspezifische Korpora dieses Limit weiter unten angesetzt ist, da das Vokabular von OSCAR sich über alle Domänen erstreckt. Ein kleineres Vokabular, gegeben durch eine spezifische Domäne, würde hier vielleicht auch unterhalb der 10MB ein sinnvolles FastText Modell trainieren können.

Die zweite Beobachtung ist, dass im Falle eines 20MB großen Korpus durch das Fitting eine Performanz eines etwa dreimal so großen Korpus erzielt wird: bei 64MB zeigt das normale FastText Modell eine Performanz von 0,27. Drittens lässt sich anhand der angezeigten Baseline zeigen, dass ein Korpus der Größe 120MB mithilfe des Fittings bereits genauso erfolgreich ist wie ein auf Wikipedia (aktuell etwa 13GB) trainiertes Word2Vec Modell.

Für das hier durchgeführte Experiment wurden keine Hyperparameter optimiert, sowohl für das Trainieren der FastText Embeddings als auch für das Anpassen mit beiden Ansätzen. Durch weiteres Anpassen der Lernrate beim Retrofitting oder bei der Wahl des Knowledge Graph Embedding Algorithmus und der Anzahl der Dimensionen für diese können möglicherweise noch bessere Ergebnisse erzielt werden. Auch die Auswahl der Tripel aus einem Graphen oder die Auswahl des Graphen an sich kann eine entscheidende Rolle spielen; prinzipiell kann diese je nach Anwendungsfall für die Word Embeddings unterschiedlich ausfallen. Wenn mit den Embeddings beispielsweise das Erkennen von Entitäten eher im Fokus steht, ist es denkbar, Tripel aus DBpedia (Lehmann et al. 2015) oder Wikidata zum Verbessern der Vektoren zu verwenden, da dort hauptsächlich Personen und Orte verzeichnet sind. Geht es eher um die Erkennung von Part-of-Speech, so kann die Zuhilfenahme eines Wörterbuches, welches morphologische Informationen zu den Wörtern beinhaltet, nützlicher sein.

Dieses Paper zeigt insgesamt, dass es sich lohnt, eine zur Verfügung stehende lexikalisch-semantische Ressource in den Erstellungsprozess von Word Embeddings zu integrieren; hier demonstriert anhand der Erkennung semantischer Wortähnlichkeiten in der deutschen Sprache. Vor allem dann, wenn wenig Daten vorhanden sind, um ein Forschungsvorhaben in einer speziellen Domäne durchzuführen, können diese zusätzliche Ressourcen ausgenutzt werden um ein Korpus inhaltlich anzureichern und somit das Trainieren eines Word Embeddings Modells unterstützen. Typische Domänen können zum Beispiel ein historisches Korpus, Dialekte, Pidgins und andere Arten von Sprachvariation oder auch ein ganz spezifisches Genre sein. Vor allem also für Germanisten, die auf Grundlage einer eher textarmen Domäne mit quantitativen Methoden arbeiten möchten (sei es beispielsweise für das Erkennen von Bedeutungsveränderungen von Wörtern mithilfe von Embeddings innerhalb einer solchen Domäne), kann das Anreichern von Textkorpora mit Wissensgraphen und eines der hier vorgestellten Verfahren von Interesse sein.

Bibliographie

- Bordes, Antoine / Usunier, Nicolas / Garcia-Durán, Alberto** (2013): „Translating Embeddings for Modeling Multi-relational Data“, in: *Advances in neural information processing systems* 2787-2795.
- Bojanowski, Piotr / Grave, Edouard / Joulin, Armand / Mikolov, Tomas** (2016): „Enriching Word Vectors with Subword Information“, ArXiv:1607.04606.
- Chodorow, Martin S. / Byrd, Roy J. / Heidorn, George E.** (1985): „Extracting Semantic Hierarchies from a Large On-Line Dictionary“, in: *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics* 299–304.
- Devlin, Jacob / Chang, Ming-Wei / Lee, Kenton / Toutanova, Kristina** (2019): „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, ArXiv:1810.04805.
- Ehrmanntraut, Anton / Hagen, Thora / Jannidis, Fotis / Konle, Leonard** (2021): „Type- and Token-based Word Embeddings in the Digital Humanities“, in: *CEUR Workshop Proceedings* 2989.
- Faruqui, Manaal / Dodge, Jesse / Jauhar, Sujay K. / Dyer, Chris / Hovy, Eduard / Smith, Noah A.** (2015): „Retrofitting word vectors to semantic lexicons“, in: *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* 1606–1615.
- Gurevych, Iryna** (2005): „Using the structure of a conceptual network in computing semantic relatedness“, in: *International conference on natural language processing* 767–778.
- Hamp, Birgit / Feldweg, Helmut** (1997): „GermaNet - a Lexical-Semantic Net for German“, in: *Automatic information extraction and building of lexical semantic resources for NLP applications*.
- Henrich, Verena / Hinrichs, Erhard W.** (2010): „GernE-diT-The GermaNet Editing Tool“, in: *ACL (System Demonstrations)* 19–24.
- Köper, Maximilian / Scheible, Christian / Schulte im Walde, Sabine** (2015): „Multilingual Reliability and ‚Semantic‘ Structure of Continuous Word Spaces“, in: *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015) -- Short Papers*.
- Lauscher, Anne / Majewska, Olga / Ribeiro, Leonardo F. R. / Gurevych, Iryna / Rozanov, Nikolai / Glavaš, Goran** (2020): „Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers“, ArXiv:2005.11787.
- Lehmann, Jens / Isele, Robert / Jakob, Max / Jentzsch, Anja / Kontokostas, Dimitris / Mendes, Pablo N. / Hellmann, Sebastian / Morsey, Mohamed / Van Kleef, Patrick / Auer, Sören / u. a.** (2015): „DBpedia--a large-scale, multilingual knowledge base extracted from wikipedia“, in: *Semantic Web* 6 167–195.
- Lengerich, Benjamin J. / Maas, Andrew L. / Potts, Christopher** (2017): „Retrofitting distributional embeddings to knowledge graphs with functional relations“, ArXiv:1708.00112.
- Leviant, Ira / Reichart, Roi** (2015): „Separated by an Uncommon Language: Towards Judgment Language Informed Vector Space Modeling“, ArXiv:1508.00106.
- Liu, Weijie / Zhou, Peng / Zhao, Zhe / Wang, Zhiruo / Ju, Qi / Deng, Haotang / Wang, Ping** (2020): „K-BERT: Enabling language representation with knowledge graph“, in: *Proceedings of the AAAI Conference on Artificial Intelligence* 2901-2908.

Oliveira, Hugo G. / Gomes, Paulo (2011): „Automatic discovery of fuzzy synsets from dictionary definitions“, in: *Twenty-Second International Joint Conference on Artificial Intelligence*.

Ortiz Suárez, Pedro J. / Romary, Laurent / Sagot, Benoît (2020): „A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages“, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sun, Zhiqing / Deng, Zhi-Hong / Nie, Jian-Yun / Tang, Jian (2019): „RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space“, ArXiv:1902.10197.

Thoma, Steffen / Rettinger, Achim / Both, Fabian (2017): „Towards holistic concept representations: Embedding relational knowledge, visual attributes, and distributional word semantics“, in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10587 LNCS: 694–710.

Xu, Chang / Bai, Yalong / Bian, Jiang / Gao, Bin / Wang, Gang / Liu, Xiaoguang / Liu, Tie Yan (2014): „RC-NET: A general framework for incorporating knowledge into word representations“, in: *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management* 1219–1228.

Yang, SungMin / Yoo, SoYeop / Jeong, OkRan (2020): „DeNERT-KG: Named Entity and Relation Extraction Model Using DQN, Knowledge Graph, and BERT“, in: *Applied Sciences* 10.

Zesch, Torsten / Gurevych, Iryna (2006): „Automatically creating datasets for measures of semantic relatedness“, in: *Proceedings of the workshop on linguistic distances* 16–24.

Zhu, Zhaocheng / Xu, Shizhen / Qu, Meng / Tang, Jian (2019): „GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding“, in: *The World Wide Web Conference* 2494–2504.

Vom gedruckten Gazetteer zum digitalen Ortsverzeichnis Das Geschichtliche Ortsverzeichnis (GOV)

Purschwitz, Anne

anne.purschwitz@geschichte.uni-halle.de
Martin-Luther Universität Halle-Wittenberg, Germany

Zedlitz, Jesper

jesper@zedlitz.de
Ministerium für Digitalisierung Schleswig-Holstein

Ortsnamen und geographische Räume wie auch ihre Positionierung spielen in der Geschichtswissenschaft eine wichtige Rolle. Vor diesem Hintergrund ist es häufig erforderlich, in mühevoller Detailarbeit Ortsnamen zu recherchieren, eine Lokalisierung und zunehmend auch eine Visualisierung auf unterschiedlichem Kartenmaterial vorzunehmen. Mittlerweile stehen dafür eine Vielzahl von Tools zur Verfügung, die sowohl die Recherche als auch die

Visualisierung (historischer) Ortsdaten erleichtern sollen. Gerade in dem breiten verfügbaren Spektrum liegen Möglichkeiten, aber auch Schwierigkeiten für Forscher*innen und interessierte Laien.

Die Besonderheiten historischer Raumzuordnungen bringen spezifische Herausforderungen bei der Aufbereitung, Präsentation und Recherche mit sich. Im Verlauf der Zeit änderten sich Schreibweisen, Namen wurden überformt oder neu vergeben, die Zugehörigkeit zu Sprachräumen konnte variieren und sich auf die Namensgebung von Wohnplätzen auswirken. Die Mehrheit der aktuell als Open-Access verfügbaren Ortsverzeichnisse berücksichtigt dieses ‚historische Werden‘ nur unzureichend bis gar nicht.

Das Geschichtliche Ortsverzeichnis (GOV) des ‘Vereins für Computergenealogie’ stellt sich den Herausforderungen in der Modellierung von Zeit und Raum (Zedlitz / Luttenberger 2014a: 220), die für geschichtswissenschaftliche Fragestellungen von entscheidender Wichtigkeit sind. Im GOV wurden dafür topologische Beziehungen modelliert, die im Vergleich zu time slices belastbarer sind (Zedlitz / Luttenberger 2014a: 220–222). Entstanden ist ein hierarchisches Modell, das zwischen Siedlung und Administration trennt, beide Ebenen aber vielfältig miteinander in Beziehung setzen kann und Mehrfachverknüpfungen ermöglicht. So kann eine Siedlung zeitlich z.B. zu unterschiedlichen politischen und kirchlichen Administrationen gehören. Das GOV geht damit deutlich über bisherige Standards hinaus (Zedlitz / Luttenberger 2014b: 36–38).

Ziel des Beitrags ist es, die Datenmodellierung im GOV vorzustellen und die Entwicklung zukünftiger Nutzungsperspektiven für die akademische und bürgerwissenschaftliche Forschung zu diskutieren. Wir beabsichtigen, das im Rahmen einer bürgerwissenschaftlichen Initiative entstandene Verzeichnis mit seinen umfassenden Funktionen und Normierungen, die weit über andere Ansätze (GND, GeoNames) hinausgehen, zu einem Hilfsmittel für Historiker*innen weiterzuentwickeln, das einen einheitlichen Zugriff auf eine Vielzahl von ortsbezogenen Daten erlaubt. Hierfür sind wir auf die Auseinandersetzung mit Historiker*innen aus verschiedenen Forschungsfeldern angewiesen. Sie können zum einen Testdatensätze zur Verfügung stellen (wie das etwa die Deutsche Auswandererbriefsammlung unter Leitung von Ursula Lehmkuhl getan hat: <http://www.auswandererbriefe.de/>), und andererseits wollen wir ihre Wünsche, Vorstellungen und Ansprüche in einem iterativen Prozess in das GOV aufnehmen. Die Analyse von Möglichkeiten zur Integration und Nutzung im geisteswissenschaftlichen Forschungskontext und die Darstellung bzw. Abgrenzung gegen andere online verfügbare Gazetteers steht im Vordergrund des Beitrags. Aufgrund der Vielzahl von Schnittpunkten erscheint es uns als besonders zielführend, die Bedarfe der Forschungsgemeinschaft zu eruieren – gleichzeitig aber auch die Stärken und Schwächen eines Crowdsourcing-Ansatzes kritisch zu diskutieren.

Raum und Zeit

‘Raum’ ist eine elementare Kategorie der Geisteswissenschaft und sollte aus diesem Grund bei der Verarbeitung von umfangreichen Datenbeständen in seinen strukturierenden und Interpretationen erleichternden Funktionen ernst genommen werden. Dafür erforderlich sind Tools, die einen Schwerpunkt auf die Koppelung von Zeit und Raum legen, denn Geschichte agiert in einem geographischen Nebeneinander und einem chronologischen Nacheinander (von Brand 1958: 22) – somit benötigt Geschichtswissenschaft eine Raum-Zeit-Kompetenz. Die Notwendigkeit einer solchen fasst der Geographie-Didaktiker Walter Sperling in fol-

genden Punkten zusammen: (Sperling 1982: 81) 1. geschichtliche Prozesse spielen sich in Räumen ab, 2. jeder Raum ist geschichtlich geworden und 3. Räume wurden in verschiedenen Zeiten unterschiedlich bewertet.

Gerade die Verbindung von Zeit und Raum stellt für die Anwendung in der Geschichtswissenschaft eine zentrale Anforderung an Ortsverzeichnisse dar.

„Orte haben eine historisch-politische Dimension, die bei einer übergreifenden Registererfassung erst sichtbar zu einem Problem wird. [...] Für die Visualisierung von Briefen etwa sind historische Karten ein Desiderat; generell auch Geodaten für Flächen. Und alle mit Geodaten versehenen Einträge müssen mit einem Zeitspielraum kombiniert sein, denn beispielsweise die Altstadt von Jerusalem ist eben heute nicht am selben Ort wie vor 2.000 Jahren.“ (Kamzelak 2018).

Gazetteers und GIS

In den vergangenen Jahren haben eine Vielzahl unterschiedlicher Disziplinen Ortsverzeichnisse (Gazetteers) erstellt, jeweils für sie relevante Informationen zusammengetragen und im Semantic Web publiziert. Die große Zahl der Verzeichnisse spiegelt dabei zum einen den offensichtlichen Bedarf an strukturierten geographischen Daten wider, zeigt andererseits aber auch die variierenden Herangehensweisen, Datenerhebungen, Standards und Interessen. Ebenso werden immer mehr Tools und Plattformen für die computergestützte Raumanalysen („spatial analysis“; Baur et. al., 2014) entwickelt, die sich auch für historisierende Fragestellungen einsetzen lassen.

An dieser Stelle muss zunächst klar zwischen geographischen Informationssystemen (GIS) einerseits und Ortsverzeichnissen, sogenannten Gazetteers, andererseits differenziert werden.

Geographische Informations-Systeme (GIS) ermöglichen die Visualisierung, Analyse und Archivierung raumbezogener Daten. Das einem GIS zugrundeliegende Datenmodell ist in der Lage, Geobjekte sowohl in Form von Vektor- als auch von Rasterdaten zu verwalten. Ein Schwerpunkt von GIS-Anwendungen liegt in den unterschiedlichen Analysetools, die es z.B. erlauben, Distanzberechnungen oder Sichtfeldanalysen vorzunehmen.

Ortsverzeichnisse verfügen im Unterschied dazu nur sehr bedingt über geographische (Visualisierungs-)Funktionen und sind nicht in der Lage, raumbezogene Prädiktoren zu berechnen. Andererseits blicken Gazetteers auf eine lange Entwicklungsgeschichte zurück. Erste Gazetteers sind bereits aus der Antike bekannt. Mit der zunehmenden Vermessung der Welt wurden sie ein immer wichtigerer Bestandteil zunächst u.a. von Atlanten. Sie finden sich seit dem 19. Jahrhundert aber zunehmend auch als eigenständige Publikationen (z.B. Rudolph 1870-1872). Ein Teil dieser gedruckten mehrbändigen Verzeichnisse hat in den letzten Jahrzehnten seinen Weg in das Semantic Web gefunden und nutzt die Vielzahl der dadurch eröffneten digitalen Funktionsweisen (z.B. *Meyers Gazetteer*).

Betont werden muss nochmals, dass Ortsverzeichnisse nicht primär auf Karten und Visualisierungen ausgerichtet sind (auch wenn sie Geokoordinaten zur Identifizierung von Orten enthalten können), sondern auf die eindeutige Identifikation von Orten, Siedlungen, Wohnplätzen etc. durch ihre Einordnung in administrative Zusammenhänge. Ein Schwerpunkt digitaler Gazetteers besteht dennoch in der Zusammenführung von Raumdaten (Koordinaten), Sachdaten (Ergänzung der Raumdaten, z.B. Kreis, Kirche, Staat) und Metadaten (Beschreibungen der Daten). Die dafür erforderlichen Informationen können punktuell für bestimmte Zeitabschnitte erhoben und aufbereitet werden. Für Historiker jedoch

zielführender ist die Kopplung unterschiedlichster Sachdaten zu einem Ort unter Berücksichtigung von Zeiträumen. Für die Erstellung historischer Ortsverzeichnisse ist es somit unerlässlich, Zeitangaben und Zeiträume zu verzeichnen, ohne damit ‚neue‘ Orte zu schaffen, sondern vielmehr mit dem Ziel, die historische Genese von Siedlungen nachvollziehbar zu machen. Aus diesem Grund gibt es eine Reihe von Projekten, die Informationen dieser Art im Semantic Web in Form von Linked Open Data (LOD) zur Verfügung stellen. Die beiden Technologien/Sprachen, die typischerweise für diesen Zweck verwendet werden, sind das Resource Description Framework (RDF; Resource Description Framework 2004) und die Web Ontology Language (OWL; OWL2 2012). Da die Darstellung sowohl von räumlichen als auch von zeitlichen Informationen mit RDF oder OWL nicht trivial ist (Hobbs / Pan 2004: 66-85; Gutierrez 2007: 207-218; Motik 2012: 3-21), gibt es sehr unterschiedliche Ansätze, wie man Verwaltungsstrukturen im Semantic Web darstellen kann.

Funktionsweise des GOV

Der bürgerwissenschaftliche Verein CompGen („Verein für Computergenealogie“) hat in den vergangenen Jahrzehnten das größte historische Ortsverzeichnis entwickelt, das mit Bezug auf Deutschland und viele andere Länder Ortsnamen (in variierenden Schreibweisen), kirchliche und staatliche Zugehörigkeiten im Zeitverlauf sowie geographische Koordinaten erfasst. Das 1992 von Heinz Augustin initiierte Projekt wurde im Jahr 1995 an den „Verein für Computergenealogie“ übergeben und dort unter Einbeziehung einer Vielzahl von Familienforschern weiterentwickelt, mit Daten gefüllt und zu einem komplexen und hochkompetenten bürgerwissenschaftlichen Projekt ausgebaut (Schnadt 2011). Im Kern besteht es heute also aus einer Datenbank mit Informationen über historische Verwaltungsstrukturen und deren Beziehungen zu Siedlungen bzw. funktionalen Gebäuden (Kirchen etc.). Eine web-basierte Anwendung steht bereits seit dem Jahr 2000 zur Verfügung. Aktuell beträgt die Abdeckung auf Ebene der Siedlungsplätze (also unterhalb der Gemeinden) für das späte Kaiserreich etwa 80%. Insgesamt verzeichnet das GOV rund 1,25 Millionen Einträge (Stand Juni 2021) - die beständig ergänzt werden. Erschließungsschwerpunkte sind bisher Europa, die USA sowie Australien.

Ziel des Geschichtlichen Ortsverzeichnisses ist es, Siedlungen, Wohnplätze, Verwaltungseinheiten etc. eindeutig zu identifizieren. Da dafür allein der Name eines Ortes nicht ausreichend ist, erfolgt die Einordnung in administrative Zusammenhänge, wobei an dieser Stelle streng zwischen politischen, kirchlichen und juristischen Zugehörigkeiten unterschieden wird, wie auch Veränderungen und Verschiebungen in der Zugehörigkeit abgebildet werden.

Bei der Entwicklung des GOV entstand die Überzeugung, dass eindeutige Identifikatoren sowohl für physische als auch für administrative Objekte benötigt werden, die in ihren Kombinationen im zeitlichen Verlauf unterschiedliche Relationen eingehen können. Im Bereich des Semantic Web sind solche Identifikatoren in Form von Uniform Resource Identifier (URI) üblich. Das GOV bietet für Wohnplätze und Verwaltungsobjekte genau solche URIs zur eindeutigen Identifizierung an. Der zentrale Mehrwert des GOV besteht darin, dass es zwischen Organisationsstruktur und ‚gebauten‘ Einheiten klar unterscheidet, durch die Vielzahl von Beiträgern deutlich über bisher vorhandene historische Ortsverzeichnisse hinausgeht, keine zeitliche oder räumliche Beschränkung aufweist, zeitlich und räumlich referenzierte Daten zur Verfügung stellt, jederzeit ergänzt, verbessert und aktualisiert werden kann und weltweit zugänglich ist.

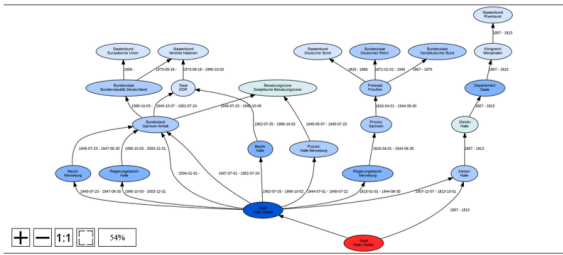


Abb. 1: Visualisierung der Zugehörigkeiten und historischen Entwicklung der Stadt Halle/Saale.

Dem Verein für Computergenealogie geht es nicht um das Sammeln möglichst vieler historischer Informationen, sondern um deren Optimierung, Anbindung und um Qualitätssicherung. Um die Datenqualität des GOV zu überprüfen und sicherzustellen, wird aktuell zusammen mit dem Verein ‚CorrelAid‘ ein Projekt durchgeführt, in dem wir mit Hilfe von Data Scientists systematisch erarbeiten, wie sich die Datenqualität des GOV aktuell gestaltet. Hieraus werden gegebenenfalls neue Ansätze zu Modellierung und Datenaufnahme abgeleitet.

GOV-Datenmodell

Im GOV werden sowohl Wohnplätze als auch Verwaltungsobjekte als GovObject modelliert. Ein GovObject besitzt eine Reihe von Eigenschaften (PropertyForObject) und Beziehungen zu anderen Objekten (Relation). Sowohl Eigenschaften als auch Beziehungen können mit Zeitangaben (von-bis) in verschiedenen Genauigkeiten und Quellenangaben versehen werden. Bei Namen (PropertyName) kann zusätzlich die Sprache in Form eines ISO-639-2 Codes angegeben werden. So ist es möglich, nicht nur verschiedene Namen zu verschiedenen Zeiten, sondern auch in verschiedenen Sprachen anzugeben. Eine weitere besondere Eigenschaft eines Objekts ist der Typ (PropertyType - aktuell verzeichnet das GOV für Siedlungen und Administrationen 276 Typen). Dieser gibt Auskunft darüber, um welche Art von Wohnplatz (z. B. Dorf, Weiler, Häusergruppe) oder Verwaltungsobjekt (z. B. Gemeinde, Stadt, Kreis, Bundesland) es sich handelt. Abhängig vom Typ kann ein Objekt eine geographische Position besitzen, um so die Anzeige auf einer Karte zu ermöglichen. Ein Vergleich mit anderen Systemen, die historische Verwaltungsinformationen im Semantic Web bereitstellen, hat gezeigt, dass das Datenmodell des GOV anderen Ansätzen deutlich überlegen ist, insbesondere durch die Möglichkeit von Zeit- sowie Quellenangaben (Zedlitz / Kluttig 2014: 290; Zedlitz / Luttenberger 2014a: 223-230 und Zedlitz / Luttenberger 2014b: 33-36).

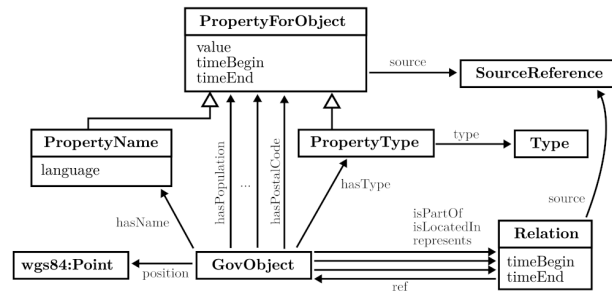


Abb. 2: UML-Modell der grundlegenden Strukturen des Geschichtlichen Ortsverzeichnisses.

Zu den Orten selbst sind folgende Informationen im GOV erfasst: geographische Lokalisation (Koordinaten und Position auf Karte) - die Angabe der Geokoordinaten im GOV orientiert sich dabei am World Geodetic System 84 (WGS84; Department of Defense 1991) oder dem Europäischen Terrestrischen Referenzsystem 1989 (ETRS89) - Eigenschaften (z.B. Bevölkerungszahlen und -entwicklung, Postleitzahl), fremdsprachige oder frühere Namen und die Zugehörigkeit zu politischen, kirchlichen und/oder rechtlichen Administrationen. Zudem wurden die Orte auch in anderen Datenbanken referenziert, so finden sich Verknüpfungen mit GeoNames oder dem *Amtlichen Ortschaften-/Ortsverzeichnis Bayern*. Eine geographische Position weisen nur Wohnplätze, also die unterste Ebene von Objekten im GOV auf. Da Verwaltungsobjekte typischerweise eine größere Fläche umfassen, wäre die Angabe einer einzelnen Punktkoordinate nicht zielführend. Gleichzeitig kann die ungefähre Position und Ausdehnung eines Verwaltungsobjekts aufgrund der zugehörigen Wohnplätze - die ja eine geographische Position besitzen - berechnet werden.

Das GOV ist somit kein GIS - obwohl es mit bestehenden geographischen Datenbanken und GIS-Systemen verknüpft werden kann.

Im Rahmen von GOV können Daten (auf der GOV-Website) kartographisch dargestellt werden, wobei unterschiedliche andere Dienste genutzt und eingebunden werden, z.B. Google Earth, historische Messtischblätter, das Virtuelle Kartenforum der SLUB und einige mehr; alternativ kann auch aus QGIS heraus ein allerdings noch rudimentärer Web Feature Service des GOV aufgerufen werden (vgl. <https://gov-dev.genealogy.net/wfs>). Alle kartographischen Darstellungen im GOV beruhen auf Punktkoordinaten; ein Ausbau zu einem polygonorientierten System ist zur Zeit nicht beabsichtigt, da eine Erfassung des Verlaufs von vielen tausend Gemeinde- und Pfarreigrenzen und ihrer Änderungen nicht realistisch erscheint.

Im GOV selbst sind andere Normdatensysteme verlinkt. Hierfür ist es allerdings wichtig, dass die Modellierungslogik des GOV und die dieser anderen Systeme aufeinander bezogen werden können. In Hinblick auf die Verknüpfungen zu anderen Normdaten bzw. Datenbeständen im Semantic Web weist GOV 770.000 Verbindungen zu GeoNames auf. Schwieriger gestaltet sich die Anbindung an die GND, da dort nicht klar zwischen Verwaltungsobjekt (z.B. Gemeinde) und Siedlungsobjekt (z.B. Dorf) unterschieden wird. Dabei handelt es sich um einen grundlegenden konzeptuellen Unterschied zum GOV. In der GND werden Orte als Geographika und als Körperschaften abgebildet, deren Geokoordinaten GeoNames entstammen. Berücksichtigt werden können bei der Erfassung von Geographika innerhalb der GND: Name, Quelle, die (administrative) Zugehörigkeit ohne zeitliche Dimension und Namensänderungen (teilweise wird dadurch jedoch eine neue Entität geschaffen; gerade dieser Modellierung

setzt das GOV eine belastbare Alternative entgegen). Aus diesem Grund existieren aktuell nur 9.200 Verbindungen mit der GND.

Nutzungsperspektiven

Im Vortrag möchten wir das Anwendungspotential in den Geisteswissenschaften und die Funktionsweise des GOV deutlich machen.

Für die Suche in GOV steht eine erweiterte und eine einfache Suchoption zur Verfügung, deren Hauptunterschied in der Eingrenzung auf bestimmte Objekttypen bei der erweiterten Suche besteht. Doch auch die Ergebnisse der einfachen Suche können durch Filterfunktionen in einem zweiten Schritt bspw. auf einzelne Objekttypen, Bundesländer etc. beschränkt werden. Eine weitere Funktion der erweiterten Suche in GOV ermöglicht die Verknüpfung von zwei Objekten. Hier kann ein übergeordnetes Objekt (z.B. Bundesland oder Landkreis) mit einem Ortsnamen verknüpft werden, was die Genauigkeit der Treffer erhöht und die nachträgliche Selektion der Treffermenge vereinfacht. Nach der Identifikation des Ortes können Geodaten und GOV-ID übernommen werden.

Für die Nutzung der *Toponymresolution* (Sen 2016) und damit der automatischen Abfrage in GOV benötigt es eine csv-Datei mit UTF-8 kodierten Zeichen unter Nutzung des Feldtrenner Tabulator. Toponyme selbst können Wildcards („?“ , „*“) und explizite Begriffe von Unschärfe („“) enthalten, ebenfalls nicht relevant ist die Beachtung der Groß- und Kleinschreibung. Wie in der Einzelabfrage besteht bereits im Vorfeld der Suchanfrage die Option sich auf bestimmte Typklassen (z.B. Wohnplatz, Gericht, Kirche) oder beliebige Kombinationen von unterschiedlichen Typenklassen zu konzentrieren, wodurch die Abfrage individuell an die Nutzeranforderungen angepasst werden kann. Ebenfalls möglich ist die Nutzung von Sprachbeschränkungen und die Eingrenzung der Suche auf bestimmte Regionen oder individuell definierte Zeiträume. Die Einbeziehung zusätzlicher Informationen zu Ortsnamen und die bereits vor der Abfrage bestehende Möglichkeit, Filterfunktionen zu nutzen, erhöht die Trefferqualität der Toponymresolution im GOV. Gleichzeitig werden mögliche Fehler in der Datenaufnahme offensichtlich, denn stimmen Ortsangabe und Bundesland nicht überein, erfolgt keine Zuweisung in anderen Bundesländern. Diese fehlenden Identifikationen können dann gezielt an die Quellen zurückgegeben und einer neuerlichen Kontrolle unterzogen werden.

Das GOV bietet bei der Einzelabfrage eine Visualisierung der möglichen Treffer auf einer OpenStreetMap-Karte. In der Detailansicht eines selektierten Ortes kann der Nutzer dann zwischen unterschiedlichem Kartenmaterial wählen und dieses auch exportieren. Zur Verfügung steht Google Earth als kml-Datei (Download), die Visualisierung in GoogleMaps, wikimapia und MapQuest (Browser), ebenfalls erfolgen kann eine Verknüpfung mit dem *Virtuellen Kartenforum 2.0*, betrieben von der SLUB Dresden. Registrierten Nutzern bietet sich zudem die Möglichkeit der Nutzung und auch Bearbeitung von historischen Messtischblättern aus der Zeit um 1900. Grundlegend muss einer Visualisierung (gleich in welcher Form) zunächst immer eine Identifizierung und Lokalisierung vorausgehen. In diesem Fall bleibt die Frage bestehen, inwiefern historische Veränderungen sich auch in den Koordinaten bzw. Vektordaten widerspiegeln sollen und müssen – an dieser Stelle hoffen wir eine Diskussion anstoßen zu können, wie die Veränderungen des Raumes in der Zeit nicht nur mit einem ‚Zeitstempel‘ versehen, sondern vielmehr in ihrer Genese transparent visualisiert werden können.

Bibliographie

Baur, Nina / Hering, Linda / Raschke, Anna Laura / Thierbach, Cornelia (2014): “Theory and Methods in Spatial Analysis. Towards Integrating Qualitative, Quantitative and Cartographic Approaches in the Social Sciences and Humanities”, in: *Historical Social Research / Historische Sozialforschung* Vol. 39, No. 2 (148), S. 7-50.

Brand, Ahasver von (1958): *Werkzeug des Historikers. Eine Einführung in die Historischen Hilfswissenschaften*, Stuttgart: Kohlhammer.

Department of Defense (1991): *World Geodetic System 1984. Its definition and relationships with local geodetic systems*, Rockville, MD.

Gutierrez, C. / Hurtado, C. A. / Vaisman, A. (2007): “Introducing time into rdf”, in: *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 2, S. 207–218.

Hobbs, J. R. / Pan, F. (2004): “An ontology of time for the semantic web”, in: *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 1, S. 66–85.

Kamzelak, Roland (2018): “Von der Raupe zum Schmetterling oder Wie fliegen lernen – Editionsphilologie zwischen Infrastruktur und Semantic Web”, in: ders. / Steyer, Timo (eds.): *Digitale Metamorphose: Digital Humanities und Editionswissenschaft* (= Sonderband der Zeitschrift für digitale Geisteswissenschaften). DOI http://dx.doi.org/10.17175/sb002_004. [Letzter Zugriff 22.04.2021].

Motik, B. (2012): “Representing and querying validity time in rdf and owl: A logic-based approach”, in: *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 12, S. 3–21.

OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax. <http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/> [Letzter Zugriff 01.07.2021].

Resource Description Framework (RDF): Concepts and Abstract Syntax. <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/> [Letzter Zugriff 30.06.2021].

Rudolph, Heinrich (1870-1872): *Vollständigstes geographisch-topographisch-statistisches Orts-Lexikon von Deutschland, so wie der unter Oesterreichs und Preussens Botmässigkeit stehenden nichtdeutschen Länder: enthaltend: alle Städte, Flecken, Pfarr- Kirch- und andere Dörfer, Ort- und Bauerschaften, Kirchspiele, Schlösser, Rittergüter, Vorwerke, Weiler, Hüttenwerke, Mühlen, Höfe, merkwürdige Ruinen, Krüge, Einschnitten, Einöden u.s.w.*, Leipzig: Zander.

Snadt, Jörg (2011): “Das genealogische Ortsverzeichnis (GOV)”, in: *Vermessung Brandenburg*, 2011/1, Seite 35-42.

Sen, Dennis (2016): *Toponymresolution on Historical Serial Sources*, Master’s Thesis, Kiel [unveröffentlicht].

Sperling, Walter (1982): “Die Stellung der Historischen Geographie in einem modernen geographischen Curriculum”, in: *Erdkunde* 26, S. 79-84.

Zedlitz, Jesper / Luttenberger, Norbert (2014a): “A Survey on Modelling Historical Administrative Information on the Semantic Web”, in: *International Journal on Advances in Internet Technology*, Col. 7 No. 3&4, S. 218-231.

Zedlitz, Jesper / Luttenberger, Norbert (2014b): “Modelling (Historical) Administrative Information on the Semantic Web”, in: *WEB 2014, The Second International Conference on Building and Exploring Web Based Environments*, S. 33-39.

Zedlitz, Jesper / Kluttig, Thekla (2014): “Das Genealogische Ortsverzeichnis (GOV), Eine Einführung”, in: *Archivar*, 67. Jahrgang, Heft 03 Juli, S. 289-292.

Von der Wolke zum Pfad

Visuelle und assoziative Exploration zweier kultureller Sammlungen

Brüggemann, Viktoria

viktoria.brueggemann@fh-potsdam.de
UCLAB, Fachhochschule Potsdam, Germany

Bludau, Mark-Jan

mark-jan.bludau@fh-potsdam.de
UCLAB, Fachhochschule Potsdam, Germany

Pietsch, Christopher

cpietsch@gmail.com
UCLAB, Fachhochschule Potsdam, Germany

Dörk, Marian

marian.doerk@fh-potsdam.de
UCLAB, Fachhochschule Potsdam, Germany

Hintergrund

In den letzten Jahren hat sich ein Forschungsfeld im Bereich von Visualisierungen kultureller Sammlungen etabliert, welche die kulturhistorischen Artefakte und Facetten von Sammlungen in Form visueller Interfaces sichtbar und erfahrbar machen (Windhager et al. 2018). Ein Großteil dieser Arbeiten widmet sich einzelnen Sammlungen, die jeweils einer spezifischen Systematik folgen und eine geringe Vielfalt an Objektgattungen und Attributen aufweisen (z.B. Glinka et al. 2017, Gortana et al. 2018). Das hier vorgestellte Forschungsprojekt widmete sich der Frage, wie mehrere Sammlungen visuell in Bezug gesetzt und auf zugängliche Weise exploriert werden können. In enger Zusammenarbeit mit Sammlungsexpert*innen sollten konkrete Ansätze zur visuellen und assoziativen Exploration von zwei unterschiedlichen Sammlungen der Staatlichen Museen zu Berlin (SMB) entwickelt werden. Entstanden ist ein funktionaler Prototyp¹², der auf konzeptionellen Ambitionen wie denen des digitalen Flanierens (Dörk et al., 2011) und der glücklichen Entdeckungen (Thudt et al., 2012) in großzügigen (Whitelaw, 2015) und explorativen Interfaces (Kreisel et al., 2017) aufbaut und die tradierten Anordnungen objektbezogener und suchbasierter Museumswebseiten aufbricht. Diese Formen der gezielten Informationssuche erfordern ein konkretes Interesse der Suchenden und werden den heterogenen und umfangreichen Beständen musealer Sammlungen nur bedingt gerecht, wenn es um ein assoziatives, Interessen-getriebenes Durchstöbern von Sammlungen und Beständen geht.

Aus der Vielzahl der Bestände und Sammlungen der SMB wurden auf Grundlage verschiedener Aspekte, wie beispielsweise dem Umfang der erfassten Objekte, der Erschließungstiefe der Objektdaten und nicht zuletzt der thematischen Heterogenität, die Bestände der Alten Nationalgalerie sowie des Museums Europäischer Kulturen aus dem 19. Jahrhundert ausgewählt. Während die

Sammlung der Alten Nationalgalerie eine der umfangreichsten Epochenansammlungen für die Kunst des 19. und frühen 20. Jahrhunderts ist, findet sich im Museum Europäischer Kulturen eine der größten Sammlungen zur Alltagskultur und Populärkunst in Europa. Die Kontraste und Gegensätze, aber auch die Gemeinsamkeiten beider Bestände sollten im Rahmen des Projekts erforscht und erfahrbar gemacht werden.

Prozess und Vorgehen

Der Forschungs- und Gestaltungsprozess folgte einem iterativen Vorgehen, in dem sich Workshops, Feedbackgespräche und Prototyping im Modus des Co-Designs abwechseln und gegenseitig beeinflussen (Dörk et al., 2020; Chen et al., 2014). So wurden Sammlungsexpert*innen und weitere Mitarbeiter*innen der Museen und des assoziierten Forschungsprojekts museum4punkt0 in den Prozess mit eingebunden. Darüber hinaus wurden fachfremde Personen beteiligt, um neben der versierten Perspektive auch die Interessen, Bedürfnisse und Anforderungen anderer Nutzer*innengruppen verstehen und berücksichtigen zu können. Visualisierung fungiert hierbei als interdisziplinäre Forschungsmethode per se, welche neue Erkenntnisse bereitstellt, aber ebenso fächerübergreifende Diskussionen anregt (Hinrichs et al., 2019) und neuartige Perspektiven auf museale Objekte und Daten eröffnet.

Zum Anfang des Projekts wurde ein Co-Design-Workshop mit den genannten Personengruppen durchgeführt. In Kleingruppen wurden Collagen (siehe Abb. 1) erstellt, auf denen bereitgestellte Bildmaterialien aus den beiden Sammlungen arrangiert und annotiert wurden. Ziel war es, sich auf ästhetischer und abstrakter Ebene mit den beiden Sammlungen zu beschäftigen und jenseits technologischer Beschränkungen Ideen für Visualisierungen zu entwickeln (Chen et al., 2014). In der anschließenden Gruppendiskussion interpretierten zunächst jene Teilnehmer*innen die Collagen, welche die jeweils zu betrachtende Collage nicht erstellt hatten, gefolgt von einer Erläuterung der Ersteller*innen. Dieser Austausch führte in eine fokussierte Diskussion über die Konkretisierung der Ansprüche und Ziele des Projekts und die praktische Umsetzbarkeit insbesondere in Hinblick auf Verfügbarkeit von Daten und die Diversität von Zielgruppen.

Ein Großteil der entstandenen Collagen wies eine assoziative Durchmischung der beiden Sammlungen auf, wobei Schlagworte und Kategorien die Arrangements erklärten. Es wurde mehrfach der Wunsch geäußert, die Verschiedenheit der Sammlungen zu achten, ohne dass dies in einem Interface eine Trennung der Objekte impliziere. Exploration wurde mit zirkulären Streifzügen durch die Sammlungen assoziiert, die entlang visueller Assoziationsketten und erklärenden Beschriftungen angeregt werden sollten. Ein anderer Teil der Diskussion drehte sich um automatische Verfahren der Bildanalyse; hier wurden Gegensätze wie Ähnlichkeit/Differenz, Kuratierung/Algorithmus und Narration/Exploration erörtert, die für den folgenden Designprozess ein fruchtbares Spannungsfeld eröffneten.

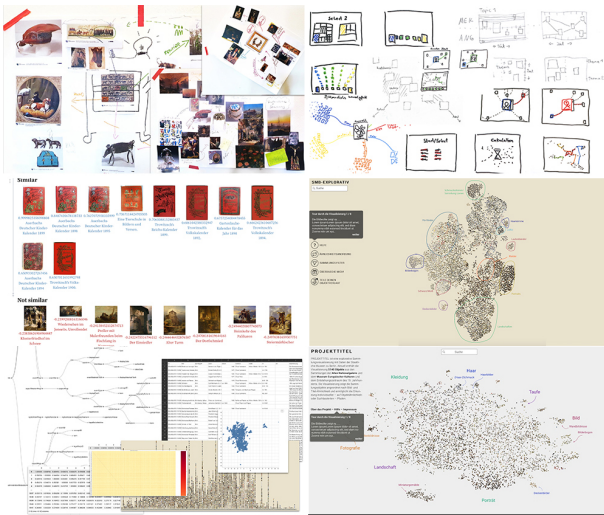


Abb. 1: Iterativer Design- und Forschungsprozess mit Collagen, Skizzen, Notebooks und Prototypen.

Ähnlichkeitsbasierte Beziehungen wurden für den weiteren Designprozess als Schwerpunkt gewählt. Eine tiefergehende Auseinandersetzung mit den Sammlungsdaten offenbarte eine hohe Heterogenität in der Datenqualität, z.B. in der Datierung und Beschreibung der Objekte. Da die manuelle Anreicherung und Angleichung der Metadaten über den Rahmen des Projekts hinausgingen, wurden Möglichkeiten der algorithmischen Ähnlichkeitsanalyse mittels maschinellen Lernens eruiert. Für die Berechnung von Ähnlichkeiten wurden verschiedene Kombinationen aus Titel- und Bilddaten herangezogen und die resultierenden Layouts auf ihre Plausibilität hin untersucht. Im Designprozess stellte sich eine kombinierte Ähnlichkeitsanalyse, also eine Mischung aus Titel- und Bilddaten in gleicher Gewichtung, für die Platzierung in einem Interface als zielführend heraus, da hier die bestmögliche Durchmischung der Sammlungen erzielt werden konnte.

Zum Ende der explorativen Phase wurde im Austausch mit den beteiligten Kuratorinnen der Sammlungen mögliche Interfacekonzepte und die kombinierte Ähnlichkeitsanalyse besprochen, wobei die Beteiligten zu verständlichen Erklärungen der Anordnung rieten, um den Einstieg in die Visualisierung zu erleichtern. Der Versuch einer automatischen Verschlagwortung mithilfe eines vortrainierten Modells schlug fehl, da hier unspezifische und unpassende Formulierungen (z.B. „Kleptomane“ als Stichwort zu einem Porträt) generiert wurden oder beispielsweise bei Kunstwerken der Fokus auf den visuell dominanteren Bilderrahmen anstatt auf das eigentliche Kunstwerk gelegt wurde. Daraufhin wurden Schlagworte für prominente Objekthäufungen manuell formuliert und im Austausch mit den Kuratorinnen im Interface platziert. Zusätzlich zu regelmäßigen Feedback-Runden mit den Kooperationspartner*innen im Laufe des Gestaltungs- und Entwicklungsprozesses wurde in der finalen Projektphase eine Evaluation nach der Think-Aloud Methode (Carpendale 2008) durchgeführt. Hier wurden Nutzer*innen mit unterschiedlichem Vorwissen zu den Sammlungen oder zur Nutzung von Datenvisualisierungen gebeten beim Explorieren der Visualisierung laut auszusprechen, was sie denken, sehen und interpretieren. Die Erkenntnisse aus der Studie sind abschließend iterativ in das Projektergebnis eingeflossen.

Visuelle Exploration: Von Wolken zu Pfaden

Das Ergebnis unseres iterativen und kollaborativen Designprozesses ist eine Visualisierung, die fließende Wechsel zwischen einer **Wolken-Ansicht** – eine ähnlichkeitsbasierte Übersicht aller Objekte (siehe Abb. 2, links) – und **Pfad-Ansichten** – von einer Objekt-Auswahl ausgehende nach Ähnlichkeit geordnete Ketten (siehe Abb. 2, rechts) – ermöglicht. Für die Ähnlichkeitsberechnungen wurden dabei zum einen die visuellen Bilddaten (also visuelle Ähnlichkeiten), als auch Titel (textliche Ähnlichkeit der Objektitel) herangezogen. Dadurch sollen nicht nur inhaltlich-thematische Verbindungen sichtbar werden, sondern auch Serendipität (Thudt et al. 2012), also glückliche Zufälle im Finden interessanter Objekte, gefördert werden. Insgesamt werden die heterogenen Objekte unterschiedlicher Sammlungen dabei ohne das Hervorheben von Sammlungszugehörigkeit basierend auf Ähnlichkeitsberechnungen in Beziehung gesetzt, um so die Neugier an den Objekten zu wecken und die Exploration anzuregen. Durch diese implizite Ähnlichkeitsdarstellung werden einzelne Objekte unterschiedlicher Sammlungen in Beziehung gesetzt, ohne vorher eine Normalisierung der Metadaten (wie z.B. Schlagworte) vorzunehmen.

Technisch wurde für die Berechnung der Bild- und Titelähnlichkeiten basierend auf maschinellern Lernen eine Merkmalsextraktion über die kombinierte Nutzung von *TensorFlow* (Abadi et al. 2016), *Universal Sentence Encoder Multilingual* (Yang et al. 2019) und *Big Transfer* (Kolesnikov et al. 2019) in *Python Notebooks* vorgenommen. Die webbasierten Visualisierungen wurden prototypisch in *Observable Notebooks* und final mit dem JavaScript Framework *Svelte*, der Datenvisualisierungs-Library *D3.js* sowie der WebGL Rendering Library *PixiJS* entwickelt.

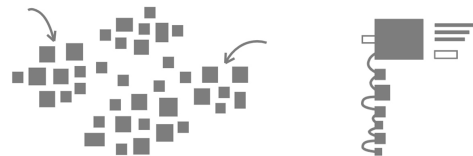


Abb. 2: Das Interface teilt sich in zwei verbundene Modi: Die Wolken-Ansicht (links) gibt eine Übersicht und die Pfad-Ansicht (rechts) bietet Details eines einzelnen Objekts und jenen, die ihm ähnlich sind.

Wolken-Ansicht

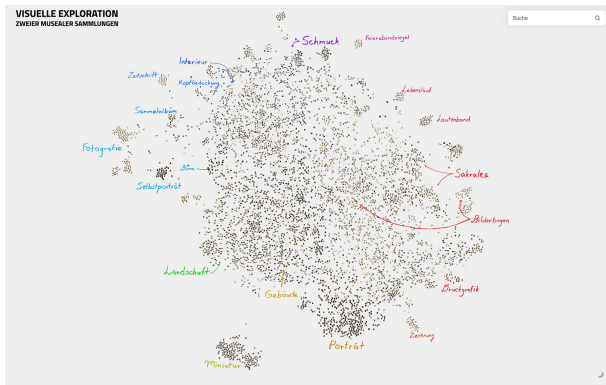


Abb. 3: Tausende von Thumbnails aus zwei musealen Beständen werden auf Basis ihrer Titel- und Bildähnlichkeit arrangiert.

Die Wolken-Ansicht bietet eine Übersicht und den Ausgangspunkt zur Exploration der beiden Sammlungen (siehe Abb. 3). Hierbei wurden die Objekte basierend auf den berechneten Bild- und Titelähnlichkeiten über *UMAP* (McInnes, Healy, and Melville 2018) – eine Technik zur Dimensionalitätsreduktion – auf einer zweidimensionalen Fläche nach Ähnlichkeit verteilt; das heißt, Objekte die sich visuell und auf den Titel bezogen besonders ähnlich sind, liegen in der Visualisierung nahe beieinander und es bilden sich einzelne Cluster besonders ähnlicher Objekte. Die manuell erzeugten Schlagworte bieten eine erste Orientierung und über Klick einen Einstieg in bestimmte Regionen des Arrangements. Die Handschriftlichkeit unterstreicht den Gegensatz zwischen der algorithmischen Anordnung und der kuratorischen Annotation der Cluster. Handschriftliche bzw. händische Annotationen in Projektionen multidimensionaler Skalierungen sind eine bereits angewandte Methode, um algorithmische Dimensionalitätsreduktionen nachvollziehbarer zu machen (z.B. Stefaner 2018, Vane 2018).

Mit der Wolken-Ansicht lässt sich über etablierte Pan+Zoom-Gesten, die Auswahl von Schlagworten oder einzelnen Objekten sowie über die Such-Funktion interagieren. Die Eingabe in der Suche (noch vor der Bestätigung der Suchanfrage) hebt die Suchergebnisse in der Ansicht hervor. So lassen sich auch jene Objekte, die der Alten Nationalgalerie oder dem Museum Europäischer Kulturen entstammen identifizieren.

Beim Klick auf ein Element oder ein Schlagwort oder über die Zoom-Funktion wird in die Visualisierung hineingezoomt und Details der einzelnen Artefakte werden sichtbar (siehe Abb. 4). Wurde ein Element ausgewählt, so wird der Titel des Objektes angezeigt und besonders ähnliche Objekte werden hervorgehoben. Ebenso werden bei der Eingabe im Suchfeld relevante Objekte in der Wolke hervorgehoben. Ein weiterer Klick auf den Button „im Pfad anzeigen“ bei der Objektauswahl oder die Betätigung der Suchanfrage löst eine Übergangsanimation aus, welche alle zur Auswahl ähnlichen bzw. alle für eine Suchanfrage relevante Objekte hervorhebt und zur Pfad-Ansicht animiert.

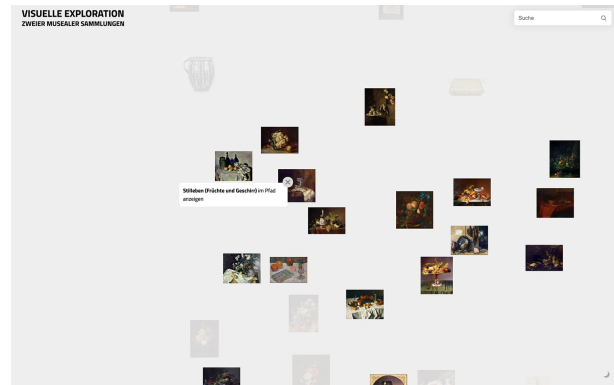


Abb. 4: In der Pfad-Ansicht werden die Details eines ausgewählten Objekts, gefolgt von ähnlichen Objekten entlang eines Fadens in abnehmender Ähnlichkeit, angezeigt.

Pfad-Ansicht

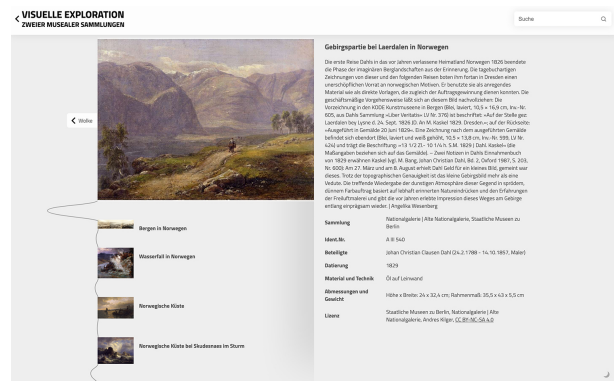


Abb. 5: In der Pfad-Ansicht werden die Details eines ausgewählten Objekts, gefolgt von ähnlichen Objekten entlang eines Fadens in abnehmender Ähnlichkeit, angezeigt.

Im Gegensatz zur Wolken-Ansicht, die Sammlungsobjekte global nach Ähnlichkeit anordnet, werden in der Pfad-Ansicht ausgehend von einem ausgewählten Objekt bzw. Suchbegriff Objekte absteigend der Ähnlichkeit bzw. Relevanz nach in einer Liste an einem Faden aufgereiht (siehe Abb. 5). Dabei folgt die Pfad-Ansicht dem Prinzip von monadischen Visualisierungen (Dörk et al. 2014), indem sie von einem Objekt ausgehende, individuelle Perspektiven auf andere Objekte der Sammlung ermöglicht. Der Ausschlag des Fadens zeigt Änderungen im Grad der Ähnlichkeit zum ausgewählten Objekt an: Ein größerer Ausschlag bedeutet dabei eine größere Differenz im Vergleich zum vorangegangenen Objekt. Zusätzlich wird der berechnete Ähnlichkeitswert in den Metadaten, ausgehend vom ersten Objekt in der Reihenfolge, angezeigt.

Das Ziel dieser Ansicht ist es, assoziative Ketten zu ermöglichen. So werden zum Beispiel Werke, die im Titel das Wort „Blume“ enthalten, mit Stilleben von Blumen, Blumenvasen oder - auf Grundlage der Bildähnlichkeit - mit floralen Schmuckstücken zusammengebracht (siehe Abb. 6). Ähnlichkeits-Pfade, die durch die Nutzung der Suchleiste generiert werden, basieren dagegen auf einer Volltextsuche über die Metadaten-Felder, sodass zum Beispiel auch Pfad-Ansichten basierend auf Materialien oder Künstler*innen angezeigt werden können.

Die Auswahl eines Objekts in einem Pfad öffnet dessen Metadaten und blendet den Button „Zeige ähnliche Objekte als Pfad“ ein, der bei Selektion dazu führt, dass der Ähnlichkeits-Pfad sich ausgehend von dem nun ausgewählten Objekt neu anordnet. Ein weiterer Button ermöglicht zudem die Bewegung zurück zur Wolke mit Fokus und Zoom auf das dabei ausgewählte Objekt.

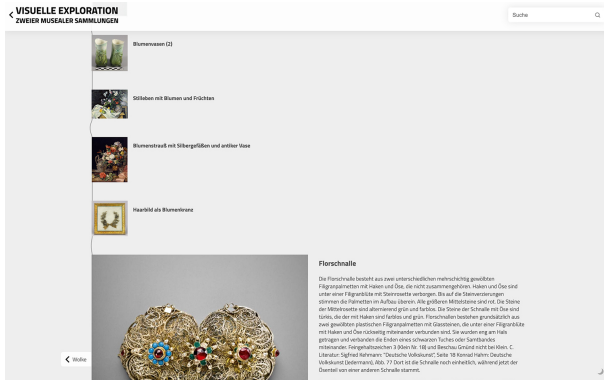


Abb. 6: Der Ähnlichkeits-Pfad ausgehend von einem Blumenstilleben enthält unter anderem auch Blumenvasen oder ein florales Schmuckstück.

Diskussion und Reflexion

Das Visualisierungskonzept wurde sowohl in Zusammenarbeit mit den Kooperationspartner*innen und Museumsmitarbeiter*innen als auch in einem strukturierten Evaluationsprozess mit externen Teilnehmer*innen reflektiert und angepasst. Dabei traten einige Herausforderungen und Fragestellungen auf, welche einerseits die algorithmischen Methoden betrafen, andererseits grundsätzliche Fragestellungen zur Exploration digitaler Sammlungen eröffneten. Zunächst wurde in der Evaluation deutlich, dass ein assoziatives Bewegen durch die Sammlungen – oftmals entlang der Schlagworte – sowohl nach einem selbst gewählten als auch nach einem vorgegebenen Interesse gut funktionierte. Tester*innen hoben dabei hervor, dass sie Entdeckungen gemacht hätten, mit denen sie nicht gerechnet hatten und dass die Visualisierung sich besonders gut zum spielerischen Erkunden eignete. Hingegen konnte die Ähnlichkeits-Anordnung in beiden Ansichten der Visualisierung von den meisten Teilnehmenden kaum vollständig erschlossen werden. Dennoch wirkten die Teilnehmenden grundsätzlich interessiert an der Neuartigkeit der Darstellung und den dadurch aufgeworfenen Fragen und versuchten, die Anordnung mittels visueller Vergleiche der Objekte und Beschreibungstexte zu verstehen. Es lässt sich jedoch auch festhalten, dass von Teilnehmenden oft auf die Suche zurückgegriffen sowie der Wunsch nach einer geführten „Tour“ neben der freien Exploration geäußert wurde.

Hinsichtlich einer Weiterentwicklung des Projektes lässt sich zunächst der Aspekt des maschinellen Lernens nennen, welcher insbesondere in einer Verfeinerung der Ähnlichkeits-Anordnung bestehen könnte, sodass mehr sichtbare Cluster und gegebenenfalls automatisch erzeugte Schlagworte als Ausgangspunkt für eine Exploration zur Verfügung stehen könnten. Ein erstes Experiment zur automatisierten Verschlagwortung zeigte jedoch, dass zumindest eine Überprüfung durch Sammlungsexpert*innen unentbehrlich bleiben wird, da beim maschinellen Lernen kontext- und sammlungsspezifische Informationen bislang nicht herangezogen werden können. Für eine potentielle Weiterentwicklung des

Projekts stellte sich darüber hinaus die Frage, ob die algorithmische Anordnung für mehrere und diverse Sammlungen skalierbar wäre; dies bezieht sich sowohl auf die Frage, ob eine automatisierte Clusterbildung mit weiteren Beständen überhaupt erfolgreich wäre, als auch ob die größere Anzahl an Objekten die Performance des Prototypen signifikant mindern würde.

Das Projekt hat wiederholt gezeigt, dass eine interdisziplinäre Zusammenarbeit und kritische Betrachtung von algorithmischen Methoden, insbesondere unter Einbezug von Sammlungsexpert*innen, unerlässlich ist. Während nach einer ersten qualitativen Auswertung festgestellt werden kann, dass der Visualisierungsprototyp, der auf einem Wechsel zwischen globalen und lokalen Ähnlichkeiten beruht, insbesondere im Bezug auf freie Exploration und unerwartete Entdeckungen sehr positives Feedback hervorgerufen hat, so wirft die Verwendung von Visualisierungstechniken auf Basis automatischer Ähnlichkeitsanalysen wichtige Fragen zur Vermittlung algorithmischer Arrangements auf.

Fußnoten

1. <https://visualisierung.smb.museum>
2. <https://uclab.fh-potsdam.de/smb/smb-demo.mp4>

Bibliographie

- Abadi, Martín / Agarwal, Ashish / Barham, Paul / Brevdo, Eugene / Chen, Zhifeng / Citro, Craig / Corrado, Greg S. / Davis, Andy / Dean, Jeffrey / Devin, Matthieu / Ghemawat, Sanjay / Goodfellow, Ian / Harp, Andrew / Irving, Geoffrey / Isard, Michael / Jia, Yangqing / Jozefowicz, Rafal / Kaiser, Lukasz / Kudlur, Manjunath / Levenberg, Josh / Mane, Dan / Monga, Rajat / Moore, Sherry / Murray, Derek / Olah, Chris / Schuster, Mike / Shlens, Jonathon / Steiner, Benoit / Sutskever, Ilya / Talwar, Kunal / Tucker, Paul / Vanhoucke, Vincent / Vasudevan, Vijay / Viegas, Fernanda / Vinyals, Oriol / Warden, Pete / Wattenberg, Martin / Wicke, Martin / Yu, Yuan / Zheng, Xiaoqiang (2016): “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”, in: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*: 265–283.
- Carpendale, Sheelagh (2008): “Evaluating Information Visualizations”, in: Kerren, Andreas / Stasko, John T. / Fekete, Jean-Daniel / North, Chris (eds.): *Information Visualization. Lecture Notes in Computer Science* 4950. Berlin, Heidelberg: Springer 19–45.
- Chen, Ko-le / Dörk, Marian / Dade-Robertson, Martyn (2014): “Exploring the Promises and Potentials of Visual Archive Interfaces”, in: *Proceedings of the 2014 IConference*: 735–741.
- Dörk, Marian / Carpendale, Sheelagh / Williamson, Carey (2011): “The Information Flaneur. A Fresh Look at Information Seeking”, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 1215–1224.
- Dörk, Marian / Comber, Rob / Dade-Robertson, Martyn (2014): “Monadic exploration: seeing the whole through its parts”, in: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*: 1535–1544.
- Dörk, Marian / Müller, Boris / Stange, Jan-Erik / Herseni, Johannes / Dittrich, Katja (2020): “Co-Designing Visualizations for Information Seeking and Knowledge Management”, in: *Open Information Science* 4, 1: 217–235.
- Glinka, Katrin / Pietsch, Christopher / Dörk, Marian (2017): “Past Visions and Reconciling Views: Visualizing Time, Tex-

ture and Themes in Cultural Collections”, in: *Digital Humanities Quarterly* 11, 2.

Gortana, Flavio / Tenspolde, Franziska von / Guhlmann, Daniela / Dörk, Marian (2018): “Off the Grid: Visualizing a Numismatic Collection as Dynamic Piles and Streams”, in: *Open Library of Humanities* 4, 2.

Hinrichs, Uta / Forlini, Stefania / Moynihan, Bridget (2019): “In Defense of Sandcastles: Research Thinking through Visualization in Digital Humanities”, in: *Digital Scholarship in the Humanities* 34: i80–i99.

Kolesnikov, Alexander / Beyer, Lucas / Zhai, Xiaohua / Puigcerver, Joan / Yung, Jessica / Gelly, Sylvain / Houlshby, Neil (2019): *Big Transfer (BiT): General Visual Representation Learning* <https://arxiv.org/pdf/1912.11370> [letzter Zugriff 15.07.2021].

Kreiseler, Sarah / Brüggemann, Viktoria / Dörk, Marian (2017): “Tracing exploratory modes in digital collections of museum Web sites using reverse information architecture”, in: *First Monday* 22, 4.

McInnes, Leland / Healy, John / Melville, James (2018): “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, in: *Journal of Open Source Software* 3.

Shneiderman, B. (1996): “The eyes have it: a task by data type taxonomy for information visualizations”, in: *Proceedings 1996 IEEE Symposium on Visual Languages*: 336–342.

Stefaner, Moritz (2018): *Multiplicity: A Collective Photographic City Portrait* <https://truth-and-beauty.net/projects/multiplicity> [letzter Zugriff 15.07.2021].

Thudt, Alice / Hinrichs, Uta / Carpendale, Sheelagh (2012): “The Bohemian Bookshelf. Supporting Serendipitous Book Discoveries through Information Visualization”, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 1461–1470.

Vane, Olivia (2018): *Visualising the Royal Photographic Society Collection: Part 2* <https://www.vam.ac.uk/blog/digital/visualising-the-royal-photographic-society-collection-part-2> [letzter Zugriff 15.07.2021].

Whitelaw, Mitchell (2015): “Generous Interfaces for Digital Cultural Collections”, in: *Digital Humanities Quarterly* 9, 1.

Windhager, Florian / Federico, Paolo / Schreder, Gunther / Glinka, Katrin / Dörk, Marian / Miksch, Silvia / Mayr, Eva (2018): “Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges”, in: *IEEE transactions on visualization and computer graphics* 25, 6: 2311–2330.

Yang, Yinfei / Cer, Daniel / Ahmad, Amin / Guo, Mandy / Law, Jax / Constant, Noah / Abrego, Gustavo H. / Yuan, Steve / Tar, Chris / Sung, Yun-Hsuan / Strope, Brian / Kurzweil, Ray (2019): *Multilingual Universal Sentence Encoder for Semantic Retrieval* <https://arxiv.org/pdf/1907.04307> [letzter Zugriff 15.07.2021].

Halling, Thorsten

thorsten.halling@hhu.de

Heinrich-Heine-Universität Düsseldorf, Medizinische Fakultät, Institut für Geschichte, Theorie und Ethik der Medizin

Holly, Eva Maria

eva.holly@hhu.de

Heinrich-Heine-Universität Düsseldorf, Abteilung Wirtschaftsgeschichte

Wieloch, Jasmin

wielocj@hhu.de

Heinrich-Heine-Universität Düsseldorf, Medizinische Fakultät, Institut für Geschichte, Theorie und Ethik der Medizin

Schnaitter, Hannes

hannes.schnaitter.1@ibi.hu-berlin.de

Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft

Balck, Sandra

balcksaa@hu-berlin.de

Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft

Plakidis, Melina

melina.plakidis@dfki.de

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Speech and Language Technology Lab

Rehm, Georg

georg.rehm@dfki.de

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Speech and Language Technology Lab

Fangerau, Heiner

heiner.fangerau@hhu.de

Heinrich-Heine-Universität Düsseldorf, Medizinische Fakultät, Institut für Geschichte, Theorie und Ethik der Medizin

Dörk, Marian

marian.doerk@fh-potsdam.de

Fachhochschule Potsdam, UCLAB

Was sehe ich? Visualisierungsstrategien für Datentransparenz in der Historischen Netzwerkanalyse

Bludau, Mark-Jan

mark-jan.bludau@fh-potsdam.de

Fachhochschule Potsdam, UCLAB

Einführung

Angelehnt an die Methodik der Sozialen Netzwerkanalyse nutzen Historiker*innen vermehrt auch graphenbasierte Visualisierungen zur (Re-)konstruktion sozialer Strukturen. Im Zentrum steht das Ziel, Muster zu erkennen um soziale Prozesse historisch verstehen und erklären zu können (Kerschbaumer et al. 2020). Konferenzserien, eine Fachzeitschrift (*Journal of Historical Network Research*), ein Handbuch (Düring et al. 2016), Lehrveranstaltungen sowie Studien zu (sozialen) Netzwerken aus al-

len Epochen der Geschichtsschreibung (Düring, M. & Grandjean, M. 2021) machen die Historische Netzwerkanalyse (HNA) zu einem inzwischen ausdifferenzierten Forschungsfeld. Zu den wichtigsten methodischen Herausforderungen gehört die Suggestionskraft von Visualisierungen: „Sie sind fehlbare, abstrahierte und interessengetriebene Modelle und mit entsprechender Vorsicht zu betrachten.“ (Düring et al. 2016: 6)

Obwohl zur Netzwerk-Visualisierung bereits zahlreiche Forschungsarbeiten und Anwendungen vorliegen, beziehen sich die hier dargestellten Techniken und Taxonomien zumeist auf allgemeine Graphvisualisierung (Hadlak et al. 2015; Lee et al. 2006; Nobre et al. 2019). Dies führt dazu, dass Anwendungen und Visualisierungen häufig nicht ausreichend spezifische Bedürfnisse, Herausforderungen und relevante Datenpraktiken der HNA-Forschung oder generell geisteswissenschaftliche Forschung beachten. Dazu gehören Probleme der Unsicherheit, Mehrdeutigkeit, Subjektivität (Drucker 2011) und wenige Methoden zur Darstellung von Provenienz (Hadlak et al. 2015). Einigen Herausforderungen in der Netzwerkforschung werden dabei unter anderem durch Nutzung von statistischen Methoden begegnet (z.B. Griffith et al. 2016; Scholtes 2017). Dennoch sehen wir auch im Bereich der Visualisierung Notwendigkeit an der Weiterentwicklung geeigneter Strategien. Bisherige Visualisierungsforschung zu Datenprovenienz dagegen konzentrierte sich auf naturwissenschaftliche Workflows (Ragan et al. 2016). Die Abbildung von Provenienz und die Offenlegung sowie die Verknüpfung mit Quellen stellen jedoch einer Metastudie zu Nutzeranforderungen von Forscher*innen der digitalen Geisteswissenschaften zufolge, zentrale Faktoren für die Glaubwürdigkeit und damit für das Vertrauen in eine Visualisierung dar (Lamqaddam et al. 2020). Weiterhin gibt es zunehmend Ansätze um diesen und ähnlichen Herausforderungen mit kritischen und für die Geisteswissenschaften geeigneten Visualisierungskonzepten zu begegnen (z.B. Kleymann & Stange 2021).

Aktuell werden im deutschsprachigen Sprachraum für HNA vor allem proprietäre Visualisierungstools verwendet, die zumeist nicht genuin für HNA-Zwecke konzipiert sind und in der Regel keine interaktiven, webbasierten Darstellungsformen ermöglichen. Eine Ausnahme bildet hier *nodegoat*, eine speziell für die Humanities entwickelte webbasierte Forschungsumgebung (van Bree & Kessels 2017). Angaben zu verwendeten Visualisierungstools fehlen in vielen Studien ebenso wie eine Begründung der Auswahl. Im Text genannte Datenquellen können den Knoten und Kanten oft nicht zugeordnet werden, ferner sind in einer statischen Darstellung facettierte Relationen ebenso wie dynamische Prozesse über die Zeit hinweg kaum abbildbar.

Verschiedene Projekte suchen daher verstärkt nach Lösungsansätzen zur Visualisierung selbst erhobener oder rekonstruierter Netzwerke (z.B. Campbell et al. 2018; Novak et al. 2014). Zu den Kernproblemen gehört die mit heterogenen Datenrepositorien verbundene, nahezu zwangsläufige Verzerrung von Visualisierungsergebnissen, insbesondere durch fehlende, ungenaue oder uneindeutige Daten (Drucker 2011). Die Datentransparenz bezüglich der Datenquellen ist für die HNA sowohl bei selbst erhobenen als auch bei sekundär genutzten Daten Voraussetzung, will sie geisteswissenschaftlichen Kriterien der Nachprüfbarkeit und Nachvollziehbarkeit von Interpretationen erfüllen.

Projekthintergrund und Vorgehen

Grundlage dieses Beitrags ist das anwendungsbezogene Forschungs- und Entwicklungsprojekt SoNAR (IDH) – Interfaces to Data for Social Historical Network Analysis and Rese-

arch¹ (Bludau et al. 2020). In diesem Projekt soll systematisch forschungsorientiert das Aufbereiten, Bereitstellen und Analysieren von Massendaten für den Aufbau einer HNA-Forschungstechnologie erprobt werden. Den Ausgangspunkt bildet die Überlegung, dass für den Bibliotheks- und Archivbereich erstellte Norm- und Metadaten bereits über eine gesicherte Qualität der Daten verfügen, deren Provenienz nachvollziehbar und nachprüfbar ist. Eine Disambiguierung, also die eindeutige Zuordnung von Entitäten wie Personen, Körperschaften oder Orten ermöglicht eine Sekundärnutzung der Daten für automatisierte historische Netzwerkanalysen. Einschränkungen ergeben sich systemimmanent aufgrund des ursprünglichen Verwendungszwecks der Daten zur Beschreibung vor allem von Bibliotheksbeständen. Dieser kann dazu führen, dass Entitäten fehlen oder unzureichend beschrieben sind. Bei der Zusammenführung heterogener Daten (GND², Kalliope³, DNB⁴, ZDB⁵ und SBB⁶) mit großen Datenmengen (~52 Mio. Knoten, ~185 Mio. Kanten) ist die Transparenz der Daten das zentrale Kriterium für die Nutzung einer solchen Forschungstechnologie.

In einem iterativen Prototypingprozess zwischen Forscher*innen mit Hintergrund in Geschichtswissenschaften, Data Science, Datenvisualisierung, Kulturwissenschaften und Informationswissenschaften haben sich Prozesse der Forschungsdesign-Entwicklung, der Datentransformation und -zusammenführung, Entwicklung von Visualisierungen und die Evaluation (siehe Balck et al. 2021) dieser Arbeitsbereiche gegenseitig beeinflusst und unterstützt. In einem Co-Design Workshop mit HNA-Forscher*innen und anderen interdisziplinären Teilnehmenden – in dem anhand von Collagen (siehe Chen et al. 2014) Zugänge zu Graph-Daten zur Ermöglichung von HNA-Forschung gestaltet wurden (siehe Abb. 1) – war Transparenz bezüglich der Daten und Visualisierungen ein wichtiger, immer wiederkehrender Diskussionspunkt. Die Forderung nach Kontextualisierung der Daten zeigte sich in den Collagen auch in der vielfältigen Verbindung und Annotation visueller Elemente.

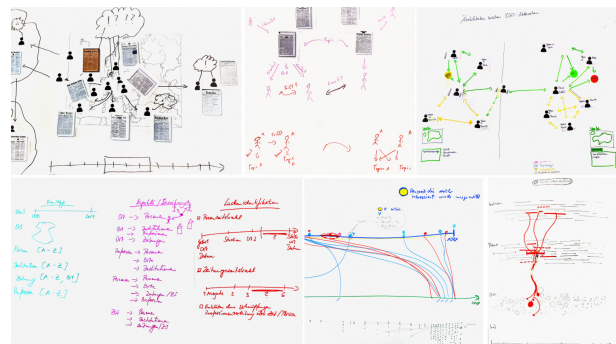


Abb. 1: Teilnehmer*innen unseres Co-Design Workshops wurden gebeten Zugänge zu den SoNAR-Daten in Form von Collagen zu gestalten mit dem Ziel, Diskussionen und Austausch über Potenziale visueller Analyse anzuregen.

Weiterhin wurde der Prozess, aufbauend auf dem Ansatz der *grounded evaluation* (Isenberg et al. 2008) durch Interviews zu Forschungsprozessen und genutzten Tools sowie durch fortlaufende Evaluierung der iterativ entwickelten Projektergebnisse auch unter Einbindung von externen HNA-Forscher*innen im Rahmen einer Nutzer*innenstudie begleitet. In beiden Studien wurden die Transparenz der Modellierungsentscheidungen, die Nachvollziehbarkeit der Datengrundlage und die Unterstützung bei der Dokumentation des eigenen Forschungsprozesses als

grundlegende Anforderungen an ein wissenschaftlich nutzbares Visualisierungswerkzeug herausgearbeitet:

„[...] da würde ich gerne wissen wo diese Informationen herkommen und wie die eigentlich miteinander verknüpft werden und wer auf die Idee gekommen ist das so zu tun.“ (P3)

Förderung von Transparenz in Graph Visualisierungen

Aus der sich in der Projektarbeit ergebenden Forderung nach Datenklarheit haben wir folgende **Designziele (DZ)** für interaktive HNA-Visualisierungen abgeleitet:

DZ1) Aufnahme und Kommunikation von Datenprovenienzen: Um die Datentransparenz auch nach Datentransformation und Zusammenführungen sicherzustellen, müssen Datenprovenienzen über Merkmale bei Knoten und Kanten unbedingt erhalten werden und über URIs auf die Ausgangsdaten verweisen.

DZ2) Dokumentation vorausgegangener Prozesse: Die konkreten Schritte der Datentransformationen und Anwendungen von Algorithmen für Visualisierungen müssen inklusive Code nachvollziehbar und frei verfügbar dokumentiert werden, um Vertrauen zu schaffen und reproduzierbare Ergebnisse sowie kritische Auseinandersetzungen zu ermöglichen. Dies beinhaltet auch eine Versionierung der Daten in allen Zwischenschritten.

DZ3) Offenhaltung der Interpretierbarkeit der Daten: Datenunsicherheiten und unterschiedliche Granularitätsstufen müssen für spätere Interpretation in den Daten erhalten bleiben und dürfen nicht durch Normalisierungen entfernt werden. Auch in Visualisierungen müssen Kodierungen verwendet werden, die fachspezifische Einschätzungen und Interpretationen erlauben. Zudem können unterschiedliche Visualisierungsformen oder zugrundeliegende Algorithmen die Interpretation beeinflussen. Um unterschiedlichste Forschungsfragen beantworten zu können und die Interpretierbarkeit gezielt zu fördern, muss dazu eine Vielzahl an An- und Übersichten mit bedarfsabhängigen Graden an Fokus und Detail bereitgestellt werden, die das Potential der Daten sowie Fehlstellen und Unsicherheiten offenlegen.

DZ4) Unterstützung von Folgeforschung: Zugriff auf die Datenquellen (z.B. Dokumente, Briefe, Publikationen) müssen direkt in der Visualisierung über URIs verfügbar sein, um weitere Recherchen zu ermöglichen. Visualisierungsergebnisse, spezifische Ansichten und die Daten selbst müssen speicherbar, zu verlinken und reproduzierbar sein.

Diese Designziele müssen in der gesamten Daten-Nutzungs-Pipeline – von der Datentransformation über die Aufbereitung der transformierten Daten durch Visualisierungen bis zur möglichen Nachnutzung der Daten und Offenlegung der Prozesse – angesetzt und mitgedacht werden. Interaktivität spielt hier eine Schlüsselrolle, um Bewegungen zwischen zusätzlichen Detailgraden und vereinfachten Darstellungen zu ermöglichen. Die folgenden Abschnitte beschreiben exemplarisch Strategien zur Förderung von Transparenz mit Verweisen auf die konkreten Designziele (DZ1–DZ4).

Unterschiedliche Ansichten für unterschiedliche Bedarfe

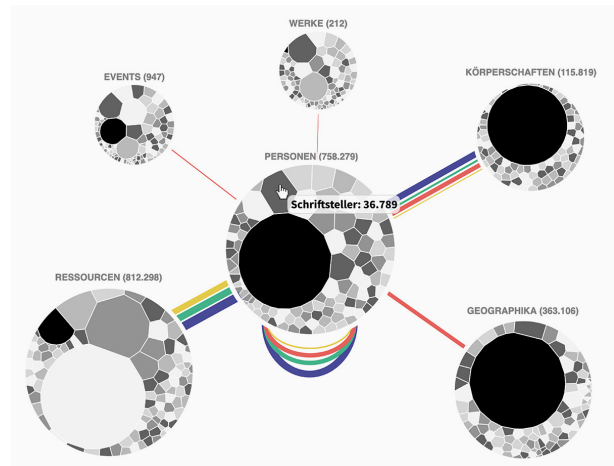


Abb. 2: Eine Übersicht bietet eine Meta-Perspektive auf ~52 Mio. Knoten und ~185 Mio. Kanten ab und zeigt z.B. akkumuliert die häufigsten Berufsgruppen von Personen für eine Zeitauswahl.

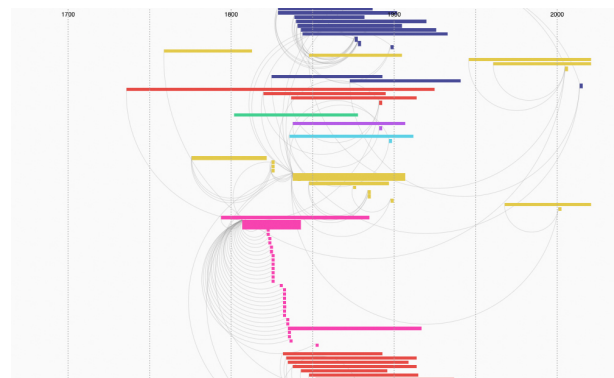


Abb. 3: Ein Force-Layout lässt sich dynamisch zu einer Zeit-basierten Ansicht von Knoten transformieren, bei der Knoten mit Hilfe eines Community-Algorithmus angeordnet werden.

In einem experimentellen *Sandcasting* Prozess (Hinrichs et al. 2019) wurden verschiedene Ansichten entwickelt, die eine Exploration der Daten aus unterschiedlichen Perspektiven und mit Fokus auf eine Vielzahl von Daten-Dimensionen ermöglichen (Dörk et al. 2017; Whitelaw 2015) (DZ3). Entstanden sind Ansichten, die als Zugang eine gesamt-datenbasierte, akkumulierte Überblicks-Ansicht nach dem Prinzip „*Overview first [...] then details-on-demand*“ (Shneiderman 1996) mit einzelnen suchbasierten Ansichten nach dem Prinzip „*Search, Show Context, Expand on Demand*“ (van Ham & Perer 2009) kontrastieren. Die übersichtsbasierte Ansicht (Abb. 2) dient dazu, die Relevanz für die Bearbeitung von Forschungsfragen ermitteln zu können, und um darzulegen, auf welchen Daten die Visualisierungen basieren (DZ1). Die suchbasierten Ansichten sollen unterschiedliche Perspektiven auf Basis gezielter Anfragen liefern, und durch die Priorisierung von unterschiedlichen Dimensionen zusätzliche Details für mehr Interpretationsspielraum bieten. Dazu gehören auch klassische Force-basierte Netzwerkdarstellungen (Abb. 5a) und zeitbasierte Ansichten (Abb. 3) (DZ3).

Entfaltung von Kanten und deren Eigenschaften

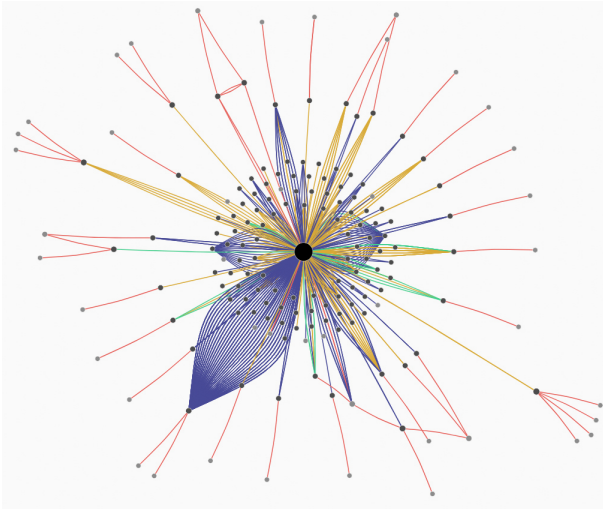


Abb. 4: Die beispielhafte Visualisierung eines Egonetzwerkes, in der alle Beziehungen separat inklusive farbiger Kodierung angezeigt werden, führt zu visuellen Überlagerungen.

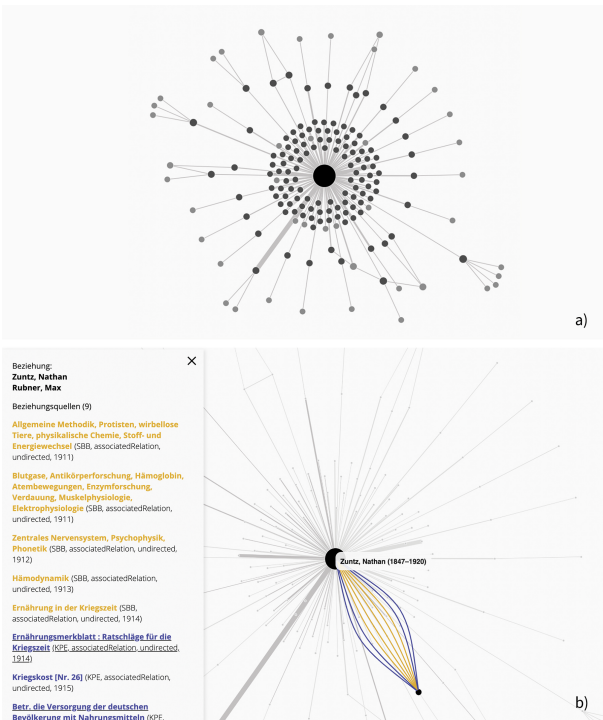


Abb. 5: Durch die iterative Entfaltung von Kanten⁷ werden in einer Grundansicht (a) mehrere Verbindungen zwischen zwei Personen zusammengefasst und die Anzahl dahinterstehender Beziehungen durch Kantenstärke dargestellt. Auswahl einer Kante führt zu bedarfsabhängigen Entfaltungen von Beziehungen (b), welche durch den gezielten Fokus Enkodierungsmöglichkeiten (z.B. für Unsicherheiten oder Kategorien) ermöglichen.

Häufig beruhen Kanten in historischen Netzwerken auf Rekonstruktionen von Beziehungen basierend auf Ressourcen wie Briefen, Tagebüchern, Protokollen oder Publikationen. Um die Datentransparenz bei der Transformation sicherzustellen, haben wir bei der Interpretation der sozialen Relationen aus den Metadaten ein Merkmal mit der ID einer Ursprungsressource an den Kanten hinzugefügt, welches eine eindeutige Identifikation zum Ursprung der Ableitung ermöglicht. Die ID referiert dabei auf einen

Metadatenatz, der als Knoten repräsentiert ist und Informationen zu Provenienzzangaben enthält (DZ1,2). Eine Herausforderung besteht in der Visualisierung von Provenienzen, da viele Beziehungen zwischen Akteuren auf mehreren Quellen basieren können, was durch die dargestellten Details zu einer Komplexitätserhöhung führen kann. Ein Netzwerk, in dem alle Beziehungen jeweils in unterschiedlichen Farben je nach Ursprungsquelle dargestellt werden, führt zu vielen Kantenüberlagerungen (Abb. 4). Unser Ansatz besteht darin, die Kanten zwischen zwei Akteuren für eine übersichtlichere Ansicht zunächst zu gruppieren (Abb. 5a), und nur bei Bedarf per Auswahl zu entfalten (Abb. 5b) (Bludau et al. 2021; Brüggemann et al. 2020). So lassen sich gezielt Details anzeigen und Quellen verlinken (DZ1,4).

Nachnutzung und Dokumentation

Die Verlinkung in der Visualisierung zu Ursprungs-URIs von Knoten und Kanten ermöglicht es Nutzer*innen direkt Daten zu überprüfen oder Recherchen fortzusetzen (DZ1,4). Weiterhin sind die Verlinkung und Speicherung von Visualisierungsansichten (inklusive Filterungen und Selektionen), aber auch der gezielte Daten-Download wichtige Bestandteile der Transparenz (DZ4). Die Nutzer*innenstudie und Interviews haben auch gezeigt, dass insbesondere für Anwender*innen mit Fokus auf quantitative Analysen der einfache Zugang zu den Daten selbst wesentlich wichtiger ist als deren Visualisierung, und dass die Dokumentation der Prozesse um das Daten-Retrieval und die Visualisierung eine hohe Priorität haben. Um weitere Analyse-Möglichkeiten bereitzustellen, setzen wir zusätzlich zu Explorations-fokussierten Visualisierung auf einen Daten-Zugang über dokumentierende und als Einführung dienende Jupyter-Notebooks. Diese sind durch die Verbindung von Code und beschreibendem Text eine in sich geschlossene Dokumentation und sind so in der Lage in der Visualisierung dargestellte Prozesse festzuhalten, sie reproduzierbar zu machen, aber auch darüberhinausgehende Analysen und statistische Methoden zu ermöglichen (DZ2). Aufbereitet in einem Curriculum für historische Netzwerkanalyse mit Jupyter Notebooks werden hierdurch auch (mit diesen Methoden unerfahrene) Nutzer*innen zur Verwendung parametrisierbarer Methoden, die über die Visualisierung hinausgehen, befähigt. Weiterhin wird der Quellcode für die Visualisierungen und die Datenzusammenführungen für Nachnutzung und Transparenz frei verfügbar gemacht (DZ2,4).

Reflexion der Ergebnisse und Diskussion

Die Diskussionen in der Forschungsliteratur zur HNA, vorausgegangene Forschungsprojekte und unser eigener Forschungsprozess bestätigen: Transparenz ist eine grundlegende Voraussetzung, um Netzwerkdaten und Visualisierungen für die HNA nutzbar zu machen und in den wissenschaftlichen Diskurs einbringen zu können. Bislang mangelt es an konkreten Strategien, um Transparenz insbesondere bei der Sekundärnutzung von Daten mit Blick auf Provenienzen, Unsicherheiten und Prozesse herstellen zu können. Zugleich führt die Einführung von zusätzlichen Daten-dimensionen für Datenprovenienzen in Visualisierungen zu einer erhöhten Komplexität, die wiederum Interpretationen erschwert.

Die hier skizzierten Strategien wurden gezielt für die Exploration und Analyse von historischen sozialen Netzwerken entwickelt, anschließend beispielhafte Umsetzungsstrategien im Rah-

men des SoNAR-Projektes in Form von Prototypen präsentiert und fortlaufend evaluierend (siehe Balck et al. 2021) begleitet. Interaktivität erwies sich dabei als wirkungsvolles Werkzeug, das helfen kann, dynamisch und Nutzer*innenspezifisch größtmögliche Transparenz herzustellen, ohne zu komplexe und dadurch unzugängliche Ansichten zu erzeugen. Weitere Lösungen müssen noch für die Darstellungen von unsicheren Beziehungen gefunden werden, um den Nutzer*innen eine erste Einschätzung zu ermöglichen, etwa durch die Anzeige von Wahrscheinlichkeitsgraden auf Grundlage von wiederum nachvollziehbaren Indikatoren und Algorithmen. Großes Potential verspricht die Dokumentation und Weiterbearbeitung über Notebooks inklusive offenem Sourcecode. Hier bieten die ergänzenden Notebooks detaillierte Erklärungen und individualisierbare Analysemethoden.

Für Forschungsumgebungen mit einer perspektivisch großen Anzahl von zusammengeführten Datenrepositorien, wie in unserem Fall als Sekundärnutzung von Bibliotheks- und Archivdaten, ist die Umsetzung von umfassenden Konzepten zur Transparenz der Datenprovenienz, -modellierung, -aufbereitung und -visualisierung für einen nachprüfbar und nachvollziehbaren Einsatz in der Historischen Netzwerkanalyse zwingend notwendig.

Danksagung

Wir danken unseren Kolleg*innen aus dem DFG-Verbundprojekt SoNAR(IDH): Michael Czolkoss-Hettwer, Katrin Getschmann, Kerstin Humm, Elena Leitner, Hans-Jörg Lieder, Sina Menzel, Gerhard Müller, Clemens Neudecker, Felix Ostrowski, Vivien Petras, Larissa Schmid, Elena Leitner, Florian Richter, Julian Moreno Schneider, Michael C. Schneider, David Zellhöfer und Josefine Zinck.

Fußnoten

1. <https://sonar.fh-potsdam.de>
2. Gemeinsame Normdatei: <https://gnd.network>
3. Kalliope: <https://kalliope-verbund.info>
4. Katalog der Deutschen Nationalbibliothek: <https://www.dn-b.de>
5. Zeitschriftendatenbank: <https://zdb-katalog.de/index.xhtml>
6. Katalog der Staatsbibliothek zu Berlin: <https://stabikat.de>
7. Demo-Video: <https://sonar.fh-potsdam.de/demos/kantenentfaltung.mp4>

Bibliographie

Balck, S. / Menzel, S. / Petras, V. / Schnaitter, H. / Zinck, J. (2021): "Fluch und Segen der Visualisierung: Unterschiedliche Zielfunktionen im Forschungsprozess der historischen Netzwerkanalyse", in: *DHd 2022 Kulturen des digitalen Gedächtnisses. Konferenzabstracts. Tagung des Verbands Digital Humanities, March 7-11, Potsdam, Germany*.

Bludau, M.-J. / Dörk, M. / Fangerau, H. / Halling, T. / Leitner, E. / Menzel, S. / Müller, G. / Petras, V. / Rehm, G. / Neudecker, C. / Zellhöfer, D. / Moreno Schneider, J. (2020): "SoNAR (IDH): Datenschnittstellen für historische Netzwerkanalyse", in: Schöch, C. (ed.), *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Poster. Konferenzabstracts. Tagung des Verbands Digital Humanities, March 2-6, Paderborn, Germany*: 360–362.

Bludau, M.-J. / Dörk, M. / Tominski, C. (2021): "Unfolding Edges for Exploring Multivariate Edge Attributes in Graphs", in: Byška, J. / Jänicke, S. / Schmidt, J. (eds.), *EuroVis 2021—Posters*. The Eurographics Association.

Brüggemann, V. / Bludau, M.-J. / Dörk, M. (2020): "The Fold: Rethinking Interactivity in Data Visualization", in: *DHQ: Digital Humanities Quarterly* 14(3). <http://www.digitalhumanities.org/dhq/vol/14/3/000487/000487.html> [letzter Zugriff 15. Juli 2021].

Campbell, S. / Yu, Z. Y. / Connell, S. / Dunne, C. (2018): "Close and Distant Reading via Named Entity Network Visualization: A Case Study of Women Writers Online", in: *Proceedings of the 3rd Workshop on Visualization for the Digital Humanities. VIS4DH*.

Chen, K. / Dörk, M. / Dade-Robertson, M. (2014): "Exploring the Promises and Potentials of Visual Archive Interfaces", in: *Proceedings of the 2014 iConference* 735–741.

Dörk, M. / Pietsch, C. / Credico, G. (2017): "One view is not enough: High-level visualizations of a large cultural collection", in: *Information Design Journal* 23(1): 39–47.

Drucker, J. (2011): "Humanities approaches to graphical display", in: *Digital Humanities Quarterly* 5(1): 1–21. <http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html> [letzter Zugriff 15. Juli 2021].

Düring, M. / Eumann, U. / Stark, M. / Keyserlingk, L.v. (2016): *Handbuch Historische Netzwerkforschung. Grundlagen und Anwendungen*, Berlin: LIT-Verlag.

Düring, M. / Grandjean, M. (2021): *Journal of Historical Network Research Bibliography* 7. <https://historicalnetworkresearch.org/bibliography/> [letzter Zugriff 15. Juli 2021].

Griffith, D. M. / Veech, J. A. / Marsh, C. J. (2016): "cooccur: Probabilistic Species Co-Occurrence Analysis in R. Journal of Statistical Software", in: *Journal of Statistical Software, Code Snippets* 69(2): 1–17.

Hadlak, S. / Schumann, H. / Schulz, H.-J. (2015): "A Survey of Multi-faceted Graph Visualization", in: *Eurographics Conference on Visualization (EuroVis) - STARS*.

Hinrichs, U. / Forlini, S. / Moynihan, B. (2019): "In defense of sandcastles: Research thinking through visualization in digital humanities", in: *Digital Scholarship in the Humanities* 34(1): 80–99.

Isenberg, P. / Zuk, T. / Collins, C. / Carpendale, S. (2008): "Grounded evaluation of information visualizations", in: *Proceedings of the 2008 Workshop on Beyond time and errors: novel evaluation methods for Information Visualization*: 1–8.

Kerschbaumer, F. / von Keyserlingk, L. / Stark, M. / Düring, M. (2020): *The Power of Networks: Prospects of Historical Network Research*. Routledge.

Kleymann, R. / Stange, J.-E. (2021): "Towards Hermeneutic Visualization in Digital Literary Studies", in: *DHQ: Digital Humanities Quarterly* 15(2). <http://www.digitalhumanities.org/dhq/vol/15/2/000547/000547.html> [letzter Zugriff 01. Dezember 2021].

Lamqaddam, H. / Moere, A. V. / Abele, V. V. / Brosens, K. / Verbert, K. (2020): "Introducing Layers of Meaning (LoM): A Framework to Reduce Semantic Distance of Visualization", in: *IEEE Transactions on Visualization and Computer Graphics* 27(2): 1084–1094.

Lee, B. / Plaisant, C. / Parr, C. S. / Fekete, J.-D. / Henry, N. (2006): "Task taxonomy for graph visualization", in: *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors Novel Evaluation Methods for Information Visualization - BELIV* 1.

Nobre, C. / Meyer, M. / Streit, M. / Lex, A. (2019): "The State of the Art in Visualizing Multivariate Networks", in: *Computer Graphics Forum* 38(3): 807–832.

Novak, J. / Micheel, I. / Melenhorst, M. / Wieneke, L. / Dürring, M. / Moron, J. G. / Pasini, C. / Tagliasacchi, M. / Fraternali, P. (2014): "HistoGraph—A Visualization Tool for Collaborative Analysis of Networks from Historical Social Multimedia Collections" in: *18th International Conference on Information Visualization*: 241–250. <https://doi.org/10.1109/IV.2014.47>.

Ragan, E. D. / Endert, A. / Sanyal, J. / Chen, J. (2016): "Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes", in: *IEEE Transactions on Visualization and Computer Graphics* 22(1):31–40.

Scholtes, I. (2017): "When is a network a network? multi-order graphical model selection in pathways and temporal networks", in: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*: 1037–1046.

Shneiderman, B. (1996): "The eyes have it: A task by data type taxonomy for information visualizations", in: *Visual Languages, 1996. IEEE Symposium on Proceedings*: 336–343.

van Bree, P. / Kessels, G. (2017): "nodegoat: Enabling Explorative Research", in: *Digital Humanities Conference*.

van Ham, F. / Perer, A. (2009): "'Search, show context, expand on demand': Supporting large graph exploration with degree-of-interest", in: *IEEE Transactions on Visualization and Computer Graphics* 15(6): 953–960.

Whitelaw, M. (2015): "Generous Interfaces for Digital Cultural Collections", in: *DHQ: Digital Humanities Quarterly* 9(1). <http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html> [letzter Zugriff 15. Juli 2021].

„Wertlose“ Taggings und ihr Nutzen für die Kunstgeschichte

Poetis, Panoria

panoria@power-group.net
Ludwig-Maximilians-Universität München, Germany

Radmacher, Emilia

emilia.radmacher@outlook.com
Ludwig-Maximilians-Universität München, Germany

Smiatek, Katharina

katharinasmiat@aol.com
Ludwig-Maximilians-Universität München, Germany

Schneider, Stefanie

stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Germany

Mit der Wiederentdeckung der Kunstbetrachtung als produktivem Prozess im 20. Jahrhundert wandelt sich auch die Beziehung zwischen Betrachter:in und Werk (Görner 2007). In Anknüpfung an die Rezeptionsforschung des vergangenen Jahrhunderts ist eine erneute Transformation der Rolle der Rezipient:innen zu aktiven „co-creator[s]“ (van de Vall 2013: 111) zu beobachten. Die produktive Kunstbetrachtung in digitalen Spielen generiert bei-

spielsweise große Mengen an Daten, die kunsthistorische Untersuchungen möglich machen. Dies stellt sich auch bei den *Games with a Purpose* der an der Ludwig-Maximilians-Universität München entwickelten Internetplattform ARTigo¹ unter Beweis, in denen „co-creator[s]“ den Grundstein der datenbasierten kunsthistorischen Untersuchung bilden. Die dort agierenden Spieler:innen vergeben in fünf Minuten möglichst viele beschreibende Schlagwörter für fünf Kunstwerke. Ziel ist es, digitale Reproduktionen von Kunstwerken zu verschlagworten und zu kategorisieren (Becker et al. 2018b). In der kunstwissenschaftlichen Auseinandersetzung mit dem dadurch generierten Datenpool stellt sich die Frage, welche Auswirkungen Irrtümer der Rezipient:innen auf die Forschungsergebnisse haben. Untersucht wurde daher, ob fehlerhafte Zuordnungen von Künstler:innennamen zu kunsthistorisch verwertbaren Ergebnissen führen. Im Folgenden werden diese Ergebnisse an einem konkreten Beispiel veranschaulicht.

Stand der Forschung

Grundsätzlich ist der Bereich der gezielten Auswertung fehlerhafter Tags noch kaum erforscht. Dennoch gibt es Publikationen, die das Thema wesentlich tangieren und eine gezielte Einordnung dieser Arbeit ermöglichen. So erörtert Hänger (2008) ein Verfahren zur schnelleren Erschließung neuer Publikationen im Rahmen der bibliothekarischen Arbeit. Dabei macht er darauf aufmerksam, dass die freie Verschlagwortung von Nutzer:innen mit systemimmanenten Problemen konfrontiert ist: „Die Erschließung von Dokumenten mit freien, nicht normierten Schlagwörtern führt zu [...] Unschärfen bei der Recherche bei möglichen Homonymen oder Synonymen“ (Hänger 2008: 66). Dieses Problem der Unschärfe prägt auch unsere Datenbereinigung im Rahmen der Untersuchung des ARTigo-Datensatzes. Hänger (2008: 64–67) macht zwar ebenfalls auf von ihm als „bad tags“ bezeichneten Fehlzusammenhänge aufmerksam, doch werden die spezifischen Qualitäten dieser Tags und mögliche Gründe für ihre Zuweisung nicht ausgewertet. Guy und Tonkin (2006) werfen zudem die Frage auf, ob die bei der Bereinigung entstehende Kategorisierung in gute und schlechte Taggings nicht vorschnell ist. Sie vermuten, dass als „schlecht“ bzw. „wertlos“ deklarierte Taggings dennoch Informationsgehalt haben und deren Tilgung eine Form der Verzerrung generieren könnte.

Anders gehen Charton et al. (2013) vor, indem sie am Beispiel von Kochrezepten *opinion mining* betreiben. Dabei stellen sie Strategien vor, die Variabilität frei vergebener Taggings maschinell zu umgehen – und damit das Kategorisierungsproblem, das bei manuell vergebenen Taggings zu oben beschriebenen Unschärfen führen kann. Auch Petz (2019), Reinel (2018) und Osherenko (2010) sind der hiesigen Forschungsarbeit dahingehend ähnlich, als die getaggten Objekte (Text, Bild etc.) eine natürliche Interpretations- und Wahrnehmungsvariabilität aufweisen, die in der Folge zu einer Menge subjektiv gefärbter Taggings führen. Neben dieser subjektiven Färbung des Datensatzes, die wertvolle Aufschlüsse über kollektive Einschätzungen und damit die Kategorisierung und Klassifizierung von Objekten geben kann, gibt es eine zweite Art der Färbung: unbeabsichtigte Irrtümer der taggenden Personen. Gegenüber dem *opinion mining* kann im Falle des ARTigo-Datensatzes damit eher von einer Form des bildbasierten *relation minings* auf Grundlage visueller Ähnlichkeit gesprochen werden. Genau diesem Desiderat wollen wir uns im Folgenden widmen.

Daten

Seit 2007 wurden in ARTigo 9,7 Millionen Taggings vergeben, die auf 295.343 Tags zurückzuführen sind (Becker et al. 2018a). Mehrheitlich stellen diese Taggings inhaltsbasierte Tags (*surface tags*) dar, d. h. sie nennen im jeweiligen Werk abgebildete Objekte oder Gegenstände. Einige Markierungen beschreiben jedoch auch von dem Werk übermittelte Emotionen, nennen die Namen der Künstler:innen oder sogar den Titel des Objekts (Bry et al. 2013: 2). Die Spieler:innen sind in ihrer Wahl der Markierungen völlig frei und werden weder vom Spiel noch von der Webseite geleitet oder beeinflusst (Schneider und Kohle 2017: 83). Unter diesen Bedingungen ist eine Klassifizierung und Kategorisierung der Tags a posteriori nicht zweifelsfrei durchführbar ist: Homonyme, Dichotomien, Schreibfehler und Inhalt-Kontext-Interferenzen verursachen mehrdeutige Tags. Da die Intention der Spielenden im Moment des Taggings nicht evaluiert werden kann, ist nicht jeder Tag korrekt zu kategorisieren.

Methoden

Ziel der Datenverarbeitung war es, Tags von Künstler:innennamen aus dem ARTigo-Korpus zu isolieren und diese nach Künstler:innen zu aggregieren. Unter Berücksichtigung der tatsächlichen Autor:innenschaft sollen Verbindungen zwischen „fälschlich“ getaggeten Künstler:innen sicht- und quantifizierbar gemacht werden.

Datenbereinigung

Da Künstler:innennamen selten vollständig und korrekt angegeben werden, müssen geringe Variationen der Rechtschreibung erkannt und denselben Künstler:innen zugeordnet werden. Für sämtliche im Datensatz vertretenen Künstler:innen wurden daher für Nachnamen und volle Namen alle möglichen case-insensitiven Tags generiert, sodass der Datensatz tatsächlich nach diesen möglichen Tags durchsucht werden konnte. Weil jeder vergebene Tag einem bestimmten Bild zugeordnet ist, lässt sich hieraus bestimmen, ob attribuierte und tatsächliche Künstler:innen übereinstimmen. Die dabei entstehende Zuordnung bestimmter Künstler:innen zu bestimmten Bildern kann auch falsch sein, beispielsweise wenn ein tatsächlich von Claude Monet stammendes Bild mit „Gauguin“ getagget wird. Diese Relation zwischen zugewiesenen Künstler:innen und tatsächlichen Urheber:innen fungiert also als Klassifikator und muss grundsätzlich in zwei Richtungen berücksichtigt werden, die im Folgenden als „Künstler:in zu Bild“ sowie „Bild zu Künstler:in“ bezeichnet werden:

1. „Künstler:in zu Bild“: Bilder, die einem:r Künstler:in per Tag zugewiesen werden, wodurch für eine:n Künstler:in alle Bilder summiert werden können, auf denen sein/ihr Name angegeben wurde. Gemein haben diese Bilder die Zuweisung einer (evtl. vermeintlichen) Autor:innenschaft: Dies betrifft beispielsweise alle Bilder, die mit „Monet“ getagget wurden, unabhängig davon, ob sie tatsächlich von Claude Monet stammen.
2. „Bild zu Künstler:in“: alle auf einem Bild fälschlich annotierten Künstler:innen, sodass zu allen Bildern eines:r Künstler:in aggregiert werden kann, welche anderen Künstler:innen häufig falsch annotiert wurden. Gemein haben diese Bilder den/ die tatsächlich schaffende:n Künstler:in: So finden sich auf

Bildern von Claude Monet beispielsweise die Tags „Manet“ oder „Gauguin“.

Während „Künstler:in zu Bild“-Verknüpfungen trivial über *matching* der generierten Namenstags extrahiert werden können, zeigt sich die Qualität der extrahierten „Bild zu Künstler:in“-Relationen als unbrauchbar, da polysemantische Nachnamen wie „Schwarz“, „Berg“ oder „Strauch“ die Ergebnisse verfälschen.² In diesen Fällen ist nicht mehr nachzuvollziehen, ob die Spieler:innen „Schwarz“ als inhaltsbasierten oder attribuierenden Tag intendiert haben. Der geringe Bekanntheitsgrad von Künstler:innen wie Gustav Schwarz und David Berg legt jedoch nahe, dass der Großteil dieser Nennungen wahrscheinlich auf inhaltsbasierte Tags zurückzuführen ist. Gestützt wird diese Annahme durch das in der Forschung bekannte Phänomen, dass ESP-Spiele wie ARTigo dazu neigen, besonders generische Tags zu produzieren (Robertson, Vojnovic und Weber 2009: 2–3). Auch die Spieler:innen von ARTigo verschlagworteten inhaltsbasierte deutlich häufiger als attribuierende Tags: Selbst die Tagvolumina häufig erkannter Künstler:innen bestehen nur zu jeweils etwa 5 Prozent aus deren Namen.

Durch Berücksichtigung der Wortfrequenz, Zuhilfenahme einer multilingualen Rechtschreiberkennung und Auswahl geschickter Nebenbedingungen ist es jedoch möglich, mehrdeutige Tags sinnvoll zuzuordnen. Grundsätzlich haben inhaltsbasierte Tags wie „schwarz“, „weiß“ oder „Berg“ eine deutliche höhere Worthäufigkeit als tatsächlich relevante Namen wie „Klee“, „Macke“ oder „Turner“. Um auch fälschlich erkannte englische Wörter tilgen zu können, wird zusätzlich die englische Worthäufigkeit berücksichtigt; wobei im Englischen vorkommende deutsche Wörter sowie in den allgemeinen Sprachgebrauch eingegangene Namen weitere Störfaktoren darstellen. Zur besseren Menschenlesbarkeit stellt das Python-Paket *wordfreq* (Speer et al. 2018) hierzu die Zipf-Häufigkeit für englische und deutsche Wörter zur Verfügung, die sich auf einer logarithmischen Skala von 0 bis 8 bewegt. Unter Berücksichtigung der oben erläuterten *trade-off*-Bedingungen wurden eine Worthäufigkeit < 2,0 für englische Wörter, < 3,5 für deutsche Wörter oder die Nichterkennung des Namens durch deutsche bzw. englische Rechtschreibkorrektur als zufriedenstellende Kriterien festgelegt. Eine geringe Anzahl dadurch fälschlich getilgter Künstler:innen wurde manuell wieder hinzugefügt (Tab. 1).

Tab. 1: Zipf-Häufigkeiten relevanter Wörter. Farblich markiert ein beispielhaftes Schwellenwertdilemma aus entweder fälschlich getilgtem Künstler:innennamen (rot) oder fälschlich beibehaltenem Wort (grün).

Wort	Zipf DE	Zipf EN	Wort	Zipf DE	Zipf EN
schwarz	5,02	2,93	Klee	3,45	2,56
Manet	–	2,55	Turner	3,81	4,09
Hügel	4,12	–	Marc	4,34	4,06

Da ein zu niedriger Schwellenwert der Worthäufigkeit relevante Namen fälschlicherweise tilgt, während ein zu hoher Schwellenwert irrelevante Namen fälschlicherweise *nicht* tilgt, ist eine optimale Trennung nicht mehr möglich. Der Vergleich in Abb. 1 zeigt gleichwohl, dass Artefakte störender polysemantischer Namen erheblich reduziert werden können.

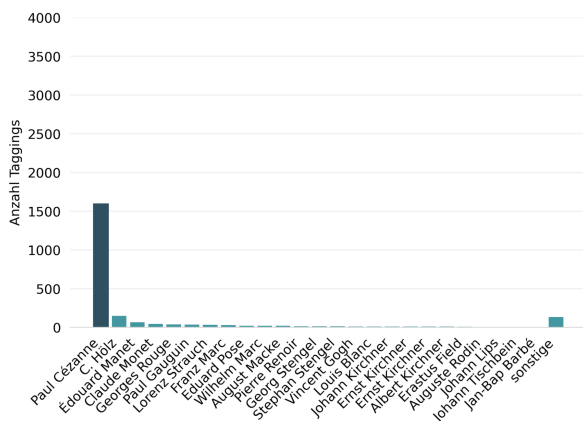
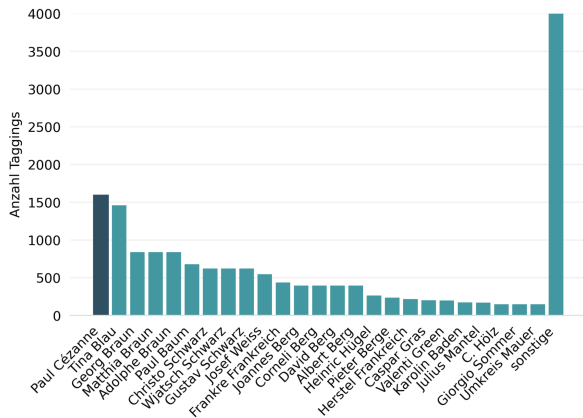


Abb. 1: „Bild zu Künstler:in“-Relationen für Paul Cézanne, vor (oben) und nach (unten) Bereinigung der Daten. Deutlich erkennbar die zunächst geringe Anzahl tatsächlicher Künstlerzuordnungen.

Ähnlichkeitsberechnung

Die so bereinigten Daten lassen für alle in der Datenbank vorhandenen Künstler:innen „Künstler:in zu Bild“- und „Bild zu Künstler:in“-Relationen berechnen. Werden diese Verteilungen gegeneinander angetragen, ergibt sich eine Konfusionsmatrix, deren Diagonale alle „korrekt“ erkannten Verbindungen darstellt; die Anzahl der Bilder also, denen ihre tatsächlichen Urheber:innen zugeschrieben wurden. Ablesbar wird dadurch im Umkehrschluss auch das Verhältnis „korrekter“ zu „falscher“ Tags, wodurch besonders oft falsch zugeschriebene oder unerkant gebliebene Künstler:innen, und in einem zweiten Schritt auch Bilder, ermittelt werden können.

Für ausgewählte Künstler:innen, vorrangig Impressionist:innen, wurde unter Verwendung der Kosinusdistanz die Ähnlichkeit der „Künstler:in zu Bild“-Verknüpfungen berechnet. Die so bestimmte Ähnlichkeit der Vektoren entspricht der Ähnlichkeit der „verwechselten“ Künstler:innen, d. h. eine größtmögliche Ähnlichkeit besteht, wenn zwei Künstler:innen denselben Künstler:innen fälschlicherweise zugewiesen worden sind. Auch hier wurden die Berechnungen durch polysemantische Namen und andere Artefakte verfälscht. Weitere Vorbedingungen konnten jedoch ein nahezu optimal gefiltertes Ergebnis liefern: der Ausschluss nicht

mindestens einmal korrekt erkannter Künstler:innen sowie der Ausschluss von Künstler:innen mit weniger als 20 Bildern im Datensatz (Abb. 2).

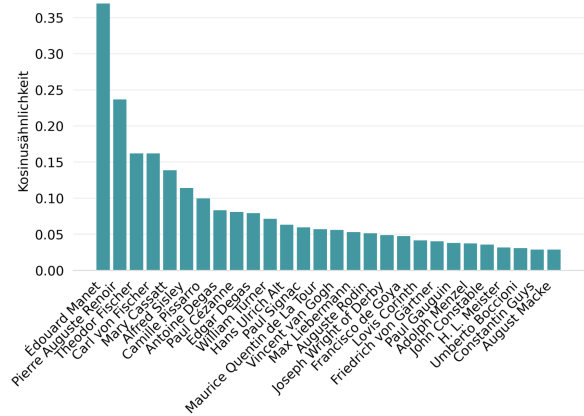


Abb. 2: Kosinusähnlichkeit der „Künstler:in zu Bild“-Relationen für Claude Monet. Polysemantische Künstler:innennamen konnten vollständig getilgt werden.

Ergebnisse

Wie auf Abb. 3 erkennbar, können bei der Untersuchung der häufigsten Fehlzuordnungen gewisse Dominanzen herausgearbeitet werden. Einerseits werden plausible Verwechslungen von Zeitgenoss:innen gleicher Epochen deutlich, wie an der Häufigkeit der Verwechslung von Turner und Constable als auch Monet und Renoir ersichtlich wird. Beide Künstlerpaare sind durch ihre nationale Zugehörigkeit, Arbeitsweise und Lebenszeit eng miteinander verknüpft. Gleiches findet sich bei Claude Monet, dessen Werke häufig mit „Manet“ und „Renoir“ getaggt werden. Andererseits veranschaulicht die Einfärbung der Namen nach Epochen eine Verflechtung romantischer und impressionistischer Strömungen: Eine Wechselbeziehung zwischen den Epochen scheint sich im Datensatz abzuzeichnen, die eine präzisere Überprüfung verlangt.

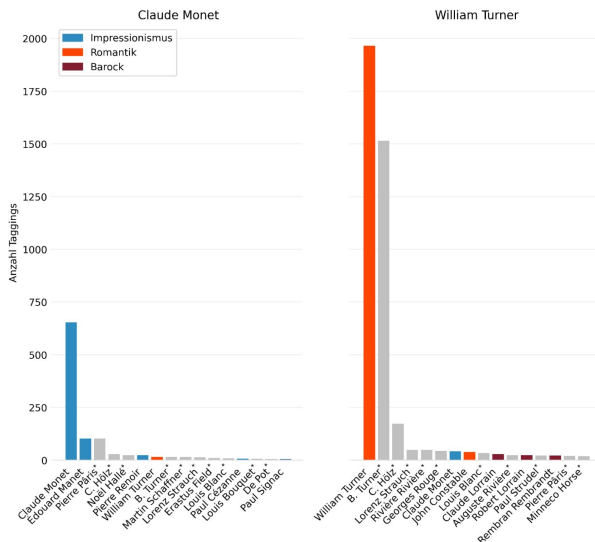


Abb. 3: Häufigkeit inkorrekt er Namenstags auf Werken von Claude Monet (links) und William Turner (rechts). Einfärbungen repräsentieren Epochen, denen die jeweiligen Künstler primär zugeordnet werden. Grau eingefärbte Künstler:innennamen sind Homonyme, die nicht aus dem Datensatz herausgefiltert werden konnten. Namen, die mit „*“ gekennzeichnet sind, zeigen ambivalente Namen an, die sowohl Homonyme als auch gezielte Namensangaben sein könnten. Eine definitive Zuordnung kann nicht vorgenommen werden, da die Intention der Spieler:innen nicht bekannt ist.

Verwechslungen von Zeitgenoss:innen, die zum Teil gemeinsam arbeiteten und die gleichen Motive malten (z. B. Monet und Renoir) werden unterbrochen von Verknüpfungen, die nicht zwangsweise hätten erwartet werden können. Evident ist, wie häufig Monet und Turner verwechselt werden: Monet wird auf Bildern von Turner noch vor Constable annotiert. Dass sich Monet intensiv mit den Werken Turners befasst hat, ist bekannt (Herrmann 2007: 244, 286; Pickeral 2007: 244, 286). Während Turner der Romantik zugeordnet wird, gilt Monet als Pionier und Wegbereiter des Impressionismus (Keller 1985: 66–81, 125; Koch 1977: 5; Wagner 2011: 117–123). Slap (1983: 183) weist allerdings darauf hin, dass Turners Spätwerk bereits erste impressionistische Züge aufweist und damit die Entwicklung des Impressionismus unwissentlich vorbereitete. Das Auslesen der Werke mit merklicher Verwechslungshäufigkeit zwischen Turner und Monet zeigt keine Dominanz einzelner Werke – die Verteilung der falschen Tags ist relativ heterogen zwischen den Werken aufgeteilt. Bei der Analyse der Entstehungsjahre der Einzelwerke offenbart sich, dass alle Arbeiten Monets, die Turner zugeordnet wurden, nach 1872 entstanden sind. Dieser Zeitpunkt koinzidiert nahezu mit der ersten Reise Monets nach London. Monet befand sich 1870 und 1871 während des Deutsch-Französischen Kriegs im Londoner Exil, wo er erstmalig mit Turners Werken in Berührung kam (Collins 2004: 709). Die Angaben der ARTigo-Crowd scheinen hier zu belegen, dass Monet nach dem Zusammentreffen mit den Bildern Turners stilistische Merkmale in seiner Arbeitsweise adaptierte (Abb. 4).

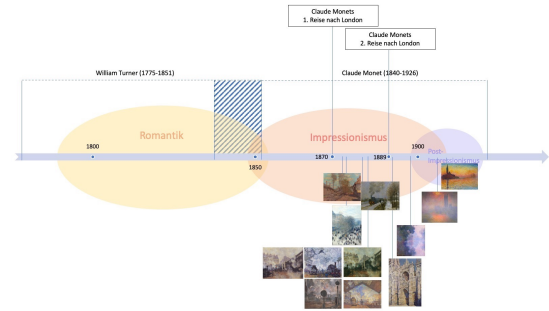


Abb. 4: Zeitstrahl mit Lebzeiten von Turner und Monet mit Verortung der korrelativen Werke Monets. Einzelbilder aus Becker et al. (2018a).

Zusammenfassung

Die Auswertung eines konkreten, kunstwissenschaftlichen Beispiels zeigt, dass falsch zugeordnete Künstler:innennamen Aufschluss über epochen- und generationsübergreifende Bezugnahmen geben können. So wiesen spezifische Einzelbildanalysen nach, dass Falschzuweisungen in Werken von Claude Monet und William Turner auf biografisch evidente Entwicklungen im Œuvre der Künstler zurückzuführen sind. Datierungen der Werke, die dominante Häufigkeiten von Verwechslungen aufweisen, geben dabei Hinweise auf konkrete Veränderungen in der Künstlerbiografie. Eine Analyse der Einzelbilder mit hoher Verwechslungshäufigkeit widerlegt dabei die Hypothese willkürlicher Verwechslung von Künstler:innen und Kunstwerken durch die ARTigo-Spieler:innen und exemplifiziert die chronologisch-biografische Evidenz der Verwechslung aufgrund stilistischer Ähnlichkeiten.

Was bedeutet dies nun für die traditionelle kunstwissenschaftliche Analyse? Unsere Untersuchungen zeigen, dass die quantitative Datenanalyse als mögliche Erweiterung der kunsthistorischen Forschungsmöglichkeiten zu begreifen ist. Sie ermöglicht eine distanzierte Form der Forschung, die übergreifende Entwicklungen als Muster erkennbar werden lässt. So erweitern die Ergebnisse aus der Makroperspektive die Detailanalyse traditioneller Forschung. Dass der kunsthistorische Blick wie hier gezeigt durch die „Weisheit der Vielen“ (Surowiecki 2007) multiperspektivisch potenziert wird, weckt die Hoffnung darauf, neue Zusammenhänge offenzulegen, die der Kunstwissenschaft dienlich sein können.

Fußnoten

1. Auf ARTigo gibt es acht Games with a Purpose, bei denen Nutzer:innen eine automatisch ausgewählte digitale Reproduktion eines Kunstwerks auf spielerische und kompetitive Art kommentieren. Wie in Luis von Ahns ESP-Spiel (von Ahn und Dabbish 2004) treten jeweils zwei Spieler gegeneinander an und versuchen Begriffe zu finden, die das abgebildete Kunstwerk und dessen Gegenstände, Figuren oder Farben beschreiben, um Punkte zu sammeln.
2. Das Phänomen der polysemantischen Künstler:innennamen verfälscht natürlich auch die matchings der „Künstler:in zu Bild“-Verknüpfungen in gleicher Weise. Da Künstler:innen mit

polysemantischen Namen jedoch meistens relativ unbekannt sind, waren diese für die kunsthistorische Analyse nicht relevant.

Bibliographie

- Becker, Matthias / Bogner, Martin / Bross, Fabian / Bry, François / Campanella, Caterina / Commare, Laura / Cramerotti, Silvia / Jakob, Katharina / Josko, Martin / Kneißl, Fabian / Kohle, Hubertus / Krefeld, Thomas / Levushkina, Elena / Lücke, Stephan / Puglisi, Alessandra / Regner, Anke / Riepl, Christian / Schefels, Clemens / Schemainda, Corina / Schmidt, Eva / Schneider, Stefanie / Schön, Gerhard / Schulz, Klaus / Siglmüller, Franz / Steinmayr, Bartholomäus / Störkle, Florian / Teske, Iris / Wieser, Christoph (2018a): *ARTigo – Social Image Tagging [Dataset and Images]* 10.5282/ubm/data.136.
- Becker, Matthias / Bogner, Martin / Bross, Fabian / Bry, François / Campanella, Caterina / Commare, Laura / Cramerotti, Silvia / Jakob, Katharina / Josko, Martin / Kneißl, Fabian / Kohle, Hubertus / Krefeld, Thomas / Levushkina, Elena / Lücke, Stephan / Puglisi, Alessandra / Regner, Anke / Riepl, Christian / Schefels, Clemens / Schemainda, Corina / Schmidt, Eva / Schneider, Stefanie / Schön, Gerhard / Schulz, Klaus / Siglmüller, Franz / Steinmayr, Bartholomäus / Störkle, Florian / Teske, Iris / Wieser, Christoph (2018b): *ARTigo – Social Image Tagging* <http://www.artigo.org> [letzter Zugriff 2. Juli 2021].
- Bry, François / Berard, Alexandre / Lagrange, Richard / Wieser, Christoph (2013): „ARTigo: Building an Artwork Search Engine with Games and Higher-Order Latent Semantic Analysis“, in: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 1: 15–20 <https://ojs.aaai.org/index.php/HCOMP/article/view/13060> [letzter Zugriff 2. Juli 2021].
- Charton, Eric / Meurs, Marie-Jean / Jean-Louis, Ludovic / Gagnon, Michel (2013): „Using Collaborative Tagging for Text Classification. From Text Classification to Opinion Mining“, in: *Informatics* 1: 32–51 10.3390/informatics1010032.
- Collins, John (2004): „Turner, Whistler, Monet. Toronto, Paris and London“, in: *The Burlington Magazine* 146: 709–710.
- Görner, Veit (2007): *Der Betrachter als Akteur. Partizipationsmodelle in der frühen Kunst des 20. Jahrhunderts*. Diss., Hochschule für Bildende Künste Braunschweig.
- Guy, Marieke / Tonkin, Emma (2006): „Folksonomies: Tidying up Tags?“, in: *D-Lib Magazine* 12 <http://www.dlib.org/dlib/january06/guy/01guy.html> [letzter Zugriff 2. Juli 2021].
- Hänger, Christoph (2008): „Good tags or bad tags? Tagging im Kontext der bibliothekarischen Wissenserschließung“, in: Gaiser, Birgit / Hampel, Thorsten / Panke, Stefanie (eds.): *Good Tags - Bad Tags. Social Tagging in der Wissensorganisation*. Münster / New York / München / Berlin: Waxmann 63–71.
- Herrmann, Luke (2007): „Turner Whistler Monet. A Superb Three-Course Feast“, in: *The British Art Journal* 6: 83–84.
- Keller, Horst (1985): *Claude Monet*. München: Bruckmann.
- Koch, Horst (1977): *William Turner*. Ramerding: Berghaus.
- Osherenko, Alexander (2010): *Opinion Mining and Lexical Affect Sensing*. Diss., Universität Augsburg https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/1469/file/Osherenko_Dissertation.pdf [letzter Zugriff 24. November 2021].
- Petz, Gerald (2019): *Opinion Mining im Web 2.0. Ansätze, Methoden, Vorgehensmodell*. Diss., Johannes Kepler Universität Linz 10.1007/978-3-658-23801-8.
- Pickeral, Tamsin (2007): *Turner, Whistler, Monet*. München: Langen Müller.
- Reinel, Dirk (2018): *Korpusbasierte Verfahren zur Generierung lexikalischer Ressourcen für das Opinion Mining. Statistische Ansätze und deren Einsatzmöglichkeiten*. Bamberg: University of Bamberg Press 10.20378/irbo-52349.
- Robertson, Stephen / Vojnovic, Milan / Weber, Ingmar (2009): „Rethinking the ESP Game“, in: *Proceedings of the 27th International Conference on Human Factors in Computing Systems* 3937–3942 10.1145/1520340.1520597.
- Schneider, Stefanie / Kohle, Hubertus (2017): „The Computer as Filter Machine. A Clustering Approach to Categorize Artworks Based on a Social Tagging Network“, in: *Art@S Bulletin* 6: 81–89 10.5282/ubm/epub.41319.
- Slap, Joseph Wm. (1983): „William Turner's Late Style: Speculation on Its Development“, in: *American Imago* 40: 175–187.
- Speer, Robyn / Chin, Joshua / Lin, Andrew / Jewett, Sara / Nathan, Lance (2018): *LuminosoInsight/wordfreq* 10.5281/zenodo.1443582.
- Surowiecki, James (2007): *Die Weisheit der Vielen. Warum Gruppen klüger sind als Einzelne*. München: Goldmann Verlag.
- van de Vall, Renée (2013): „Transformations in Perception and Participation: Digital Games“, in: Thissen, Judith / Zwijnenberg, Robert / Zijlmans, Kitty (eds.): *Contemporary Culture. New Directions in Arts and Humanities Research*. Amsterdam: Amsterdam University Press 110–125.
- von Ahn, Luis / Dabbish, Laura (2004): „Labeling Images With a Computer Game“, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 319–326 10.1145/985692.985733.
- Wagner, Monika (2011): *William Turner*. München: C.H.Beck.

What's in a name?

Die Rolle der Sprache zur Kultivierung von inklusiven Zugängen zu Kulturerbe

High-Steskal, Nicole

nicole.high-steskal@donau-uni.ac.at
Donau-Universität Krems, Austria

Seit geraumer Zeit wird bereits darauf hingewiesen, dass die digitalen Geisteswissenschaften trotz internationaler Ausrichtung ein von westlichen Ländern geprägtes Feld sind. Dies zeigt sich einerseits in der Kartierung von DH-Netzwerken (Russell 2014; Fiormonte 2015, 2017), andererseits auch im Methodenkoffer der digitalen Geisteswissenschaften, der hauptsächlich auf westliche Sprachen ausgelegt ist (Dombrowski 2020; Galina Russell 2014). Diesen Umstand versuchen mehrere Initiativen auf unterschiedliche Weisen sichtbar zu machen und zu beheben, u.a. das Netzwerk „multilingual dh“, Global Outlook::Digital Humanities, Programming Historian oder auch das Forschungsprojekt „New Languages for NLP. Building Linguistic Diversity in the Digital Humanities“ der Princeton University. Diese Initiativen besitzen unterschiedliche Schwerpunkte, doch sprechen vor allem GO::DH und Programming Historian eine wesentliche Hürde in der Zugänglichkeit von digitalisiertem Kulturerbe und der internationalen Vernetzung digital-tätiger Wissenschaftler*innen an, nämlich

Sprache. Sowohl GO::DH als auch Programming Historian haben sich zur Aufgabe gemacht, sämtliche Texte und Unterlagen in möglichst viele Sprachen zu übersetzen, um den Zugang zu Wissen im Bereich der digitalen Geisteswissenschaften zu verbessern. Im vorliegenden Beitrag wird diskutiert, welche Voraussetzungen erfüllt werden müssen, damit die Zugänglichkeit von Daten durch Mehrsprachigkeit erhöht werden kann, und welche Rolle dabei offene Systeme, wie etwa Wikidata, spielen können. Es gilt dabei Wege zu finden, wie speziell die Dokumentation von Kulturerbe, und damit auch das kulturelle Gedächtnis, besser zugänglich gemacht und zur inklusiven und interkulturellen Zusammenarbeit zwischen heterogenen, internationalen Wissenschaftsgruppen eingesetzt werden kann, wie dies etwa auch in den CARE-Prinzipien zum ethischen Umgang mit Kulturdaten gefordert wird (Carroll et al. 2020).

Der Beitrag beruht auf Erfahrungen, die in Folge des CELSUS-Projektes gemacht wurden, das von der Autorin 2018 am Österreichischen Archäologischen Institut der Österreichischen Akademie der Wissenschaften in Wien begonnen wurde und nun die Grundlage eines an der Donau-Universität Krems in Beantragung befindlichen Projektes bildet. Ausgangspunkt des Projektes war es, die weitgehend deutsche Literatur zur archäologischen Stätte Ephesos (Türkei) als open access digital zur Verfügung zu stellen, damit dieses Gedächtnis auch türkischen Partner*innen besser zugänglich gemacht werden kann. Es wurde jedoch klar, dass eine technologische Aufbereitung allein nicht ausreicht, sondern der Faktor Sprache wesentlich ist, um Zugänge inklusiv zu gestalten.

Inklusion durch Sprache

Sprache bestimmt Teilhabe. Auch wenn sich die englische Sprache zur lingua franca der Digital Humanities entwickelt hat, kann Sprache trotzdem noch eine Barriere darstellen (Dombrowski 2020). Gerade um die Zusammenarbeit mit Wissenschaftler*innen aus anderen Sprachgruppen zu stärken, ist es wichtig, Sprache und damit einhergehende Hürden zu bedenken. Für den Abbau von Sprachhürden hat Isabel Galina Russell zwei Alternativen vorgeschlagen: entweder man beginnt Daten in mehreren Sprachen zu publizieren oder Englisch als lingua franca der digital humanities inklusiver zu gestalten (Galina Russel 2014: 314). Dieser Vorschlag mag für die wissenschaftliche Bearbeitung von manchen Themen funktionieren, doch sind häufig kulturelle Daten im Fokus geisteswissenschaftlicher Forschung, wo es wichtig sein kann, Daten in ihrer ursprünglichen Sprache zu publizieren. Beispielsweise hat Roopika Risam (2018, 2019) zuletzt sehr eindrücklich die Verbindung zwischen Postkolonialismus und den digitalen Geisteswissenschaften aufgezeigt, wo sie die Teilhabe lokaler Bevölkerungen an der Aufarbeitung ihres Kulturguts und somit auch technischen und sprachlichen Zugang zu ihren Kulturdaten und ihrem digitalen Gedächtnis fordert.

Im Hinblick auf postkolonialen Argumentationen und der besseren Einbindung von unterrepräsentierten Gruppen wurde im Projekt daher entschieden die primäre Organisationsform der Archäologie, nämlich Toponyme, mehrsprachig aufzuarbeiten, damit die deutschen Begriffe für Kolleg*innen in der Türkei einfacher auffindbar sind. Die deutschen Toponyme waren bereits weitgehend bekannt, doch gab es keine strukturierte Liste der Begriffe in ihren türkischen und englischen Varianten, weshalb es notwendig war nicht nur die deutschen Toponymlisten ins Türkische zu übersetzen, sondern zusätzlich auch die lokal verwendeten, aber in der Wissenschaft nicht rezipierten, Begriffe zu suchen. Ephesos bietet eine zusätzliche Schwierigkeit: als UNESCO-Kulturerbestätte und beliebtes Touristenziel publizieren nicht nur Wissenschaft-

ler*innen zu diesem Ort, sondern es existiert auch eine Vielzahl an populärwissenschaftlichen Publikationen. Man befindet sich somit im Spannungsfeld zwischen eingebürgerten, touristisch verwendeten Begrifflichkeiten und wissenschaftlich fundierten - aber sonst unbekannten - Termini. Um hier größtmögliche Teilhabe zu ermöglichen, wurde der Entschluss gefasst, alle Begriffe, sowohl wissenschaftliche als auch touristische Toponyme, zu dokumentieren.

Das Projekt fokussierte zunächst auf die Abstimmung der deutschen, englischen und türkischen Begriffe. Für eine möglichst einheitliche Grundlage wurde der im Ephesos-Führer von Peter Scherrer publizierte archäologische Plan herangezogen (Scherrer 1995, Scherrer - Bier 2000, Scherrer 2000; Sun et al. 2020: 5), da er besonders weit verbreitet ist, von Wissenschaftler*innen und Tourist*innen gleichermaßen verwendet wird und in englischer und türkischer Übersetzung vorliegt. Die Kartenlegenden der unterschiedlichen Übersetzungen wurden gescannt, mittels OCR digitalisiert – wobei dies nur für die deutschen und englischen Legenden zu guten Ergebnissen geführt hat, türkische Begriffe mussten manuell verbessert werden – und mit OpenRefine normalisiert. Die entstandene Liste wurde mit weiteren Karten abgeglichen, die seitdem entstanden sind und ebenfalls auf Deutsch, Englisch und Türkisch publiziert wurden, wodurch die Liste teilweise erweitert werden konnte. Ein automatisierter Abgleich mit GeoNames, Wikidata, und dem Pleiades Gazetteer hat keine nennenswerte Erweiterung des Datensatzes erbracht. Auch eine Suche nach georeferenzierten Toponymen auf Wikidata war erfolglos. Das Endergebnis war eine Liste mit 117 Toponymen in deutscher, englischer und türkischer Sprache.

Inklusion durch Netzwerke

Sprache bestimmt Netzwerke. Sprache führt nicht nur dazu, dass Hürden im Verständnis entstehen können, sondern Sprache bestimmt oft auch, mit wem man kommuniziert und welche Stimmen man erfassen kann. Eine Auswertung von Gil und Ortega (2016: 23-5) hat etwa ergeben, dass Publikationen von Personen außerhalb der Ballungszentren von Wissenschaftler*innen im "global north" oft schlichtweg nicht wahrgenommen werden und dadurch manche Forschungsbereiche und Fragestellungen aus anderen Sprach- und Kulturregionen nicht rezipiert werden. Das Netzwerk GO::DH versucht, durch eine offene Publikationsplattform interkulturellen und transdisziplinären Ansätzen aus unterrepräsentierten Regionen eine Bühne zu bieten und dadurch verstärkt die Bildung von Netzwerken zu unterstützen. In manchen Fällen kann hier aber bereits durch die Einbindung mehrsprachiger Normdaten eine gewisse Hilfestellung geboten werden.

Innerhalb des Projektes führte der zufällige Fund einer finnischen Übersetzung der Ephesos-Karte von Scherrer auf einen reichen Fundus an zusätzlichen ephesischen Toponymen in anderen Sprachen auf Wikipedia und Wikidata. Die unterschiedlichen Artikel waren zu einem großen Teil nicht miteinander verlinkt, hatten keine zusätzlichen Informationen und sind dadurch schwer auffindbar. Eine tineye-Suche nach der Ephesos-Karte von Scherrer hat eine weitere französische Übersetzung der Legenden zutage gefördert, die ebenfalls über OpenRefine in die Liste eingepflegt werden konnte. Innerhalb kürzester Zeit konnte so eine Liste an deutschen Toponymen mit türkischen, englischen, finnischen und französischen Begriffen ergänzt werden. Der zufällige Fund weiterer Begriffe hat Gils Erfahrung bestätigt und aufgezeigt, dass oft eingeschränkt innerhalb von kleinen Sprachgruppen gearbeitet wird und diese Arbeit – trotz aufwendiger Suche – sehr schwer zu finden sein kann.

Die Datengrundlage von vielen Toponymen in Wikidata hatte einen weiteren Fehler: sie waren größtenteils nicht georeferenziert und auch nicht definiert (fehlende Beschreibung, Geokoordinaten und "instance of - P31"-Felder). Popescu et al. (2009: 58) haben in ihrer Arbeit festgelegt, dass mehrsprachige Gazetteers drei Elemente unbedingt benötigen, damit große Toponym-Datensätze über Sprachgrenzen hinweg zusammengeführt werden können: 1. eine Bezeichnung, 2. GPS-Koordinaten, 3. einen Typ. Die Liste wurde daher mit GPS-Daten, geographischer Zuordnung und Kurzbeschreibung weiter ergänzt. Da viele der Grunddaten bereits – zwar verteilt – auf Wikidata zur Verfügung standen, wurde beschlossen, dass die Ergänzungen und Änderungen auf Wikidata eingespielt werden und die Plattform zur Normalisierung von unterschiedlichen Gazetteers eingesetzt werden kann. Mittels OpenRefine konnten die Änderungen automatisiert in Wikidata eingespielt werden, wodurch nicht nur die Datengrundlage auf Wikidata verbessert wurde, sondern auch etliche unverknüpfte Wikipedia-Seiten zueinander in Beziehung gesetzt wurden. Von den ursprünglich 117 Begriffen, die im Scherrer-Führer genannt werden, konnten somit 56 Begriffe auf Wikidata eingespielt, verknüpft und in fünf Sprachen ergänzt werden. Wikipedia und Wikidata wurde dadurch dazu genutzt, um die Vernetzung von Wissen in anderen Sprachen zu verbessern.

Inklusion durch Technologie

Zur Teilhabe in den digitalen Geisteswissenschaften sind infrastrukturelle Voraussetzung notwendig, wie etwa verlässliche Stromversorgung und stabile Internetverbindung, sowie Computerzugänge und Lizenzen für bestimmte Programme. Zugleich müssen Wissenschaftler*innen, die mit unterrepräsentierten Sprachen arbeiten, sehr viel Grundlagenarbeit leisten, um überhaupt Datensätze und digitale Methoden für Fragestellungen in unterrepräsentierten Sprachen anwenden zu können. Gerade im Umgang mit Kulturdaten sind Ansätze des minimal computing hilfreich, wie von Gil und Ortega (2016: 26) vorgeschlagen, aber auch offene Systeme, die durch die lokale Bevölkerung ohne aufwendige IT-Infrastruktur und technisches Vorwissen ergänzt werden können. Unserer Erfahrung nach hat sich gerade Wikidata und Wikipedia für diese Umsetzung ausgezeichnet geeignet. Dies zeigt sich auch dadurch, dass seit der Bereinigung der Datenlage im Dezember 2020 die Begriffe von anderen Nutzer*innen weiter bearbeitet wurden. Es sind für manche Bereiche sowohl sprachliche Ergänzungen (etwa Arabisch und Russisch) als auch Verweise auf Normdaten in anderen Sprachen hinzugekommen. Diese Ergänzungen wurden hauptsächlich bei Toponymen durchgeführt, die sehr bekannt sind, etwa Artemistempel oder House of Virgin Mary, eröffnen wiederum neue Möglichkeiten für das Projekt und machen es möglich, Publikationen und Quellen in Türkisch aber auch anderen (noch nicht antizipierten) Sprachen zu Ephesos zu erfassen. Zusätzlich hat sich auch gezeigt, dass einige Änderungen durch User mit mobilen Endgeräten durchgeführt wurden und somit ein niederschwelliger Zugang gerade für User mit alternativen Internetzugängen möglich war (siehe z.B. <https://www.wikidata.org/w/index.php?title=Q43018&action=history>; letzter Zugriff: 1. Dezember 2021). Die Anzahl der Änderungen ist leider nicht quantifizierbar.

Fazit

Der Aspekt der Inklusion und Teilhabe durch türkische Partner*innen in der Wissenschaft und in der Lokalbevölkerung wurde im Projekt anfangs nur im Hinblick auf technologische Zugänglichkeit gedacht. Das Anliegen des Projektes war es jedoch die Zugänglichkeit wissenschaftlicher Dokumentation einer Kulturerbestätte zu erhöhen, weshalb schnell klar wurde, dass mehr Arbeit notwendig war als nur Digitalisate online zu stellen und die Mehrsprachigkeit der nächste logische Schritt war. Obwohl das Augenmerk zunächst nur auf der türkischen Sprache lag, konnten per Zufallsfund weitere Sprachen nach einem einheitlichen Workflow hinzugefügt werden, wobei die weite Verbreitung einer einzelnen Karte die Datenlage unterstützte. Die Erfahrung zeigte, dass teilweise bereits sehr viele Daten vorhanden sind, diese aber erst gesucht und zusammengeführt werden müssen. Wikidata und Wikipedia hatten für das Projekt den Vorteil, dass die bereinigten Datensätze von anderen schnell gefunden werden können, und durch einheitlich Ansprache und Normdaten verknüpft sind. Wikidata hatte den zusätzlichen Vorteil, dass die Wikimedia-Plattform kostenfrei und niederschwellig in der Nutzung und Einpflegung von Daten ist. Die Ergebnisse des Projektes zeigen, dass dieser Prozess zur verbesserten Wahrnehmung von Forschung aus unterrepräsentierten Gruppen führen und dadurch ein multikulturelles und inklusives Gedächtnis entstehen kann.

Bibliographie

- Carroll, Stephanie / Russo, Ibrahim / Garba, Oscar L. / Figueroa-Rodriguez, Jarita / Holbrook, Raymond / Lovett, Siameon / Materechera, Mark Parsons / u. a. (2020).** "The CARE Principles for Indigenous Data Governance". *Data Science Journal* 19 (1): 1-12. <https://doi.org/10.5334/dsj-2020-043>.
- Dombrowski, Quinn** (2020): "What's a „Word“: Multilingual DH and the English Default." <https://www.quinndombrowski.com/?q=blog/2020/10/15/whats-word-multilingual-dh-and-english-default> [letzter Zugriff 15. Juli 2021]
- Fiormonte, Domenico** (2015): "Towards Monocultural (Digital) Humanities?" in: *Infolet*. <https://infolet.it/2015/07/12/monocultural-humanities/>. [letzter Zugriff 15. Juli 2021]
- Fiormonte, Domenico** (2017): "Digital Humanities and the Geopolitics of Knowledge." in: *Digital Studies/Le champ numérique* 7: 1-18.
- Galina Russell, Isabel** (2014): "Geographical and Linguistic Diversity in the Digital Humanities." in: *Literary and Linguistic Computing* 29: 307-316.
- Gallon, Kim** (2016): "Making a Case for the Black Digital Humanities" in: *Debates in the Digital Humanities*. <https://dhdebates.gc.cuny.edu/read/untitled/section/fa10e2e1-0c3d-4519-a958-d823aac989eb>. [letzter Zugriff 15. Juli 2021]
- Gil, Alex / Ortega, Élika** (2016): "Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing" in: Crompton, Constance / Jane, Richard J. / Siemens, Ray *Doing Digital Humanities*. London / New York: Routledge 22-34.
- Laurini, Robert** (2015): "Geographic Ontologies, Gazetteers and Multilingualism" in: *Future Internet* 7: 1-23.
- Laurini, Robert** (2017): "Gazetteers and Multilingualism" in: *Geographic Knowledge Infrastructure*. London: ISTE Press Ltd 157-182.

Piller, Ingrid/Takahashi, Kimie (2011): "Linguistic Diversity and Social Inclusion" in: *International Journal of Bilingual Education and Bilingualism* 14: 371–381.

Popescu, Adrian/Grefenstette, Gregory/Bouamor, Houda (2009): "Mining a Multilingual Geographical Gazetteer from the Web" in: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. Milan, Italy*: IEEE. 58–65. <http://ieeexplore.ieee.org/document/5284918/>.

Risam, Roopika (2018): "Decolonizing the Digital Humanities in Theory and Practice" in: Sayers, Jentery (ed.): *The Routledge Companion to Media Studies and Digital Humanities*. New York: Routledge, Taylor & Francis Group, 78–86.

Risam, Roopika (2019): *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston, Illinois: Northwestern University Press.

Scherrer, Peter (1995): *Ephesos - der neue Führer: 100 Jahre österreichische Ausgrabungen; 1895 - 1995*. Wien: Österreichisches Archäologisches Institut.

Scherrer, Peter (2000): *Efes: Rehberi*. Istanbul: Ege Yayınları.

Scherrer, Peter / Bier, Lionel (2000): *Ephesus: The New Guide*. Istanbul: Ege Yayınları.

Sun, Kai / Hu, Yingjie / Song, Jia / Zhu Yunqiang (2020): "Aligning Geographic Entities from Historical Maps for Building Knowledge Graphs" in: *International Journal of Geographical Information Science*: 1–30.

“Wie Wölkchen im Morgenlicht” Zur automatisierten Metaphern- Erkennung und der Datenbank literarischer Raummetaphern laRa

Schumacher, Mareike

schumacher@linglit.tu-darmstadt.de

Technische Universität Darmstadt, Germany

Bisher waren Bemühungen, theoretische Ansätze zur Metaphernforschung in praktischen informationstechnologischen Anwendungen, wie z.B. künstlichen Intelligenzen, zu implementieren wenig erfolgreich (vgl. Thaller 2021: 90). Auch der vorliegende Beitrag wird diese Lücke nicht schließen, dokumentiert aber den Versuch und zeigt auf, warum dieser scheitern musste. Mit dem Aufbau der relationalen Graphdatenbank literarischer Raummetaphern *laRa* wurde ein alternatives Recherche-Tool entworfen, das einerseits Einblicke in die Entwicklung, Etablierung und Konventionalisierung literarischer Raummetaphern gibt und andererseits literaturwissenschaftliche Fallstudien, bei denen Raummetaphern eine Rolle spielen, sinnvoll ergänzt. Der vorliegende Beitrag dokumentiert Training und Tests eines auf maschinellem Lernen basierenden Raum-Classifiers (Schumacher 2021a), der auch die Kategorie "Raummetapher" umfasst, und führt beispielhaft eine Metadaten-Netzwerkanalyse anhand der Raummetapher "Weg" durch. Auf diese Weise werden zwei digitale methodische Zugänge zum Phänomen literarischer Raummetaphern kontrastiert und gezeigt, wie Mixed-Methods-Ansätze

und Cross-Validierung einer Methode durch eine andere gewinnbringend eingesetzt werden können.

Über (Raum-)Metaphern

Zwei grundlegende Probleme bei der Betrachtung von Metaphern im Vergleich zu nicht-metaphorischen Ausdrücken sind Uneigentlichkeit und Variabilität. Metaphern werden grundsätzlich aus drei Größen konstruiert: dem sprachlichen Ausdruck, dem, was der sprachliche Ausdruck im Wortsinne bezeichnet und etwas Ähnlichem (vgl. Wenz 1997: 32). Dabei unterscheidet sich die Grammatik metaphorischer Ausdrücke meist in nichts von wörtlich gemeinten Phrasen (vgl. Thaller 2021: 91). Metaphern sind dynamische Konstrukte, die eine Entwicklung durchlaufen von Einführung, über Etablierung zur Konventionalisierung und schließlich bis hin zum Übergang in den eigentlichen Sprachgebrauch (vgl. Thaller 2021), d.h. Ausdrücke können die einmal aufgebaute Metaphorik auch wieder verlieren. Metaphern sind aber nicht nur ein interessantes sprachliches Phänomen, sondern können auch als prägende Ausdrücke menschlichen Handelns fungieren (vgl. Blumenberg 1971: 213 und Wenz 1997: 33). Sie sind eine wesentliche Basis menschlichen Denkens.

Der methodische Fokus der Digital-Humanities-Forschung zu Metaphern in deutschsprachigen Texten liegt derzeit auf manueller Annotation (vgl. Herrmann 2018), halbautomatischem Annotieren (vgl. Majoros 2013) und Automatisierungsvorbereitenden Ansätzen (z.B. Do, Gerloff und Nunez 2016). Mit MIPVU (Metaphor Identification Procedure VU University Amsterdam) wurde eine sprachunabhängige Methodik zur manuellen Annotation von Metaphern entwickelt, die Annotator*innen dabei unterstützt, intersubjektiv Metaphern zu erkennen und zu annotieren (vgl. Steen et al 2010, Majoros 2013: 68, Herrmann 2018: 185). Eine vollautomatische Erkennung literarischer Metaphern, die sowohl konventionalisierte Phrasen als auch kreative Neuschöpfungen umfasst, wurde bisher nicht entwickelt und vereinzelt sogar als unmöglich bezeichnet (vgl. Gehring und Gurevych 2014: 103). Dennoch ist die Problematik ausgesprochen wichtig, da sich Metaphern bei der computationalen Erfassung anderer Phänomene als erheblicher Störfaktor erweisen können.

Raummetaphern als Störfaktor automatisierter Erkennung und Kategorisierung literarischen Raumes

Im Rahmen meiner Dissertation *Orte und Räume im Roman* (Schumacher im Erscheinen) habe ich eine Methode zur automatisierten Erkennung und Klassifizierung von Raumreferenzen in Erzähltexten entwickelt. Die Automatisierung fußt auf einem theoriebasierten Modell, das sieben Kategorien umfasst: Orte, Relationen, relationale Verben, Raumthemen, Raumbeschreibungen, Raumhinweise und Raummetaphern. Raummetaphern beinhalten zwar Raumausdrücke, bezeichnen aber nicht-räumliche Phänomene. So steht das hier gewählte Beispiel der Raummetapher "Weg" (wie in "Weg zum Glück" oder "Lebensweg") für eine Reihe von Entscheidungen, nicht für eine geographische begeh- oder befahrbare Strecke. Im Gegensatz zu Raumsymbolen und -Motiven, bei denen räumliche Aspekte zwar zusätzlich zu ihrer Raumreferenz mit nicht-räumlicher Bedeutung aufgeladen werden, weisen Raummetaphern also keinerlei semantische Verbindung mit räumlichen (wie z.B. geographischen) Größen auf. Steht in

einem Erzähltext z.B. eine Figur auf einem Berg und dieser Berg steht symbolisch für einen Erkenntnisgewinn, so ist die Geografie des Berges innerhalb der erzählten Welt nach wie vor gegeben. Wird von einer Figur in einem Roman gesagt, sie habe "einen Berg von Arbeit vor sich", so handelt es sich bei "Berg" um eine Raummetapher, die eigentlich für eine große Menge steht. Da Raummetaphern in ihrer grammatischen Form und im sprachlichen Kontext der Referenzierung von Raum aber sehr ähneln, wurde zunächst der Versuch unternommen, sie mithilfe eines kontextsensitiven Machine-Learning-Trainings in die automatische Erkennung zu integrieren. Genutzt wurde der methodische Rahmen der Named Entity Recognition (vgl. Schumacher 2018). Entwickelt wurde ein Classifier (Schumacher 2021a), der Ausdrücke erkennt, die mehr oder weniger explizit Raum referenzieren und sie den oben genannten Kategorien zuweist. Über alle Kategorien hinweg wurde eine durchschnittliche Gesamterkennungsgenauigkeit von 75,65 % F1-Score¹ (vgl. Schumacher 2021a) erreicht. Mit 7,74 % hat die Kategorie der Raummetaphern mit Abstand die schlechteste Quote.

Trainingsprozess und Testergebnisse

Named Entity Recognition (NER) ist ursprünglich eine computeringuistische Methode zur automatischen Erkennung und Klassifizierung klar benannter Einheiten (vgl. Schumacher 2018: §1). Die am häufigsten in Named-Entity-Recognition-Tools implementierten Kategorien sind Personen, Orte und Organisationen. Für die literaturwissenschaftliche Nutzung von NER bedarf es allerdings einer Domänenadaption, bei der sowohl die implementierten Kategorien als auch die Trainingsdaten angepasst werden müssen. Die Methode wurde bereits erfolgreich für die Erkennung literarischer Figuren adaptiert (vgl. Jannidis et al. 2015) und auch eine Unterklassifizierung nach Genderzuweisungen ist möglich und für literaturwissenschaftliche Forschung gewinnbringend (vgl. Schumacher und Flüh 2020). Die Kategorie des Ortes ist für literarische Texte nahezu ebenso relevant wie die der Person bzw. Figur. Statt eines komplexen Raumkonzeptes, das in mehrere Unterkategorien aufgeteilt wird, werden bei der linguistischen Nutzung von NER-Tools lediglich Ortsnamen erkannt. Um ein NER-Tool so zu adaptieren, dass es Raumreferenzen erkennen und in eine von sieben Kategorien literarischen Raumes einsortieren kann, wurde ein Machine-Learning-Training durchgeführt. Im Folgenden werden diejenigen Ausschnitte des Trainings vorgestellt, die zeigen, inwiefern die automatisierte Erkennung von Raummetaphern dabei gescheitert ist.

Das NER Training – Testumgebung

Die Wahl des Tools fiel auf den in den Digital Humanities gut etablierten Stanford Named Entity Recognizer (Finkel, Grenager und Manning 2005), in dem kontextsensitive Conditional-Random-Fields-Algorithmen (zu CRF-Algorithmen vgl. Sutton und McCallum 2010) implementiert sind (Manning et al. 2014). Das Trainingskorpus besteht aus Ausschnitten aus 80 Romanen aus vier Jahrhunderten (18–21). Aus jedem Jahrhundert wurden 20 Erzähltexte integriert, sodass das Trainingskorpus einen gleichmäßigen Aufbau aufweist². Aus jedem Text wurden 4.000 Tokens extrahiert und ins Trainingskorpus überführt. Dabei handelt es sich um Anfangspassagen, da Romananfänge eine expository-

sche Funktion haben (vgl. z.B. Miller 1965: 9, Retsch 2000: 138, Richardson 2008: 4 und Herrmann 2018: 171) und eine Häufung von Raumreferenzen dadurch besonders wahrscheinlich ist³. Das Trainingskorpus und insgesamt acht Testtexte wurden in einem iterativen Prozess anhand detaillierter Guidelines (vgl. Schumacher forthcoming) manuell annotiert⁴. Aus den Testtexten wurden jeweils 10.000 Tokens betrachtet. Anschließend an Jannidis et al. (2015) wurde zunächst mit einem kleinen Trainingskorpus von 40.000 Tokens gearbeitet, das dann wiederum in einem iterativen Prozess auf 320.000 Tokens ausgeweitet wurde⁵. Insgesamt wurden 3 unterschiedliche Varianten des Trainings durchgeführt: mit jahrhundertspezifischen Daten (vgl. Abb. 1 links), mit Daten aus zwei Jahrhunderten (dem, aus dem der Testtext stammt und dem folgenden - vgl. Abb. 1 Mitte) und mit kumulierten Trainingsdaten aus allen vier Jahrhunderten (vgl. Abb. 1 rechts). Aus jedem Jahrhundert wurden insgesamt 80.000 Tokens ins Trainingskorpus überführt. Die Daten wurden diachron hinzugefügt.

Die Testergebnisse können wie in Abb. 1 visualisiert werden:

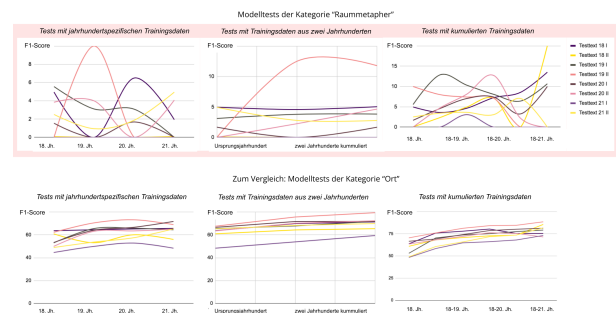


Abb. 1: Vergleich einiger Testergebnisse für das Training von Raummetaphern- und Ortserkennung

Der Vergleich mit der insgesamt am besten erkannten Kategorie "Ort" zeigt, dass sich für die Testergebnisse der Kategorie "Metapher" keine Regelmäßigkeiten ergeben. Für einzelne Testtexte zeigen jahrhundertspezifische Trainingsdaten oder Trainingsdaten aus zwei Jahrhunderten den größten Trainingserfolg. Für andere ist ein Trainingskorpus aus einem ganz anderen zeitlichen Kontext passender. Eine schrittweise Ausweitung des alle vier Jahrhunderte umfassenden Trainingsmaterials zeigt keinerlei regelmäßigen Zuwachs der Erkennungsgenauigkeit. Einen viel typischeren Trainingsverlauf zeigt die Ortskategorie. Hier ist hauptsächlich die Größe des Trainingskorpus ausschlaggebend. Je mehr Trainingsdaten eingesetzt werden, desto höher die Erkennungsquote. Auch der zeitliche Kontext ist hier nur bei wenigen Testtexten bedeutend. Insgesamt zeigt die Kumulierung der Trainingsdaten eine zwar langsame aber kontinuierliche Steigerung. Das Training der Metaphernerkennung gleicht hingegen einem Glücksspiel: Mal führt das Hinzufügen neuer Trainingsdaten zu einer Verbesserung des Classifiers, mal wird dadurch alles buchstäblich zurück auf Null gesetzt.

laRa - Datenbank literarischer Raummetaphern

Als alternatives Recherche-Tool und um besser zu verstehen, warum die Automatisierung von Metaphernerkennung so problematisch ist, wurde die relationale Graphdatenbank literarischer

Raummetaphern *laRa* (Schumacher 2021b) aufgebaut. Dazu wurden die im Trainingsprozess des maschinellen Lernens generierten Daten genutzt. Das heißt, die manuell annotierten Raummetaphern im Trainingskorpus wurden in eine Neo4J-Graphdatenbank (Graph Data Modeling Concepts and Techniques for Neo4J 2021) übertragen und dabei manuell mit Metadaten angereichert. Zu jeder Metapher wurde im Sinne der Kernmetapher von Lakoff und Johnson (1998: 9) ein Kernwort bestimmt, das dem am stärksten raumreferentiell besetzten Ausdruck entspricht. Außerdem wurde festgehalten, ob es sich um eine Ein-Wort-Metapher (Metapher, die aus einem einzigen Wort besteht), eine raummetaphorische Phrase (kurze metaphorische Phrase von ca. 3-5 Wörtern) oder ein Raumbild (komplexe metaphorische Konstruktion, die sich mindestens über einen halben Satz erstreckt) handelt. Zu jeder Metapher wurde ebenfalls manuell und in Interpretation des Kontextes eine Deutungsmöglichkeit hinzugefügt. Auf diese Weise wurde aus jeder Metapher ein Knotenpunkt mit drei Propertyts. Jeder dieser Knotenpunkte wurde über Relationen mit den Quellen verbunden, in denen die jeweilige Raummetapher vorkommt. Die Quellen bilden also ebenfalls Knotenpunkte, die mit den Propertyts "Titel", "Autor*in" und "Erscheinungsjahr" versehen wurden (vgl. Abb. 2).

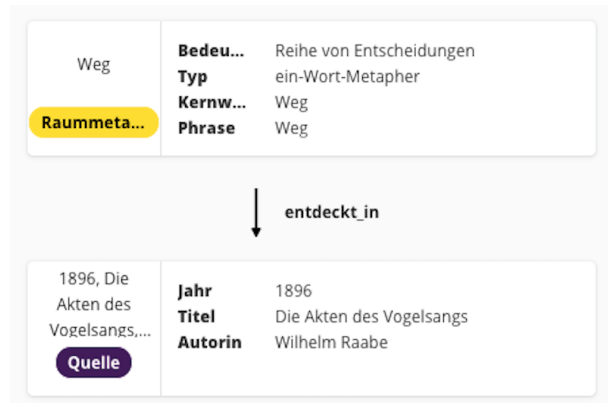


Abb. 2: Beispiel der Verbindung einer Raummetapher mit vier Propertyts mit einer Quelle mit drei Propertyts

Durch die Anreicherung mit Propertyts kann die Datenbank auf vielfältige Weise durchsucht werden. Wenn sowohl das Wortmaterial als auch die Bedeutung einer Metapher extrem ähnlich waren, wie z.B. der Fall bei "in schlechte Hände geraten" und "in schlechten Händen sein", wurden die Phrasen mit einem zweiten Typ von Relation untereinander verbunden (Relationstyp *Variation*). Insgesamt sind in *laRa* rund 800 Raummetaphern mit über 1.000 Relationen verzeichnet, die aus 80 Romanen stammen, die die Zeitspanne vom 18.–21. Jahrhundert regelmäßig abbilden, d.h. aus jedem Jahrhundert sind 20 Romane in den Aufbau der Datenbank einbezogen worden. Der Vorteil dieses systematischen Aufbaus ist, dass sich eine zwar kleine aber gleichmäßige Datenbasis ergibt, die den diachronen Vergleich der Nutzung von Raummetaphern unterstützt. Die in *laRa* verzeichneten Raummetaphern können als Netzwerk wie folgt visualisiert werden:

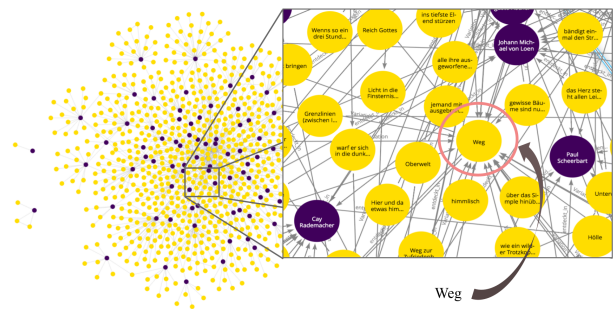


Abb. 3: Netzwerk der literarischen Raummetaphern mit Fokus auf "Weg" (helle Knoten stehen für Raummetaphern, dunkle für Quellen)

Das Netzwerk der Raummetaphern in Abb. 3 zeigt, dass es nur wenige zentrale Metaphern gibt, d.h. Metaphern, die sowohl viele Varianten aufweisen als auch in vielen Texten vorkommen. Viele Texte bilden mit ihren Raummetaphern eigene Cluster, die vielfach nur über Variationen mit anderen Metaphern, also nur indirekt mit anderen Texten verknüpft sind. Manche Text-Metaphern-Cluster sind gar nicht mit dem Hauptnetzwerk verbunden. Noch deutlicher wird das Gefüge, wenn eine Raummetapher einzeln betrachtet wird. Die Wahl fiel auf die Ein-Wort-Metapher "Weg", die keine der zentralsten Metaphern ist, sondern einen mittleren Vernetzungsgrad aufweist. Abb. 4 zeigt diese Raummetapher mit ihren Varianten und den Erzähltexten, in denen sie vorkommen:



Abb. 4: Raummetapher "Weg" und Varianten in diachronen Vergleich

Die *laRa*-Abfrage der Metaphern mit dem Kernwort "Weg" zeigt sowohl den Variantenreichtum als auch das Vorkommen von Weg-Metaphern in literarischen Texten im diachronen Zeitverlauf. Der Variantenreichtum ist mit neun Varianten im Teilkorpus des 18. Jahrhunderts am höchsten. Im 19. Jahrhundert sind "Weg"-Raummetaphern in vergleichsweise vielen Texten zu finden. Abgesehen von der Ein-Wort-Metapher "Weg" kommt kaum eine Variante jahrhundert- oder auch nur textübergreifend vor. Nur die Metapher des "Lebensweges" kann über die Varianten "Lebenspfad" und "Pfad des Lebens" sowohl dem 18. als auch dem 19. und 20. Jahrhundert zugeordnet werden. Diese Charakteristik von Raummetaphern bleibt auch über Abfragen anderer zentraler Knotenpunkte stabil. Immer sind Häufigkeit und Variantenreichtum in den frühen Jahrhunderten am höchsten. Selten wirken Metaphern jahrhundertübergreifend. Meistens findet sich eine Variante einer Metapher nur in einem einzelnen Text. Die einzigen über alle vier Jahrhunderte wirkenden Raummetaphern sind generische Ein-Wort-Metaphern wie "Welt", "Weg" oder "Himmel".

Zusammenführung von Machine-Learning-Tests und Datenbank-Abfragen

Die Analysen, die mit Hilfe der Datenbank literarischer Raummetaphern *laRa* durchgeführt wurden, bieten eine grundlegende Erklärung für das Scheitern des Machine-Learning-Prozesses: Literarische Raummetaphern sind sehr spezifisch für begrenzte sprach- und literaturgeschichtliche Phasen. Oft führen Autor*innen eigene Varianten ein oder entwickeln komplexe metaphorische Raumbilder. Es reicht darum nicht aus, das Trainingsmaterial in kleinerem Umfang auszuweiten. Ist zufällig innerhalb eines neuen Abschnittes der Trainingsdaten dieselbe oder eine ähnliche Raummetapher vorhanden wie im Testtext, so kann die Erkennung zwar sprunghaft ansteigen, schon das Hinzufügen eines weiteren Datensatzes kann die Erkennung aber wieder absinken, wenn hier eine ähnliche Phrase in einem nicht-metaphorischen Kontext steht.

Mit der Datenbank *laRa* ist ein erster Schritt zu einer systematischen Erfassung literarischer Raummetaphern getan. Durch den gleichmäßigen Aufbau der Datenbasis gibt *laRa* schon jetzt gute Hinweise auf Konstruktion und Funktionsweise literarischer Raummetaphern. Sie kann dazu genutzt werden, andere methodische Zugänge einer Cross-Validierung (oder Cross-Falsifizierung) zu unterziehen. *laRa* kann außerdem die literaturwissenschaftliche Analyse einzelner Begriffe oder Raumsymbole wie z.B. dem des Weges gut ergänzen.

Fußnoten

1. Beim F1-Score handelt es sich um ein in der Computerlinguistik gängiges Maß, um die Erkennungsgenauigkeit automatischer Klassifikation zu erfassen. Mathematisch kombiniert der F1-Score die beiden Werte Precision (wie viele der klassifizierten Ausdrücke wurden richtig annotiert?) und Recall (wie viele der im Text befindlichen relevanten Ausdrücke wurden annotiert?).
2. Eine tabellarische Auflistung der zum Training genutzten Romane und der Quellen ihrer txt-Versionen findet sich in der Datei *Texte Trainingsdaten Raum-Classifizier.xlsx* im GitHub-Repository zum Raum-Classifizier (vgl. Schumacher 2021a).
3. Außerdem zeigt Herrmann (2018), dass in Erzählanfängen zuverlässig Metaphern vorkommen. In dem von ihr untersuchten Erzählanfangskorpus werden in keinem Text weniger als 5,79% der Wörter metaphorisch gebraucht, im Durchschnitt sind es 14,1% (vgl. Herrmann 2018: 188).
4. Da in der Trainingsphase die Methode der manuellen Annotation genutzt wurde, konnten historische Varianzen der Schreibweise mit berücksichtigt werden, wenn diese in den generell meist normalisierten Texten aus dem TextGrid-Repository noch vorhanden waren.
5. Soweit rechtlich möglich wurden die annotierten Trainings- und Testdaten in einem GitHub-Repository zugänglich gemacht (vgl. Schumacher 2021a).

Bibliographie

Do Dinh, Erik-Lân / Malte Gerloff / Alexandra Núñez (2017): „Metaphern digital – Auf dem Weg von der Annotation

zur automatischen Detektion“ in: *Modellierung - Vernetzung – Visualisierung: Die Digital Humanities als fächerübergreifendes Forschungsparadigma*. 3. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHD 2016), Leipzig, Mai 2. <https://doi.org/10.5281/zenodo.4645046>.

Blumenberg, Hans (1971): „Beobachtungen an Metaphern“ in: *Archiv Für Begriffsgeschichte* 15: 161-214 <http://www.jstor.org/stable/24358391> [letzter Zugriff 8. Juli 2021].

Dunn, Jonathan (2013): „Evaluating the Premises and Results of Four Metaphor Identification Systems“ in: *Lecture Notes in Computer Science* 471–86.

Finkel, Jenny Rose / Trond Grenager / Christopher Manning (2005): „Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling“ in: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* 363–370.

Gehring, Petra, und Iryna Gurevych (2014): „Suchen als Methode? Zu einigen Problemen digitaler Metapherndetektion“ in: *Journal Phänomenologie - Schwerpunktthema: Metaphern als strenge Wissenschaft* 41: 99-109.

Herrmann, J Berenike (2018): „Eine quantitative Metaphernanalyse deutschsprachiger Erzählanfänge zwischen 1880 und 1926“ in: Köppe, Tilmann / Singer, Rüdiger (eds.): *Show, don't tell: Konzepte und Strategien anschaulichen Erzählens* 167-212.

Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank (2015): „Automatische Erkennung von Figuren in Deutschsprachigen Romanen“ in: *Von Daten Zu Erkenntnissen* 1–6 <http://gams.uni-graz.at/o:dh-d2015.abstracts-vortraege> [letzter Zugriff 8. Juli 2021].

Lakoff, George / Johnson, Mark (1998): *Leben in Metaphern Konstruktion und Gebrauch von Sprachbildern*. Heidelberg: Carl-Auer-Verlag.

Majoros, Krisztián (2013): „Metapher und Kookkurrenz. Eine alternative ‚Trichter‘-Methode zur korpus- basierten Untersuchung metaphorischer Ausdrücke in öffentlich zugänglichen elektronischen Zeitungskorpora am Beispiel der Wissenschaftsberichterstattung im Bereich der Zellbiologie“ in *Sprachtheorie und germanistische Linguistik* 23: 65–110.

Manning, Christopher D. / Surdeanu, Mihai / Bauer, John / Finkel, Jenny / Bethard, Steven J. / McClosky, David (2014): „The Stanford CoreNLP Natural Language Processing Toolkit“ in: *Association for Computational Linguistics (ACL) System Demonstrations* 55–60. <http://www.aclweb.org/anthology/P/P14/P14-5010> [letzter Zugriff 8. Juli 2021].

Mason, Zachary J. (2004): „CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System“ in: *Computational Linguistics* 30: 23–44. <https://doi.org/10.1162/089120104773633376>.

Miller, Norbert (1965): *Romananfänge*. Berlin: Verlag Literarisches Colloquium.

Retsch, Annette (2000): *Paratext und Textanfang*. Würzburg: Königshausen & Neumann.

Richardson, Brian (2008) *Narrative Beginnings: Theories and Practices*. Lincoln: University of Nebraska Press.

Schumacher, Mareike (2018): „Named Entity Recognition (NER) | ForTEXT“ in: Gius, Evelyn / Meister, Jan Christoph / Schumacher, Mareike / Gerstorfer, Dominik / Meister, Malte / Blaess, Sandra / Flüh, Marie / Horstmann, Jan / Jacke, Janina (eds.): *ForTEXT - Literatur digital erforschen*. <https://fortext.net/routinen/methoden/named-entity-recognition-ner> [letzter Zugriff 8. Juli 2021]

Schumacher, Mareike (2021a): *Raum-Classifizier (kompatibel Mit StanfordNER)*. Zenodo: doi:10.5281/zenodo.4992662.

Schumacher, Mareike (2021b): *laRa - Die Datenbank für literarische Raummetaphern*. Zenodo. doi:10.5281/zenodo.4987844.

Schumacher, Mareike / Flüh, Marie (2020): "m*w Figurengender zwischen Stereotypisierung und Literarischen und theoretischen Spielräumen. Genderstereotype und -bewertungen in der Literatur des 19. Jahrhunderts". Zenodo. doi:10.5281/zenodo.4621892.

Schumacher, Mareike (im Erscheinen): *Orte und Räume im Roman. Ein Beitrag zur digitalen Literaturwissenschaft*.

Steen, Gerard / Dorst, Lettie / Herrmann, Berenike / Kaal, Anna / Krennmayr, Tina / Pasma, Trijntje (2010): *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam: John Benjamins Publishing. <https://doi.org/10.1075/celcr.14>.

Sutton, Charles / McCallum, Andrew (2010): *An Introduction to Conditional Random Fields*. <http://arxiv.org/abs/1011.4088> [letzter Zugriff 8. Juli 2021].

Thaller, Manfred (2021): "Über Metaphern und die Voraussetzungen für ihre Verwendung in der Informationstechnologie". In: Flüh, Marie, Horstmann, Jan, Jacke, Janina und Schumacher, Mareike (Hrsg.): *Toward Undogmatic Reading. Narratology, Digital Humanities and Beyond*. Hamburg: Hamburg University Press.

Wenz, Karin (1997): *Raum, Raumsprache und Sprachraum. Zur Textsemiotik der Raumbeschreibung*. Tübingen: Narr.

Doctoral Consortium

Adnominale Possession in einem Bibel-Parallelkorpus

Fleischmann, Florian

Florian.Fleischmann@itg.uni-muenchen.de
LMU München, Germany

Empirische Sprachwissenschaft erarbeitet Erkenntnisse auf Basis sprachlichen Original-Materials. Für Untersuchungen über Unterschiede zwischen verschiedenen diachronen Stufen oder dialektalen Varietäten einer Sprache muss sprachliches Material aus diversen Quellen miteinander verglichen werden. Textsorte, Inhalt und Art der Quellen addieren hierbei weitere Variablen zur Untersuchung. Eine Möglichkeit, diesem Umstand zu begegnen, bieten Parallelkorpora. Sie haben den Vorteil, dass hier nur ein einziger Text in verschiedenen Sprachstufen bzw. -varietäten verglichen wird. Varianz in der Textsorte scheidet so als Erklärung für etwaige beobachtete linguistische Unterschiede aus. Um ein umfassendes Parallelkorpus zu erstellen, muss der untersuchte Text in möglichst vielen diachronen und/oder dialektalen Fassungen vorliegen. Einen Text, für den dies zutrifft, stellt die Bibel dar. Bibelübersetzungen ins Deutsche existieren bereits seit dem Althochdeutschen (z.B. Evangelienharmonie des Tatian, ca. 830 (Sievers 1872)) und sind lückenlos bis zum Neuhochdeutschen überliefert. Auch dialektal besitzen sie synchron eine weite Verbreitung. Es existieren bspw. (Teil-)Übersetzungen ins Hessische (Mieth 2011), Kärntnerische (Bünker 2007), Niederdeutsche (Jessen 1984), Pennsylvania Dutch (o. A. 2002), Schwäbische (Paul 1997), Walliserdeutsche (Theler 2011) oder Zürichdeutsche (Weber 2011). Ein weiterer Vorteil der Bibel liegt darin, dass hier eine möglichst textgetreue Wiedergabe im Sinne des Übersetzers liegt. Anpassungen an die individuelle Varietät erfolgen möglichst behutsam und konservativ, insbesondere was Syntax und Lexik angeht. Das Erstellen eines derartigen Parallelkorpus ist Kern meiner Promotionsschrift.

Besonderes Augenmerk liegt dabei auf einer nachhaltigen Aufbereitung der zugrundeliegenden Daten im Sinne der FAIR-Prinzipien. In einem ersten Schritt müssen die Bibeln digitalisiert werden. In vielen Fällen ist dies bereits durch Bibliotheken geschehen, meistens jedoch nur als Bild-PDFs. Zur weiteren Verarbeitung werden diese mittels automatischer Texterkennung (OCR) in maschinenlesbare Form gebracht. Bei älteren Texten stellt dies aufgrund der verwendeten gebrochenen Schriftarten (wie Fraktur) ein Problem dar. Vortrainierte Texterkennungsmodelle versagen hier oft (Baierer 2020, Springmann 2017: 3), so dass selbstständig neuronale Netze zur Erkennung trainiert werden müssen. Mit OCR4all (<https://www.ocr4all.org>) steht ein leistungsfähiges Software-Paket hierfür zur Verfügung. OCR4all basiert auf dem OCRopus-Derivat Calamari, das im Vergleich zu anderen OCR-Lösungen wie OCRopy, Tesseract oder OCRopus die besten Erkennungsraten zeigt (Wick 2020). OCR4all vereint weiterhin die Schritte von Pre-Processing, Character Recognition und Post-Processing in einem Tool und sorgt so für einen effizienten Arbeitsablauf. Bei Pilotversuchen ließen sich für ausgewählte historische Bibeln Erkennungsraten von 98 % und mehr erzielen. Mit OCR-D ist eine vergleichbare Lösung in Entwicklung, zumindest momentan liegen dessen Erkennungsraten jedoch noch niedriger (Baierer 2020).

Die Ausgabe liegt zunächst als reiner Text vor. Diese werden in eine SQL-Datenbank importiert und nach ihren Bibelversen anno-

tiert. Auf diese Weise lässt sich in einem späteren Schritt eine einfache Weboberfläche entwerfen, die eine alignierte Darstellung von Bibelversen zulässt. Dadurch können auch Forschungszweige außerhalb der Sprachwissenschaften (im Rahmen von UrhG § 60d) auf das Parallelkorpus zugreifen. Im Kern soll das Parallelkorpus als Grundlage für linguistische Fragestellungen dienen. Hierfür sind weitere Verarbeitungsschritte notwendig, um die Daten entsprechend aufzubereiten. Es ist angezeigt, die Annotation der Texte um POS-Tags zu erweitern. Für Nicht-Standard Varietäten muss hierfür wieder auf das Training eigenständiger Tagger durch neuronale Netze zurückgegriffen werden. Obwohl diese sich im NLP bewährt haben, sind die Voraussetzungen für einen erfolgreichen Einsatz im linguistischen Kontext nicht vollständig klar. Im Rahmen dieser Arbeit sollen Einflussvariablen identifiziert und deren Auswirkung auf die Arbeit mit neuronalen Netzen vermessen werden. Ziel der Untersuchung ist es deshalb weiterhin, Verfahren zu verbessern und Parameter einzugrenzen, wie Neuronale Netze optimal zur Erkennung sprachlicher Strukturen genutzt werden können.

Ein möglicher beispielhafter Untersuchungsgegenstand ist die Possession, eine grundlegende, sprachübergreifende Kategorie, um Besitzverhältnisse auszudrücken. Possession kann durch unterschiedliche Konstruktionen realisiert werden. Diese können in Konkurrenz stehen oder parallel existieren. Eine Unterkategorie dieser Möglichkeiten bilden die adnominalen Konstruktionen. Im Deutschen umfassen diese (vgl. Kasper 2017: 300):

- possessiver Genitiv: Marias Kind, das Kind Marias .
- possessiver Dativ: Maria ihr Kind .
- von -Konstruktion: das Kind von Maria .
- Possessivpronomen: ihr Kind .

Es liegen zahlreiche Arbeiten zur Possession allgemein (Seiler 1983; Heine 1997; Stolz et al. 2008; McGregor 2009; Börjars et al. 2013) oder zu Teilaspekten vor: ihre Verwendung in einzelnen Dialekten des Deutschen (z.B. für Hessen: Kasper 2017), die Konkurrenz zwischen Genitiv und von -Konstruktionen in Abhängigkeit der Textsorte (Lang 2018), den frühkindlichen Erwerb possessiver Phrasen (Eisenbeiß et al. 2009) oder Possession im Sprachvergleich – beispielsweise Deutsch und Koreanisch (Shin 2004). Nicht vorhanden ist hingegen eine longitudinale Studie, die die gesamte deutsche Sprachgeschichte abdeckt. Und obwohl für Einzeldialekte Untersuchungen zur Possession existieren, fehlt eine umfassende empirische Studie, die eine größere Anzahl an dialektalen Varietäten des Deutschen abdeckt. Die adnominale Possession eignet sich deshalb besonders, um anhand des geschaffenen Parallelkorpus hinsichtlich ihrer diachronen Entwicklung und der Realisierung in dialektalen Varietäten untersucht zu werden.

Bibliographie

- Baierer, Konstantin et al.** (2020): "OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative", in: *Bibliothek Forschung und Praxis* 44/2: 218-230.
- Börjars, Kersti / Denison, David / Scott, Alan**, (eds.) (2013): *Morphosyntactic Categories and the Expression of Possession*. Amsterdam, Philadelphia: John Benjamins.
- Bünker, Michael / Lager, Sepp** (Übers.) (2007): *Es wead ana kemmen*. Das Markusevangelium auf Kärntnerisch. Übersetzt von Michael Bünker und Sepp Lager. Klagenfurt: Heyn.
- Eisenbeiß, Sonja / Matsuo, Ayumi / Sonnenstuhl, Ingrid** (2009): „Learning to encode possession“. In: McGregor, William (ed.): *The expression of possession*. Berlin [u.a.]: de Gruyter, S. 143–213.

Heine, Bernd (1997): *Possession. Cognitive sources, forces, and grammaticalization*. Cambridge: Cambridge University Press.

Jessen, Johannes (Übers.) (1984): *Das Ole und das Nie Testament in unser Moderspraak*. Übersetzt von Johannes Jessen. Göttingen: Vandenhoeck & Ruprecht.

Kasper, Simon (2017): „Adnominale Possession“. In: *SyHD-Atlas*.

Lang, Kristine (2018): *Possession. Empirisch-funktionale Untersuchungen zu Genitivattribut und Präpositionalphrase mit von*. München: Iudicium.

McGregor, William (2009): *The expression of possession*. Berlin [u.a.]: de Gruyter.

Mieth, Klemens (Übers.) (2011): *Das Neue Testament uff Hesisch*. Übersetzt von Klemens Mieth. Norderstedt: Books on Demand GmbH.

o. A. (2002): *Es Nei Teshtament. Mitt Di Psaltah un Shpricha*. South Holland, IL: The Bible League.

Paul, Rudolf (Übers.) (1997): *D Bibel für Schwoba. s Matthäus-Evangelium*. Ens Schwäbische übersetzt vom Pfarrer Rudolf Paul. Tübingen: Silberburg.

Shin, Yong-Min (2004): *Possession und Partizipantenrelation. Eine funktional-typologische Studie zur Possession und ihren semantischen Rollen am Beispiel des Deutschen und Koreanischen*. Bochum: Brockmeyer.

Seiler, Hansjakob (1983): *Possession as an Operational Dimension of Language*. Tübingen: Narr.

Sievers, Eduard (ed.) (1872): *Tatian. Lateinisch und altdeutsch mit ausführlichem Glossar*. Paderborn: Schöningh.

Springmann, Uwe / Lüdeling, Anke (2017): „OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus“, in: *Digital Humanities Quarterly* 11/2.

Stolz, Thomas / Kettler, Sonja / Stroh, Cornelia / Urdze, Aina (2008): *Split possession. An areal-linguistic study of the alienability correlation and related phenomena in the languages in Europe*. Amsterdam, Philadelphia: Benjamins.

Theler, Hubert (Übers.) (2011): *Ds Niww Teschtamänt uf Waliseritsch*. Übersetzt von Hubert Theler. Visp: Rotten.

Weber, Emil (Übers.) (2011): *S Nöi Teschtamänt Züritüütsch*. Us em Griechische. Übersetzt von Emil Weber. Zürich: Jordan.

Wick, Christoph / Reul, Christian / Puppe, Frank (2020): „Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition“, in: *Digital Humanities Quarterly* 14/2.

Das mediale und politische Framing von Extremismusformen im Zeitraum der Jahre 1999 – 2021

Feldmüller, Tim

tim.feldmueller@uni-leipzig.de
Universität Leipzig, Germany

Das Dissertationsvorhaben widmet sich der linguistischen, korpusbasierten Analyse des bundesdeutschen Extremismuskurses der Jahre 1999 – 2021.

Unter inhaltlichem Gesichtspunkt untersucht das Projekt, wie Extremismus(-varianten)-Frames in Deutschland auf der medialen und politischen Diskursebene verhandelt werden. Aus einer linguistisch-kulturwissenschaftlichen Perspektive ist dabei insbesondere von Interesse, wie einzelne epistemologische Entitäten (Frames) von den Diskursteilnehmenden sprachlich (re-)produziert und variiert (geframet) werden. Dabei soll durch einen Abgleich des Framings auf zwei unterschiedlichen Diskursebenen (Medien, Politik), zu unterschiedlichen Zeitpunkten im Analysezeitraum sowie zwischen einzelnen Akteur:innen innerhalb der Diskursebenen (Parteien, Zeitungen) sprachliches Handeln in Bereichen sichtbar werden, die von zentraler Bedeutung für gesellschaftliche Meinungsbildungsprozesse sind und für die bereits strategische Neuausrichtungen des Extremismusbegriffs beschrieben worden sind (Ackermann et al. 2015). Für die Untersuchung wurde ein Korpus erstellt, das den medialen Diskurs durch den Online-Artikelbestand der Zeitungen *Welt*, *Spiegel* und *Taz* des Zeitraumes 1999 – 08/2021 abbildet (pro Zeitung ca. 425 Mio. Token), während die politische Diskursebene in Form von Parlamentsdebatten des gleichen Zeitraums (Blaette 2020, Deutscher Bundestag 2021) repräsentiert ist.

Als Deutungsrahmen ermöglichen Frames eine sinnstiftende Einordnung sprachlicher Daten. Semantische Frames im Sinne Busses (2012) sind – u.a. Fillmore, Minsky und Barsalou folgend – in einer Slot-Filler-Struktur organisiert. Ein Extremismus-Frame könnte etwa Slots für Akteur:innen, Handlungen oder Ziele der Handlungen aufweisen. Bei jeder Instanziierung des Frames können diese mit Werten gefüllt werden, um z.B. ein konkretes extremistisches Gewaltereignis zu beschreiben. Die damit einhergehende potenziell strategische Auswahl der im Text realisierten Slots und Filler fasse ich in Anlehnung an Klein (2018) und Ziem et al. (2018) als *Framing*.

In methodischer Hinsicht entwickelt das Projekt erstens ein Verfahren zur korpusbasierten Frame-Identifizierung und versucht zweitens, zwei scheinbar widersprüchliche Paradigmen der korpusbasierten Diskursforschung zu verbinden: die Analyse thematischer Diskurse und die datengeleitete, d.h. *corpus driven* (Tognini-Bonelli 2001), Diskursanalyse. Letztere reklamiert für sich ein hohes Maß an Unvoreingenommenheit, steht einer Fokussierung auf thematisch definierte Diskurskorpora jedoch eher ablehnend gegenüber (Bubenhof 2009: 36; Scharloth et al. 2013). Eine Vereinbarung der beiden Paradigmen kann m.E. gelingen, indem der Prozess der Korpuszusammenstellung von einer Auswahl relevanter Texte durch die Forschungsperson (Busse & Teubert 1994) zu einer algorithmisch unterstützten, möglichst induktiven und datengeleiteten Selektion thematisch passender Texte weiterentwickelt wird. Der analytische Zugang zum Diskurs ist dabei keineswegs frei von Vorprägungen durch die Forschungsperson – sei es bei der Korpus- oder Methodenwahl. Er kann jedoch auf ungleich größere und somit eher repräsentative Diskurskorpora erweitert werden. Die thematische Reduktion des Korpus trägt einerseits der semantischen Kontextgebundenheit einzelner Wortformen Rechnung – so bedeuten etwa ‚links‘ und ‚rechts‘ in der Sportberichterstattung etwas anderes als im Extremismuskurs – andererseits werden datengeleitete Zugänge wie Keyword-Analysen so erst ermöglicht.

Einzelne Lösungsansätze liegen hier bereits vor; so wird als Kriterium für die Reduktion eines themenübergreifenden Korpus auf ein thematisches Diskurskorpus häufig das Vorkommen eines oder mehrerer repräsentativer Schlagwörter in den Texten angesetzt (vgl. aus dem Bereich der Frame-Forschung etwa Baker et

al. 2020; Storjohann & Schröter 2011; Ziem et al. 2018). Dass eine Vielzahl sprachlicher Muster einzelne Deutungsrahmen aufrufen und perspektivieren, ist jedoch eine Kernannahme der linguistischen Frame-Theorie (vgl. etwa die Rolle der *Lexical Units* in Fillmores FrameNet, Ruppenhofer et al. 2016). Ähnlich haben Kozłowski et al. (2019) unter Anwendung von *word embeddings* gezeigt, dass kulturelle Kategorien über die Zeit stabil bleiben, die assoziierten Wortvektoren jedoch „in constant flux“ (S. 929) sind. Während einzelne Arbeiten auch dies berücksichtigen, indem erst ganze Begriffsfelder empirisch ermittelt und dann in einem nicht-thematischen Korpus abgefragt werden (Czulo et al. 2020), bleiben die diskursanalytischen Potenziale neuerer Verfahren des *Natural Language Processing* weitgehend ungenutzt. Hier können insbesondere unüberwachte Algorithmen wie *Topic Modeling* (vgl. zu Potenzialen für die Diskursanalyse Murakami et al. 2017; kritisch Brookes & McEnery 2019) oder *Word Embeddings* (vgl. zu Anwendungen in diskursanalytischer Forschung etwa Bubenhofer et al. 2020; Kozłowski et al. 2019) einerseits eine Eingrenzung des im Rahmen der Dissertation verwendeten multithematischen Korpus, andererseits einen Zugang zu einzelnen Framings in verschiedenen Diskursbereichen ermöglichen. Auf vorliegende Forschung zur *Semantic Frame Induction* (vgl. etwa die im Kontext von QasemiZadeh et al. 2019 veröffentlichten Konferenzbeiträge) kann in dieser Hinsicht nur bedingt aufgebaut werden, da diese zumeist bereits vorliegende FrameNet-Frames zu identifizieren versucht, anstatt epistemologisch komplexere Deutungsrahmen zu rekonstruieren (vgl. zu einer Kritik an FrameNet aus diskurslinguistischer Sicht Busse 2012: 210-213).

Mit dem hier skizzierten Dissertationsprojekt wird erstmals eine breit angelegte Untersuchung des bundesdeutschen Extremismuskurses seit der Jahrtausendwende geleistet. Einzelne als Frames modellierte Wissensbestände dieses Diskurses werden – einschließlich der sprachlichen Praktiken, die sie formen – sichtbar. Die methodische Verankerung in den Digital Humanities ermöglicht dabei neben einer unvoreingenommenen und besonders repräsentativen Modellierung des Forschungsgegenstandes auch einen völlig neuen Blick auf die „complex geometry of culture“ (Kozłowski et al. 2019: 931).

Bibliographie

- Ackermann, Jan / Behne, Katharina / Buchta, Felix / Drobot, Marc / Knopp, Philipp (2015): *Metamorphosen des Extremismusbegriffes*. Diskursanalytische Untersuchungen zur Dynamik einer funktionalen Unzulänglichkeit. Wiesbaden: Springer VS.
- Baker, Paul / Brookes, Gavin / Atanasova, Dimitrinka / Flint, Stuart W. (2020): "Changing frames of obesity in the UK press 2008–2017", in: *Social Science & Medicine* 264.
- Blaette, Andreas (2020): *GermaParl*. Linguistically Annotated and Indexed Corpus of Plenary Protocols of the German Bundestag. CWB corpus version 1.0.6. <https://doi.org/10.5281/zenodo.3735141>.
- Brookes, Gavin / McEnery, Tony (2019): "The utility of topic modelling for discourse studies: A critical evaluation", in: *Discourse Studies* 21 (1): 3-21.
- Bubenhofer, Noah (2009): *Sprachgebrauchsmuster*. Korpuslinguistik als Methode der Diskurs- und Kulturalanalyse. Berlin: De Gruyter.
- Bubenhofer, Noah / Knuchel, Daniel / Sutter, Livia / Keltenberger, Maaïke / Bodenmann, Niclas (2020): „Von Grenzen und Welten: Eine korpuspragmatische COVID-19-Diskursanalyse“, in: *Aptum* 16 (2/3): 156-165.
- Busse, Dietrich (2012): *Frame-Semantik*. Ein Kompendium. Berlin: De Gruyter.
- Busse, Dietrich / Teubert, Wolfgang (1994): "Ist Diskurs ein sprachwissenschaftliches Objekt?", in: Busse, Dietrich / Hermanns, Fritz / Teubert, Wolfgang (eds.): *Begriffsgeschichte und Diskursgeschichte*. Opladen: Westdeutscher Verlag: 10-28.
- Czulo, Oliver / Nyhuis, Dominic / Weyell, Adam (2020): "Der Einfluss extremistischer Gewaltereignisse auf das Framing von Extremismen auf SPIEGEL Online". In: *Journal für Medienlinguistik* 3 (1): 14-45.
- Deutscher Bundestag (2021): *Open Data*. <https://www.bundestag.de/services/opendata> [letzter Zugriff 01. Dezember 2021]
- Klein, Josef (2018): "Frame und Framing: Frametheoretische Konsequenzen aus Praxis und Analyse strategischen politischen Framings", in: Ziem, Alexander / Inderelst, Lars / Wulf, Detmer (eds.): *Frames interdisziplinär*. Modelle, Anwendungsfelder, Methoden. Düsseldorf: Düsseldorf University Press: 289-330.
- Kozłowski, Austin / Taddy, Matt / Evans, James A. (2019): "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings", in: *American Sociological Review*. 10.1177/0003122419877135.
- Murakami, Akira / Thompson, Paul / Hunston, Susan / Vajn, Dominik (2017): "What is this corpus about?: Using topic modelling to explore a specialised corpus", in: *Corpora*. 10.3366/cor.2017.0118.
- QasemiZadeh, Behrang / Petruck, Miriam R. L. / Stodden, Regina / Kallmeyer, Laura / Candito, Marie (2019): "SemEval-2019 Task 2: Unsupervised Lexical Frame Induction", in: *Proceedings of the 13th International Workshop on Semantic Evaluation* 16-30.
- Ruppenhofer, Josef / Ellsworth, Michael / Petruck, Miriam R. L. / Johnson, Christopher R. / Baker, Collin F. / Scheffczyk, Jan (2016): *FrameNet II*. Extended Theory and Practice <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf> [letzter Zugriff 01. Dezember 2021].
- Scharloth, Joachim / Eugster, David / Bubenhofer, Noah (2013): "Das Wuchern der Rhizome: Linguistische Diskursanalyse und Data-driven Turn", in: Busse, Dietrich / Teubert, Wolfgang (eds.): *Linguistische Diskursanalyse*. Neue Perspektiven. Wiesbaden: Springer Fachmedien Wiesbaden: 345-380.
- Storjohann, Petra / Schröter, Melani (2011): "Die Ordnung des öffentlichen Diskurses der Wirtschaftskrise und die (Un-) Ordnung des Ausgeblendeten", in: *Aptum* 7 (1): 32-53.
- Tognini-Bonelli, Elena (2001): *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Ziem, Alexander / Pentzold, Christian / Fraas, Claudia (2018): "Medien-Frames als semantische Frames: Aspekte ihrer methodischen und analytischen Verschränkung am Beispiel der ‚Snowden-Affäre‘", in: Ziem, Alexander / Inderelst, Lars / Wulf, Detmer (eds.): *Frames interdisziplinär*. Modelle, Anwendungsfelder, Methoden. Düsseldorf: Düsseldorf University Press: 155-182.

Digitale Methodenkritik

Die Integration computergestützter Textanalyseverfahren in den Werkzeugkasten der Historiker:innen

Althage, Melanie

melanie.althage@hu-berlin.de

Humboldt-Universität zu Berlin, Germany

Das vorliegende Proposal basiert auf dem Dissertationsprojekt “Mining the Historian’s Web – Methodenkritische Reflexion quantitativer Verfahren zur Analyse genuin digitaler Quellen am Beispiel der historischen Fachkommunikation”. Im Zentrum steht die Untersuchung der Adaptierbarkeit etablierter Textanalysemethoden der Digital Humanities und Computerwissenschaften für historische Quellen und Forschungsfragen sowie die Entwicklung von Strategien zur Integration dieser Methoden in den Werkzeugkasten der Historiker:innen.

Die digitale Durchdringung der Alltags- und Arbeitswelt hat die Art und Weise des Forschens in den Geistes- beziehungsweise Geschichtswissenschaften modifiziert: Nicht nur das *womit*, sondern auch das *worüber* geforscht wird, ist zunehmend geprägt von “Digitalität”, also dem Umstand, dass die Quellen als maschinell (re-)produzierte wie verarbeitbare Daten vorliegen, d.h. als “in Zahlen gefasste Informationen” (Emich 2019: 213), die abstrakte Repräsentationen und Rekonstruktionen von Objekten, Konzepten oder Ereignissen darstellen (Schöch 2013; Drucker 2011; Owens 2011). Die (kommunikations-)technologischen Entwicklungen seit Mitte des 20. Jahrhunderts begünstigten einerseits die von umfangreichen Digitalisierungsprojekten angestoßene freiere und ortsunabhängige Zugänglichkeit zum digitalisierten kulturellen Erbe über bspw. Archiv- oder Bibliotheksdatenbanken im Web.¹ Andererseits führten sie zu einem grundlegenden Wandel der gebräuchlichen Kommunikationsverfahren sowie der Mediennutzung im Beruflichen wie im Privaten. Dadurch entstanden neuartige, multimediale und genuin digitale Quellengattungen, die nicht zuletzt für die Zeitgeschichte von essenzieller Bedeutung sind (u.a. Milligan 2019; Haber 2012; Patel 2011).

Digitale Quellen, gleich ob es sich um *digitized*, *born-digital* oder *reborn-digital* handelt (nach Brügger 2012), sind und werden aus dem geschichtswissenschaftlichen Forschungs- und Erkenntnisbildungsprozess nicht mehr wegzudenken (sein). Entsprechend erfährt die Frage zunehmend Aufmerksamkeit, ob und inwiefern die klassische Quellenkritik auf diese neuen Quellentypen angewendet werden kann (u.a. Föhr 2019; Margulies 2009).² Indes sind Untersuchungen, die sich damit auseinandersetzen, was es heißt, mit digitalen und insbesondere *genuin digitalen* Objekten als Forschungsressource zu arbeiten, bislang sowohl in praktischer als auch in theoretischer Perspektive in den Geschichtswissenschaften noch unterrepräsentiert (Aufschläge: Hiltmann et al. 2021; Fickers 2020; Milligan 2019; Schreiber 2012).

Eine wesentliche Säule des Dissertationsprojekts ist daher die Untersuchung der epistemologischen Konsequenzen der computergestützten Analyse textbasierter, originär digitaler Quellen für den historischen Forschungsprozess: Was heißt es, Text zunächst als Daten und erst in zweiter Instanz als Bedeutungsträger zu ver-

arbeiten? Welche epistemologischen Veränderungen gehen mit der Untersuchung digitaler Objekte einher, wenn sie nicht lediglich als Surrogat für ein physisches Objekt verstanden werden, sondern die Digitalität als ihre substantielle Eigenschaft berücksichtigt wird? Dies erfolgt am Beispiel der digitalen historischen Fachkommunikation und damit anhand genuin digitaler Quellen.

Ihre spezifischen Eigenschaften als maschinell prozessierbare Daten legen den Einsatz digitaler Werkzeuge und Methoden nahe. In den Digital Humanities haben sich dementsprechend eine Reihe von digitalen Methoden zur Auswertung von Textdaten etabliert wie bspw. *Topic Modeling*, die auch in den Geschichtswissenschaften nachgenutzt werden und in nutzerfreundlichen Werkzeugen wie *DARIAH-DE TopicsExplorer* (Simmler et al. 2019) oder *Voyant Tools* (Sinclair/Rockwell 2016) implementiert sind. Solche Methoden stammen allerdings nicht selten aus fachfremden Disziplinen mit je eigenen theoretisch-methodologischen Annahmen respektive Erkenntnisinteressen, die im Kontrast zu geschichtswissenschaftlichen Forschungstraditionen stehen können; sie sind daher nicht ohne Weiteres auf historische Anwendungsfälle übertragbar. Folglich ist die Frage zu stellen, wie die Kluft zwischen historischer Fachdisziplin und fachfremder Methode identifiziert und überwunden werden kann, um sie produktiv in den Werkzeugkasten der Historiker:innen zu integrieren.

Die epistemologischen Implikationen, die mit diesen Methoden und *Tools* für die Arbeit mit historischen Quellen sowie für die Wissensproduktion zusammenhängen, werden erst in jüngerer Zeit aus spezifisch geschichtswissenschaftlicher Perspektive intensiver erforscht (Hiltmann et al. 2021; Fickers 2020; Braake et al. 2016; Wettlaufer 2016). Da eine systematische Werkzeug- und Methodenkritik, die den verantwortungsvollen Umgang mit digitalen Methoden begleiten muss, allerdings bislang für die Geschichtswissenschaften weitestgehend Desiderat geblieben ist, ist es Ziel der Dissertation, in Anlehnung an Diskussionen rund um *Tool* und *Algorithmic Criticism* (Es/Schäfer/Wieringa 2021; Dobson 2019; Ramsay 2011) zu ihrer Ausbildung einen Beitrag zu leisten. Dafür ist es notwendig, die Verfahren hinsichtlich der in sie eingeflossenen Annahmen einzuordnen, kritisch zu reflektieren und gegebenenfalls entsprechend des Erkenntnisziels anzupassen oder Alternativen aufzuzeigen; hierbei sind insbesondere die Erkenntnisgrenzen für historische Forschungsvorhaben zu dokumentieren. Dabei gilt es zu reflektieren, inwiefern die computergestützten Berechnungen den interpretativen Akt als Kerngeschäfts der Geschichtsschreibung selbst beeinflussen.

Der Vortrag wird anhand der Methode *Topic Modeling* erste anwendungsbezogene und methodenkritische Erkenntnisse zu ihrer produktiven Integration in den historischen Forschungsprozess präsentieren, und dabei speziell die für die Geschichtswissenschaften zentralen Aspekte der Historizität und Relationalität berücksichtigen.

Fußnoten

1. Die Potenziale ihrer Verknüpfung zeigen Projekte wie *Europeana* : <https://www.europeana.eu/de> [letzter Zugriff 29.11.2021].

2. Dass dies nicht allein die Geschichtswissenschaften betrifft, zeigt die interdisziplinär angelegte Workshopreihe im Rahmen der vDHd 2021 von Jonathan D. Geiger et al.: “Digitale Quellenkritik: Ein neues Kapitel”, in: *vDHd 2021* (Blog), erschienen am: 28.01.2021, URL: <https://vdhd2021.hypotheses.org/288> [letzter Zugriff 29.11.2021].

Bibliographie

Braake, Serge ter / Fokkens, Antske / Ockeloen, Niels / Son, Chantal van (2016): "Digital History: Towards New Methodologies" in: Bozic, Bojan / Mendel-Gleason, Gavin / Debruyne, Christophe / O'Sullivan, Declan (eds.): *Computational History and Data-Driven Humanities*. Cham: Springer 23–32.

Brügger, Niels (2012): "When the Present Web is Later the Past: Web Historiography, Digital History and Internet Studies", in: *HSR* 37, 4: 102–117.

Dobson, James E. (2019): *Critical Digital Humanities. The Search for a Methodology*. Urbana (Illinois): University of Illinois Press.

Drucker, Johanna (2011): "Humanities Approaches to Graphical Display", in: *DHQ* 5, 1: <http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html> [letzter Zugriff 29.11.2021].

Emich, Birgit (2019): *Geschichte der Frühen Neuzeit (1500–1800) studieren*. München 2. völlig überarb. Aufl.: UKV Verlag.

Es, Karin van / Schäfer, Mirko T. / Wieringa, Maranke (2021): "Tool Criticism and the Computational Turn. A "Methodological Moment" in Media and Communication Studies", in: *Medien & Kommunikationswissenschaft* 69, 1: 46–64.

Fickers, Andreas (2020): "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?", in: *Zeithistorische Forschungen/Studies in Contemporary History* 17, 1: 157–168.

Föhr, Pascal (2019): *Historische Quellenkritik im Digitalen Zeitalter*. Glückstadt: Verlag Werner Hülsbusch.

Haber, Peter (2012): "Zeitgeschichte und Digital Humanities", in: *Docupedia-Zeitgeschichte* 24.09.2012: http://docupedia.de/zg/haber_digital_humanities_v1_2012 [letzter Zugriff 29.11.2021].

Hiltmann, Torsten / Keupp, Jan / Althage, Melanie / Schneider, Philipp (2021): "Digital Methods in Practice. The Epistemological Implications of Applying Text Re-Use Analysis to the Bloody Accounts of the Conquest of Jerusalem (1099)", in: *Geschichte und Gesellschaft* 46, 1: 122–156 10.13109/gege.2021.47.1.122.

Margulies, Simon B. (2009): *Digitale Daten als Quelle der Geschichtswissenschaft. Eine Einführung* (Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik, Bd. 2). Hamburg: Verlag Dr. Kovač.

Milligan, Ian (2019): *History in the Age of Abundance? How the Web is Transforming Historical Research*. Montreal: McGill-Queen's University Press.

Owens, Trevor (2011): "Defining Data for Humanists: Text, Artifact, Information or Evidence?", in: *Journal of Digital Humanities* 1, 1: <http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/> [letzter Zugriff 29.11.2021].

Patel, Kiran K. (2011): "Zeitgeschichte im digitalen Zeitalter. Neue und alte Herausforderungen", in: *Vierteljahrshefte für Zeitgeschichte*, 59, 3: 331–351 10.1524/vfzg.2011.0019.

Ramsay, Stephen (2011): *Reading Machines. Toward an Algorithmic Criticism*. Urbana (Illinois): University of Illinois Press.

Schöch, Christof (2013): "Big? Smart? Clean? Messy? Data in the Humanities", in: *Journal of Digital Humanities* 2, 3: <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/> [letzter Zugriff 29.11.2021].

Schreiber, Catherina (2012): "Genuine Internetdaten als historische Quellen – Entwurf einer korrealistischen Quellentheorie", in: *Zeitschrift für digitale Geschichtswissenschaft*

ten 1: <http://universaar.uni-saarland.de/journals/index.php/zdg/article/view/292> [letzter Zugriff 29.11.2021].

Simmmler, Severin / Vitt, Thorsten / Pielström, Steffen (2019): "Topic Modeling with Interactive Visualizations in a GUI Tool", in: *Proceedings of the Digital Humanities Conference* <https://dev.clariah.nl/files/dh2019/boa/0637.html> [letzter Zugriff 29.11.2021].

Sinclair, Stéfan / Geoffrey Rockwell (2016): "Voyant Tools". Web. <http://voyant-tools.org/>.

Wettlaufer, Jörg (2016): "Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern", in: *ZfdG* 1: 10.17175/2016_011.

Kontextwissen zu historischen Quellen im Semantic Web

Die computergestützte Analyse heraldischer Wand- und Deckenmalereien mit Hilfe von Background Knowledge

Schneider, Philipp

philipp.schneider.1@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Das Dissertationsvorhaben "Coat of Arms in Context. The Aggregation and Analysis of Heraldic Data on Wall and Ceiling Paintings in French and German Speaking Areas (1300-1600) with Semantic Web Technologies" hat zum Ziel, mittels Semantic Web Technologien Wappensammlungen auf Wandmalereien zu erschließen, zu modellieren, und zu analysieren.¹ Dabei soll insbesondere untersucht werden, wie aus unterschiedlichen, heterogenen Provenienzen stammendes historisches Kontextwissen und Metadaten zu den Quellen in eine Analyse eingebunden werden müssen und welche Konsequenzen sich daraus für den Erkenntniswert solcher Datenauswertungen ergeben. Für die Analyse ist eine geschichtswissenschaftliche Forschungsperspektive maßgebend. Hierbei werden Funktion und Bedeutung heraldischer Wandmalereien für die visuelle Kommunikation von sozialem Aufstieg in den Blick genommen. Grundlage ist ein auszubauender Knowledge Graph eines Drittmittelprojekts.

Anforderungen an eine Ontologie heraldischer Wandmalereien und ihrer Kontexte

Insbesondere heraldische Wandmalereien zeichnen sich durch eine hohe Intermedialität aus (Pastoureau 1979). Diese prägt bereits einzelne Wappen – sie sind einem Akteur zugeordnet und tragen eine Bedeutung, die von Überlieferungszeit, -ort und im Besonderen vom Verwendungskontext abhängt, in dem das Wappen dargestellt und nachgenutzt wurde. Sind Wappen Teil einer

Wandmalerei, ist darüber hinaus auch der Raum, in dem sie sich befinden, ihre dortige Position und Relation zu anderen Abbildungen, die Funktion des Raums und dessen Wechselwirkung mit den ihn nutzenden Akteuren, sowie das Bauwerk, in dem sich der Raum befindet, von analytischer Relevanz (Meier 2007).

Diese Kontexte müssen sowohl in die digitale Repräsentation als auch in die Analyse heraldischer Wandmalereien eingebunden werden. Auf Modellierungsebene wird hierfür eine flexible und modulare Ontologie (Eide 2019) entwickelt, mit der sich die Wappen einer Wandmalerei in ihrem architektonischen, zeitlichen und personalen (bzw. institutionellen) Kontext verorten lassen. Ebenso muss berücksichtigt werden, dass einige Wandmalereien nur aus sekundären Nachweisen rekonstruiert werden können – einige ursprüngliche Werke sind zerstört, können aber z.B. über historische Zeichnungen nachgewiesen werden (Hablot 2020). In dieses Datenmodell wird eine Ontologie integriert, mit der sich einzelne Wappen formal beschreiben lassen (Hiltmann und Riechert 2020). Somit sind auch auf ikonografischer Ebene Bezüge zwischen einzelnen Teilen unterschiedlicher Wandmalereien sowie zwischen ihren Rezipient:innen und den Träger:innen der Wappen analysierbar.

Neben der Repräsentation von Multidimensionalität und Kontextwissen, müssen heterogene Datenbestände integriert werden. Nachgenutzt werden Datenbanken (z.B. (Armma), (Literatur und Wandmalerei)), Denkmallisten und analoge Verzeichnisse (z.B. (de Mérindol 2000)), die mit jeweils eigenen Vorannahmen, Modellierungsentscheidungen und Anwendungszielen erstellt wurden. Diese haben z.T. erhebliche Auswirkungen auf die Interpretierbarkeit, Vergleichbarkeit und Aussagekraft der Daten (Beretta 2021).

Methoden zur Analyse multidimensionaler Daten und ihrer Kontexte: Ansätze mit Semantic Web Technologien

Daneben hat das Dissertationsvorhaben zum Ziel, diese Daten auch für die Beantwortung konkreter geschichtswissenschaftlicher Fragestellungen nutzbar zu machen. Methodisch und technisch erfordert dies die Integration des zu modellierenden multidimensionalen historischen Kontextwissens, in konkrete Analyseverfahren. In der Regel besteht Datenanalyse aus dem Erkennen von Mustern, die dann mit Hilfe von Domänenwissen von der Forscher:in kontextualisiert und interpretiert werden – die Einbindung von Kontextwissen und die eigentliche Auswertung sind somit voneinander getrennt. Mit Hilfe von Semantic Web Technologien und Linked Data wurden in der Informatik mehrere Ansätze zur direkten Einbeziehung von Kontextwissen in Datenauswertungen entwickelt (Ristoski und Paulheim 2016). Zu nennen sind hier beispielsweise automatisches Ableiten von implizitem Wissen durch *Inferencing*, *Semantic Similarity*, oder die Berechnung von auf Kontextwissen basierenden Erklärungsvorschlägen für in den Daten identifizierte Cluster mittels induktiver logischer Programmierung (Tiddi 2018). Zentral sind außerdem Potenziale des *Graph Embedding*, bei dem Graphdaten in einen Vektorraum transformiert werden (Ristoski et al. 2019). Derartige Verfahren sollen im Rahmen der Dissertation erstmals für historische Daten erprobt werden. Sie können bestimmte historische (v.a. intermediale) Quellentypen überhaupt erst computergestützt analysierbar machen, im Sinne des Serendipitätsprinzips bei der

Interpretation angewandtes Domänenwissen ergänzen, um neue Zusammenhänge zu erfassen, sowie Analysen komplexer Graphdatenbanken stärker skalieren.

Analyse von Knowledge Graphen in den Humanities

Diese Vorteile wurden hinsichtlich der Erforschung von Kulturdaten schon in der Anfangszeit des Semantic Web diskutiert (Lin 2008). Eine Integration derartiger Ansätze in geisteswissenschaftliche Arbeiten, und insbesondere in den geschichtswissenschaftlichen Methodenkanon, ist jedoch nicht erfolgt. Dadurch sind auch die epistemologischen Auswirkungen einer computergestützten Integration von Daten als Kontextwissen in Analysen von Knowledge Graphen bislang nicht untersucht (Hogan et al. 2021).

Der Vortrag wird die beschriebenen Fragen und Herausforderungen des Dissertationsprojekts präsentieren. Dies erfolgt auf Grundlage einer ersten Version der Ontologie und einer ersten exemplarischen Datenauswertung. Zentral sind dabei die Modellierungsentscheidungen zur Repräsentation des multidimensionalen historischen Kontextwissens, deren Konsequenzen für die erprobten Analyseverfahren, sowie die Interpretierbarkeit ihrer Ergebnisse. So können erste Überlegungen hinsichtlich der Erkenntnispotenziale der vorgestellten Methoden für die Erforschung von Kulturobjekten diskutiert werden.

Fußnoten

1. Grundlage hierfür ist der, auszubauende, Knowledge Graph des von der Volkswagenstiftung geförderten Projekts *Die Performanz der Wappen*; <http://www.digitalheraldry.org/> (zuletzt aufgerufen am 29.11.2021).

Bibliographie

- “Armma. ARmorial Monumental du Moyen-Âge“. Letzter Zugriff: 13. Mai 2021. <https://armma.sapat.fr/>.
- Beretta, Francesco (2021): “A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)”, in: *Semantic Web* 12 (2) 279–294. <https://doi.org/10.3233/SW-200416>.
- Eide, Øyvind / Christian-Emil Smith Ore (2019): “Ontologies and Data Modeling” in: Flanders, Julia / Jannidis, Fotis (eds.): *The Shape of Data in the Digital Humanities. Modeling Texts and Text-Based Resources*. Abingdon 178–196.
- Hablot, Laurent (2020): “Le Cycle héraldique Du Couvent Des Jacobins de Poitiers. Un Armorial Des Morts de La Bataille de Poitiers 1356?” in: Hiltmann, Torsten / de Seixas, Miguel Metelo (eds.): *Heraldry in Medieval and Early Modern State Rooms* (= *Heraldic Studies* 3). Ostfildern 257–276.
- Hiltmann, Torsten / Thomas Riechert (2020): “Digital Heraldry. The State of the Art and New Approaches Based on Semantic Web Technologies” in: Balouzat-Loubet, Christelle (ed.): *L'édition En Ligne de Documents d'archives médiévaux*. Turnhout 102–125.
- Hogan, Aidan / Eva Blomqvist / Michael Cochez / Claudia D'amato / Gerard de Melo et al. (2021): “Knowledge Graphs” *ACM Computing Surveys* 54.4: 71:1–71:37. <https://doi.org/10.1145/3447772>.

Lin, Chia-Hung / Jen-Shin Hong / Martin Doerr (2008): "Issues in an Inference Platform for Generating Deductive Knowledge: A Case Study in Cultural Heritage Digital Libraries Using the CIDOC CRM", in: *International Journal on Digital Libraries* 8 (2): 115--132. <https://doi.org/10.1007/s00799-008-0034-0>.

"Literatur und Wandmalerei. Erscheinungsformen 'höfischer' Kultur Und Ihre Träger Im Mittelalter". Letzter Zugriff: 13. Mai 2021. <http://wandmalereien.imareal.sbg.ac.at/>.

Meier, Hans-Rudolf (2007): "Funktion Und Fiktion von Raumdekorationen. Zur Raumsymbolik Im Mittelalterlichen Profanbau." in: Staubach, Nikolaus / Johannerwage, Vera (eds.): *Außen Und Innen. Räume Und Ihre Symbolik Im Mittelalter* (= Tradition - Reform - Innovation. Studien Zur Modernität Des Mittelalters 14). Frankfurt a. M. 251--264.

Mérindol, Christian de (2000): *La Maison Des Chevaliers de Pont-Saint-Esprit. Les décors Peints. Corpus Des décors Monumentaux Peints Et Armoriés Du Moyen Âge En France*. Vol. 2. Pont-Saint-Esprit.

Pastoureau, Michel (1979): *Traité d'héraldique*. Paris.

Ristoski, Petar / Heiko Paulheim (2016): "Semantic Web in Data Mining and Knowledge Discovery: A Comprehensive Survey." *Journal of Web Semantics* 36: 1--22. <https://doi.org/10.1016/j.websem.2016.01.001>.

Ristoski, Petar / Jessica Rosati / Tommaso di Noia / Renato de Leone / Heiko Paulheim (2019): "RDF2Vec: RDF graph embeddings and their applications" *Semantic Web* 10.4: 721-752. <http://doi.org/10.3233/SW-180317>.

Tiddi, Ilaria (2018): *Explaining Data Patterns Using Knowledge from the Web of Data* (= Studies on the Semantic Web 34). Berlin. <https://www.iospress.nl/book/explaining-data-patterns-using-knowledge-from-the-web-of-data/>.

Relating the Unread Modellierungen der Literaturgeschichte

Brottrager, Judith

judith.brottrager@tu-darmstadt.de
TU Darmstadt, Germany

Von Beginn an war die Inklusion einer größeren Textmenge und damit auch nicht-kanonisierter Werke eines der Hauptargumente für digitale Ansätze in der Literaturwissenschaft: Durch Distant Reading, könne, so Moretti, eine alternative Literaturgeschichte erfasst werden, die auch das sogenannte „Great Unread“ mit einschließt (Moretti 2013: 48f). Forschungsprojekte der letzten Jahre haben jedoch gezeigt, dass der Mehrwert der digitalen Literaturwissenschaft nicht nur in der rein quantitativen Erweiterung der Untersuchungsgegenstände, sondern auch in der tiefgreifenden Kontextualisierung der untersuchten Texte in einer, wie es Bode nennt, „data-rich literary history“ (Bode 2018: 37–57) liegt. Diese kontextreichen Ansätze, wie beispielsweise Analysen zum Verhältnis von Prestige und Popularität (Underwood/Sellers 2016; Algee-Hewitt et al. 2016; Porter 2018; Underwood 2019: 68–110), zeigen, wie quantitative Methoden trotz nötiger Formalisierungen und Abstraktionen die Komplexität des literarischen Systems (cf. Bode 2017: 91) beschreiben können.

Mein Dissertationsprojekt folgt dieser Tradition, indem kanonisierte und nicht-kanonisierte literarische Werke auf unterschiedlichen Ebenen mit Rückgriffen auf literaturhistorische Daten

miteinander in Beziehung gesetzt werden. Die für die Untersuchungen erstellten Korpora und Datensätze umfassen etwa 1.200 englisch- und deutschsprachige Texte von 1688 bis 1914, wodurch diachrone und synchrone Vergleiche von Kanonisierungsprozessen und -mustern über Sprachgrenzen hinweg ermöglicht werden. Durch diese Datenvielfalt sollen einerseits etablierte literaturhistorische Narrative untersucht und quantitativ überprüft werden und andererseits literaturwissenschaftliche Kategorien wie Kanonisierung und Wertung so operationalisiert werden, dass sie mit computationellen Methoden zur Bestimmung von Textähnlichkeiten gewinnbringend kombiniert und diese Ähnlichkeit schließlich als Netzwerkmodelle dargestellt werden können.

Korpusaufbau

Der Korpusaufbau folgt einem systematisch angepassten Ansatz von Algee-Hewitt und McGurl, der darauf abzielt, von einem vorgefundenen zu einem maßgeschneiderten Korpus zu gelangen, indem Bestenlisten, Bestsellerlisten und von Expert*innen kuratierte Literaturlisten kombiniert werden, um ein repräsentatives Korpus für die englischsprachige Literatur des 20. Jahrhunderts zu erstellen (Algee-Hewitt/McGurl 2015). Durch die Kombination dieser Listen decken Algee-Hewitt und McGurl drei Ebenen der literarischen Produktion ab: den normativ-exklusiven Kanon, populäre Texte und von Expert*innen für Postkoloniale und Feministische Literaturwissenschaft vorgeschlagene Werke. Für die Umsetzung für die Zeitspanne von 1688-1914 wurde dieser Ansatz systematisch adaptiert, indem narrative Literaturgeschichten, Anthologien und (spezialisierte) Sekundärtexte, die diese Ebenen abdecken, identifiziert und als bibliografische Quellen für die Korpuserstellung genutzt wurden. Der Workflow umfasst Web-scraping, X-Technologien/Transformationen und Retro-Digitalisierungen.

Metadatenätze

Analog zur Korpuserstellung wurden Daten zur Kontextualisierung der jeweiligen Korpustexte, aber auch der gesamten literarischen Produktion gesammelt. Durch diese Daten können die Korpora mit den von Algee-Hewitt et al. als „the published“¹ und „the archive“² (Algee-Hewitt et al. 2016: 2) bezeichneten Ebenen der Literaturgeschichte verglichen werden, wodurch wiederum eine Einordnung der Untersuchungsergebnisse und eine Einschätzung der unvermeidlichen Selektionsprozesse bei der Korpuserstellung möglich sind (cf. Bode 2017: 85). Die Daten reichen von Publikationslisten und Leihbibliothekskatalogen bis hin zu Rezensionen und Daten zu Zweitaufgaben; der Aufbau der Datensätze orientiert sich am Beispiel der von Garside zur Verfügung gestellten Datenbank *British Fiction 1800-1829* (Garside 2011). In Anlehnung an erfolgreiche Metadatenanalysen (z.B. Jockers 2013: 35–62) sollen diese Daten zur Überprüfung von Hypothesen zum Zusammenhang zwischen dem Aufschwung des Romans als Gattung und geänderten Kanonisierungsprozessen verwendet werden (cf. Watt 1957; Raven 1987; Raven/Forster 2000; Tuchman/Fortin 2012).

Operationalisierungen

Die gesammelten Daten werden neben diesen Metadatenanalysen auch für die Operationalisierungen der literaturwissenschaft-

lichen Konzepte der Kanonisierung und Wertung eingesetzt. Aufbauend auf die theoretischen Grundlagen von Heydebrand und Winko werden Kanonisierung und Wertung als Scores implementiert, die ausdrücken, wie hoch die Wahrscheinlichkeit ist, dass ein bestimmter Text sehr kanonisiert ist beziehungsweise zur Entstehungszeit sehr gut rezipiert wurde (Heydebrand/Winko 1996). Ein besonderes Augenmerk liegt hierbei auf der Differenzierung der Konzepte und der Einbindung der Rezeptionsebene über Marker für Publikumsinteresse (wie Einträge in Leihbibliothekskatalogen und Zweitaufgaben innerhalb einer Generation) und sprachliche Werturteile, die durch Sentiment Analysen vergleichbar werden.

Ausblick

Unter Verwendung der generierten Scores soll schließlich untersucht werden, ob Kanonisierung und Wertung mit textintrinsic Merkmalen in Verbindung gebracht werden können. Stilometrische Berechnungen, Topic Modeling und Word Embeddings sowie wortartenbasierte Ansätze sollen dabei als allein-stehende Analysen der Textähnlichkeiten durchgeführt werden. Als zusätzliche Analysen- und Visualisierungsmethode dienen Netzwerkmodelle, die anhand der Ergebnisse der Textähnlichkeitsberechnungen erstellt werden, zur Exploration von Ähnlichkeitsstrukturen. Besonders auf dieser Ebene soll der Bezug zur literaturwissenschaftlichen Forschung durch die Identifikation von dichten stilistischen Ähnlichkeitsgruppen und durch aus den Modellen abgeleitete Einzeltextanalysen hergestellt werden.

Fußnoten

1. i.e. die Menge aller veröffentlichten Texte
2. i.e. die Menge der in Bibliotheken und Archiven gesicherten Texte

Bibliographie

- Algee-Hewitt, Mark / Allison, Sarah / Gemma, Marissa / Heuser, Ryan / Moretti, Franco / Walser, Hannah** (2016): "Canon/Archive. Large-scale dynamics in the literary field", *Pamphlets of the Stanford Literary Lab* 11. <https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf> [letzter Zugriff 15. Juli 2021].
- Algee-Hewitt, Mark / McGurl, Mark** (2015): "Between canon and corpus: Six perspectives on 20th-century novels", *Pamphlets of the Stanford Literary Lab* 8. <https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf> [letzter Zugriff 15. Juli 2021].
- Bode, Katherine** (2017): "The equivalence of "close" and "distant" reading; or, toward a new object for data-rich literary history", in: *Modern Language Quarterly* 78, 77–106.
- Bode, Katherine** (2018): *A world of fiction: Digital collections and the future of literary history*. Ann Arbor: University of Michigan Press.
- Garside, Peter** (2011): *British fiction 1800-1829. A database of production, circulation & reception*. <http://www.british-fiction.cf.ac.uk/> [letzter Zugriff 15. Juli 2021].
- Heydebrand, Renate von / Winko, Simone** (1996): *Einführung in die Wertung von Literatur: Systematik - Geschichte - Legitimation*. Paderborn: Schöningh
- Jockers, Matthew Lee** (2013): *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.

Moretti, Franco (2013): *Distant reading*. London / New York: Verso.

Porter, J.D. (2018): "Popularity/Prestige", *Pamphlets of the Stanford Literary Lab* 17. <https://litlab.stanford.edu/LiteraryLab-Pamphlet17.pdf> [letzter Zugriff 15. Juli 2021].

Raven, James (1987): *British fiction: 1750 – 1770. A chronological check-list of prose fiction printed in Britain and Ireland*. Newark: University of Delaware Press.

Raven, James / Forster, Antonia (eds.) (2000): *The English novel 1770 - 1829: A bibliographical survey of prose fiction published in the British Isles*. Oxford: Oxford University Press.

Tuchman, Gaye / Fortin, Nina E. (2012): *Edging women out: Victorian novelists, publishers and social change*. London: Routledge.

Underwood, Ted (2019): *Distant horizons: digital evidence and literary change*. Chicago: The University of Chicago Press.

Underwood, Ted / Sellers, Jordan (2016): "The Longue Durée of Literary Prestige", in: *Modern Language Quarterly* 77, 321–344.

Watt, Ian (1957): *The rise of the novel*. Berkeley: University of California Press.

Selektion und Nutzer*innen-Position in traditionellen und Internet- Informationsintermediären

Leyrer, Katharina

katharina.leyrer@fau.de
FAU Erlangen-Nürnberg, Germany

Motivation und Relevanz

Angesichts der Menge an Informationen, die uns täglich zur Verfügung stehen, müssen wir unweigerlich eine Auswahl treffen, um entscheidungs- und handlungsfähig zu sein. Bezüglich der Informationsauswahl im Internet ruft dies heftige Diskussionen und Bedenken hervor: So wird vor einer Polarisierung der Internetnutzer*innen und der Fragmentierung der Öffentlichkeit gewarnt. Zugleich wird befürchtet, dass Rezipient*innen an Autonomie verlieren, weil sie nicht mehr selbst entscheiden, welche Inhalte sie wahrnehmen, sondern in ihrer Informationsauswahl von Algorithmen beeinflusst werden (Pariser 2011, Schweiger 2017, Lobe 2018, Ngyuen 2018).

Dabei stehen vor allem Suchmaschinen und Soziale Netzwerk-Seiten in der Kritik. Als Informationsintermediäre vermitteln sie zwischen den Produzierenden und den Rezipierenden von Inhalten, indem sie Informationen auswählen und gewichten. Dies bietet Nutzer*innen einerseits Orientierung in der Informationsfülle, nimmt andererseits aber Einfluss auf die Inhalte, aus denen Nutzer*innen auswählen können (Jürgens & Stark 2017). In der Diskussion wird dabei oft ausgeklammert, dass die Vorauswahl von Information durch Intermediäre kein Phänomen der Digitalisierung ist, sondern ebenso durch traditionelle Intermediäre wie

publizistische Medienangebote, Verlage, Buchhandlungen oder Bibliotheken vorgenommen wird (Hagenhoff 2017).

Forschungsstand

Weder für Internet-Intermediäre, noch für traditionelle Intermediäre ist ausreichend untersucht, welche Faktoren und Normen ausschlaggebend dafür sind, wie Inhalte ausgewählt und für Nutzer*innen sichtbar gemacht werden. Daher kann auch keine Aussage darüber getroffen werden, wie sich die Selektion der Internet-Intermediäre von den Praktiken der traditionellen Intermediäre strukturell unterscheiden (Leyrer 2018). Darüber hinaus gibt es bislang keine vergleichenden Analysen, die zeigen, wie sich der Wechsel von traditionellen zu Internet-Intermediären auf die Autonomie der Nutzer*innen auswirkt (Hagenhoff 2020).

Forschungsfragen

Diese Arbeit widmet sich daher den Fragen:

- Welche Selektionslogiken und Normen bestimmen die Auswahl von Inhalten in Informationsintermediären im traditionellen und Internet-Kontext?
- Wie unterscheiden sich die Selektionslogiken in traditionellen Intermediären von den Selektionslogiken in Internet-Intermediären?
- Wie unterscheidet sich die Position der Nutzer*innen von Internet-Intermediären im Vergleich zur Position der Nutzer*innen traditioneller Intermediäre?

Theoretischer Rahmen und methodische Herangehensweise

Um sich diesen Fragen zu nähern, greift die Arbeit verschiedene Typen traditioneller und Internet-Informationsintermediäre beispielhaft heraus: So werden die beiden Internet-Intermediärstypen Suchmaschinen und Soziale Netzwerk-Seiten in den Blick genommen, da sie im Zentrum der Diskussion um die Informationsauswahl im Internet stehen. Als traditionelle Intermediärstypen werden Buchhandlungen, Verlage und Bibliotheken untersucht, da deren Selektionslogiken eine deutliche Forschungslücke darstellen (v.a. im Vergleich zu publizistischen Medienangeboten). Darüber hinaus wird die Recherche in Bibliotheken als Vorläufer oder Parallele zur Suche in Suchmaschinen gesehen (Nissenbaum 2010, Zimmer 2008), sodass ein Vergleich dieser beiden Intermediärstypen vielversprechend ist.

Selektionslogiken und Informationsnormen

Zuerst wird analysiert, welche Selektionslogiken die Auswahl von Inhalten im traditionellen Intermediärstyp *Bibliothek* und im Internet-Intermediärstyp *Suchmaschine* jeweils bestimmen. Dazu wird für jeden Intermediärstyp anhand des Filter-Modells von Bozdag (2013, Abb.1) dargestellt, welche Faktoren und Akteur*innen den Informationsfluss in den verschiedenen Stufen der Informationsverarbeitung beeinflussen. Da bisher nicht erforscht ist, welche Selektionslogiken die Medienauswahl in Bibliotheken bestimmen, werden Expert*innen-Interviews mit Erwerbsbi-

bliothekar*innen geführt und ausgewertet. Die Selektionslogiken, die in Suchmaschinen zum Einsatz kommen, werden anhand aktueller Forschungsliteratur herausgearbeitet (Lewandowski 2021).

Vergleich der Selektionslogiken

Anschließend werden die Selektionslogiken der Internet-Intermediäre *Suchmaschinen* mit denjenigen der traditionellen Intermediäre *Bibliotheken* verglichen. Die theoretische Basis bietet dafür die *Contextual Integrity Decision Heuristic* (CIDH) von Helen Nissenbaum (2010): Sie geht davon aus, dass jeder Kontext von handlungsweisenden Informationsnormen bestimmt wird, die festlegen, welche Praktiken angemessen sind. Um eine neue Praxis zu bewerten, werden Informationsnormen als Ausgangspunkt herangezogen, die sich in einem bereits bestehenden, vergleichbaren Kontext etabliert haben. Damit ermöglicht die CIDH eine Aussage darüber, ob die Selektionslogiken in Internet-Intermediären die Informationsnormen verletzen, die sich in traditionellen Intermediären etabliert haben.

Position der Nutzer*innen

Schließlich wird untersucht, wie sich die Digitalisierung auf die Autonomie und die Position der Nutzer*innen von Informationsintermediären auswirkt. Auf Basis der *Network Gatekeeping Theorie* von Barzilai-Nahon (2008) wird für jeden Intermediärstyp untersucht, welche Beziehung der*die Nutzer*in (*Gated*) zum Intermediär (*Gatekeeper*) hat. Die Position der *Gated* gegenüber dem *Gatekeeper* wird dabei anhand von vier Attributen beschrieben: politische Macht, Informationsproduktion, Beziehung zum *Gatekeeper* und Alternativen. Anhand der Ausprägung dieser vier Attribute wird verglichen, wie sich die Position der Nutzer*innen bei traditionellen und Internet-Gatekeepern unterscheidet.

Ziele und erste Ergebnisse

Ziel der Arbeit ist es, eine systematische empirische Analyse als Beitrag zur Debatte um die Informationsauswahl und die Nutzer*innen-Autonomie im Kontext von Internet-Informationsintermediären zu leisten. Erste Ergebnisse zeigen beispielsweise, dass sich die Position der Nutzer*innen gegenüber Informationsintermediären durch die Digitalisierung zwar verändert, aber nicht verschlechtert hat.

Bibliographie

Barzilai-Nahon, Karine (2008): „Toward a theory of network gatekeeping: A framework for exploring information control“, in: *Journal of the American Society for Information Science and Technology* 59(9): 1493–1512.

Bozdag, Engin (2013): „Bias in algorithmic filtering and personalization“, in: *Ethics and Information Technology* 15(3): 209–227.

Hagenhoff, Svenja (2017): „»Außer Kontrolle«: Alte und neue Informationsfluten im Publikationswesen“, in: Freiburg, Rudolf (ed.): *D@tenflut: Erlanger Universitätstage 2016*. Erlangen: FAU University Press 77–98.

Hagenhoff, Svenja (2020): „Digitale Souveränität“: Kontextualisierung des Phänomens in der Domäne der medial vermittelten öffentlichen Kommunikation unter besonderer Berücksichtigung von Reader Analytics. Erlanger Beiträge zur Medienwirtschaft 14 urn:nbn:de:bvb:29-opus4-150157.

Jürgens, Pascal / Stark, Birgit (2017): „The power of default on reddit: A general model to measure the influence of information intermediaries“, in: *Policy & Internet* 9(4): 395–419.

Leyrer, Katharina (2018): *Selektion und Bias in traditionellen und Internet-Informationsintermediären. Forschungsstand*. Erlanger Beiträge zur Medienwirtschaft 10 urn:nbn:de:bvb:29-opus4-102405.

Lewandowski, Dirk (2021): *Suchmaschinen verstehen*, 3. Aufl., Berlin, Springer Vieweg.

Lobe, Adrian (2018): „Wenn die Filterblase platzt“, in: *Süddeutsche Zeitung*, 10. Dezember <https://www.sueddeutsche.de/medien/filterblase-facebook-youtube-soziale-netzwerke-1.4245243-2> [Letzter Zugriff am 9 Dezember 2020].

Nguyen, C. Thi (2018): „Escape the echo chamber“, in: *Aeon*, 10. April <https://aeon.co/essays/why-its-as-hard-to-escape-an-echo-chamber-as-it-is-to-flee-a-cult> [Letzter Zugriff am 9 Dezember 2020].

Nissenbaum, Helen F. (2010): *Privacy in context: Technology, policy, and the integrity of social life*. Stanford, CA: Stanford Law Books.

Pariser, Eli (2011): *The filter bubble: What the Internet is hiding from you*. London: Viking.

Schweiger, Wolfgang (2017): *Der (des)informierte Bürger im Netz*. Wiesbaden: Springer Fachmedien.

Zimmer, Michael (2008): „Privacy on Planet Google: Using the theory of »Contextual Integrity« to clarify the privacy threats of Google's quest for the perfect search engine“, in: *Journal of Business & Technology Law* 3: 109–126.

The Remembered A Global Study of Literature Dissertations' Bibliography

Gutiérrez De la Torre, Silvia Eunice

silviaegt@gmail.com

Leipzig Universität, Deutschland

My doctoral project aims to build a map of contemporary literary research practices in German universities. To do so, I will apply reference mining techniques on a corpus of ca. 1,000 full-text electronic doctoral dissertations (ETDs) from Literature Departments in Germany (2000-2020) compiled via a research agreement with the National Library. In the following lines I will explain why ETDs are a necessary source to gain a much-needed strategic bird sight of the field, but also, what the potential problems of data heterogeneity in this dataset are, and how they will be addressed.

Dissertations are the default “rite of passage” through which students become researchers. Doctoral research typically passes through several phases of approval by an institutional community: the proposal acceptance, the internal upgrade/review, and the external panel or examination. Thus, they are a good compass of what different institutions consider “satisfactory knowledge” to become an academic. Moreover, other than briefer genres of academic writing such as journal articles, dissertations tend to require

in Germany, on average, 5 years (Jaksztat et al., 2012). Hence, it is safe to say that the reference lists contained in 1,000 Literature ETDs represent the readings of at least 5,000 years of human research that has not been analyzed until now.

Citation Analysis (CA) is the area of bibliometrics that studies the relationship between a cited and a citing document, and it has been increasingly used to “study, map, and evaluate academic research” (Hammarfelt, 2012). The basic input for CA is a citation database and Reference Mining (RM), a Natural Language Processing (NLP) task focused on the “detection, extraction and classification of bibliographic references and their constituent components” (Rodrigues Alves et al., 2018), has been a useful computational method to obtain this information as data.

However, as noted by Rodrigues Alves et al. (2018), unlike scientific publications (which have been the target of most RM methods), Humanities texts are significantly less structured. The following three reasons should be considered: 1) Humanities research uses both primary and secondary sources, and the former are, by definition, more varied; 2) references can happen anywhere in the text (footnotes, image captions, etc.); and lastly 3) “the variety of publication venues, languages, scholarly communities in the arts and humanities are broader, making reference practices and styles less uniform” (Rodrigues Alves et al., 2018). To this list we should add the fact due to the lack of strict editorial guidelines, dissertations tend to have a less structured reference list than in established publications.

Available methods (Tkaczyk et al., 2018) rely on a painstaking tagging process which nonetheless only works with texts that are close to the original training dataset (Grennan & Beel, 2020). Moreover, these methods require complex, black-box-like, and highly carbon-emitting computer power. In this presentation I will showcase a method that combines exploratory data analysis with readable machine learning results to automatically detect pages with reference lists on which NLP techniques such as Name Entity Recognition (NER) and fuzzy matching will be used to pair reference strings with richer metadata.

My hypothesis is that by paying attention to the citation patterns in a nation-wide corpus of literary dissertations, it is possible to reveal patterns of literary research in at least three dimensions: broad regional or institutional trends; interdisciplinary connections; and genre defined behaviors. With this in mind, this research proposes to answer the following questions:

1. What are the regional or institutional trends of literary research? (i.e. who are the most cited authors in Bavarian institutions, and how are they different or similar from those cited in other parts of Germany?)

2. How intercultural are these approaches? How often is Jorge Luis Borges cited along his German influences: Hölderlin, Silesius, Goethe?

3. Is it true that multidisciplinary approaches are becoming more popular and if so, what are their characteristics? For example, do medieval studies have a canon that includes publications from different sciences?

4. Is it possible to create topology of the citation networks of different genres and subgenres? For instance, which patterns of co-citation emerge in feminist fiction?

Reference mining Humanities dissertations is a challenging task that will require much more coordinated effort. Yet, this proposal offers, on the one hand, a comprehensive technique to extract, at least coarsely, the bibliographic “bricks” upon which PhD students build “new knowledge”; and on the other, a map of what and how (in which bibliographic interconnections) students in Germany have been analyzing literature in the last 20 years.

Bibliography

Hammarfelt, B. (2012). "Harvesting footnotes in a rural field: Citation patterns in Swedish literary studies." *Journal of Documentation*, 68 (4): 536–58.

Jaksztat, S., Preßler, N., & Briedis, K. (2012). *Promotionen im Fokus: Promotions- und Arbeitsbedingungen Promovierender im Vergleich*. https://www.dzhw.eu/pdf/pub_fh/fh-201215.pdf

Rodrigues Alves, D., Colavizza, G. and Kaplan, F. (2018). "Deep Reference Mining From Scholarly Literature in the Arts and Humanities". *Frontiers in Research Metrics and Analytics*, 3. Frontiers doi:10.3389/frma.2018.00021. <https://www.frontiersin.org/articles/10.3389/frma.2018.00021/full> (accessed 23 October 2020).

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). "Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers". *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 99–108. <https://doi.org/10.1145/3197026.3197048>

Transformation der Geschichtsschreibung? Der Einsatz von digitalen Medien und Forschungsmethoden in der historischen Praxis und dessen Folgen

Siebold, Anna

asiebold@mpiwg-berlin.mpg.de

Max-Planck-Institut für Wissenschaftsgeschichte und Carl von Ossietzky Universität Oldenburg

Forschungsthema: Das Dissertationsprojekt beschäftigt sich mit der Frage, inwieweit datengetriebene Forschungsansätze den historischen Arbeitsprozess transformieren und sich auf die Produktion von historischem Wissen auswirken.

Hintergrund: Ausgangspunkt ist die Beobachtung, dass die Verfügbarkeit digitaler Objekte und datengetriebener Forschungsansätze zunehmend die historisch arbeitenden Geisteswissenschaften prägen. Dies zeigt sich unter anderem in der Herausbildung von Disziplinen wie etwa der Digital Humanities (DH), der Digital bzw. Computational History und der Digital Art History, welche sich in Form von Lehrstühlen, Studiengängen, Journals, einer Vielzahl an Einführungsliteratur sowie Verbänden und Arbeitsgruppen professionalisiert und institutionalisiert haben. Ob und inwieweit die dort verfolgten digitalen Ansätze den historischen Forschungsprozess verändern, war lange umstritten (Alves 2014, 3). In der DH-Gemeinschaft lässt sich jedoch jüngst ein wachsendes Interesse an den epistemologischen Implikationen der datengetriebenen Forschungsansätze beobachten. Dass sich der historische Arbeitsprozess durch sie transformiert, scheint demnach zunehmend vorausgesetzt zu werden. Diskussionen um eine digitale Quellenkritik, die Forderung eines „Updates für die Hermeneutik“ und die Feststellung, die historische Forschung erfahre einen Medienwechsel, veranschaulichen

dies (vgl. Fickers, Andreas 2020; Föhr 2017; Hiltmann et al. 2021). Umfassende detaillierte Untersuchungen, die den veränderten historischen Forschungsprozess sowie den potentiell damit einhergehenden epistemologischen Wandel zu greifen versuchen, sind bislang allerdings kaum unternommen worden.¹

Forschungsdesign: Zur Untersuchung der Forschungsfrage werden komparative Fallstudien durchgeführt. Etwas enger gefasst fragt das Dissertationsprojekt demnach, inwieweit datengetriebene historisch operierende Forschungsprojekte *im Vergleich zu* Forschungsprojekten, in denen etablierte analoge Ansätze der Geschichtsforschung angewandt werden, zu veränderten Forschungszugängen und -prozessen führen. Jedes Fallbeispiel besteht aus einer Gegenüberstellung von zwei historischen Forschungsprojekten, die den gleichen oder einen ähnlichen Forschungsgegenstand untersuchen und dabei unterschiedlich verfahren: einmal digitalen und einmal etablierten analogen Forschungsansätzen folgend. Anhaltspunkte für die Analyse bieten: (1) die jeweils gewählte Forschungsfrage, (2) das historische Ausgangsmaterial, (3) die angewandten Verfahren, (4) die gewonnenen historischen Erkenntnisse und (5) deren Darstellungsweisen. Das Untersuchungsmaterial des Dissertationsprojekts bilden zum einen die Materialien, die die Historiker:innen während ihrer Forschung produziert haben: etwa Vorträge, mündliche und schriftliche Präsentationen von Zwischenständen, Webseite-Präsentationen und einschlägige Veröffentlichungen. Zum anderen werden leitfadengestützte Interviews mit den Historiker:innen geführt, in denen Fragen zur Entstehung, Wahl des Forschungsdesigns und zu den Möglichkeiten und Grenzen der eingesetzten Verfahren gestellt werden.

Vorgesehen ist die Analyse von insgesamt drei Fällen. Den Ausgangspunkt bilden DH-Forschungsprojekte deutscher Institutionen. Um Repräsentativität zu gewährleisten, werden Projekte ausgewählt, die Verfahren und Technologien anwenden, welche die DH-Forschung maßgeblich prägen. Der Auswahl geht deshalb eine empirische Untersuchung voraus, die ermittelt, welche von ihnen am weitesten verbreitet sind und/oder in den vergangenen Jahren an Relevanz gewonnen haben.² Für die Auswahl des jeweiligen analogen Pendants ist die Rezeption des Forschungsprojekts entscheidend.

Forschungsziel: Ziel ist es, zu untersuchen, inwieweit datengetriebene Ansätze den Forschungsprozess verändern und sich auf die Produktion von historischem Wissen auswirken – was ermöglichen, was verhindern sie? Schaffen digitale Forschungsansätze spezifische historische Perspektiven? Lassen sich allgemeine Tendenzen aufzeigen? Behauptungen, die in verschiedenen Kontexten bereits aufgestellt wurden, sind etwa, DH-Forschung geschehe „bottom-up“ (Braake et al. 2016), ermögliche ein „Distant Reading“ der Quellen (Moretti 2013), erlaube es, Forschungsfragen auf lange Zeiträume und große geographische Räume zu skalieren (Alexandrakis / Walther 2021), neige zu Longue durée-Perspektiven (Guldi / Armitage 2014), verfolge explorative Ansätze (Röhle 2014, 167–68; Gibbs / Owens 2011), sei interdisziplinär und ermögliche Selbstreflexivität (Krämer 2018, 10). Das Dissertationsprojekt möchte eine Überprüfung und Erweiterung dieser Folgen für die historisch arbeitenden Geisteswissenschaften leisten, sie in Relation zu konkreten Verfahren und Technologien setzen und kritisch reflektieren.

Stand der Arbeit (Beginn des zweiten Dissertationsjahres): Zunächst erfolgte die Beschäftigung mit der DH-Landschaft, ihren einschlägigen Verfahren und Technologien sowie deren Klassifizierung. Die Erhebung der am weitesten verbreiteten Verfahren und Technologien ist nahezu abgeschlossen. Sie bildet die Grundlage für die Auswahl der DH-Forschungsprojekte, die zeitnah stattfindet.

Fußnoten

1. Vgl. u.a. (Braake et al 2016, 25) sowie (Hiltmann et al. 2021, 122). Hiltmann et al. legen eine Untersuchung vor, die die epistemologischen Folgen eines datengetriebenen Ansatzes auf Basis eines Vergleichs mit einem analogen Verfahren herausarbeitet. Dabei handelt es sich sowohl im Analogen als auch im Digitalen um die Suche nach wiederverwendetem Text („syntactic text re-use“ und „semantic text re-use“) in zwei Textkorpora.
2. Hierzu werden verschiedene Datensätze herangezogen: (1) von der DFG geförderte geisteswissenschaftliche Projekte, die datengetriebene Ansätze verfolgen, (2) E-Mails der Mailingliste der DHd, (3) die Books of Abstracts der DHd-Konferenzen.

Bibliographie

- Alexandrakis, Katja / Walther, Daniel** (2021): „Gastbeitrag: ‚Digital Humanities. Geistes- und Kulturwissenschaften im Fokus der Digitalisierung.‘ Digitale Geisteswissenschaften in der Forschungs-, Arbeits- und Medienwelt von heute und morgen – ein Interview mit Dr. Mareike König (Deutsches Historisches Institut Paris)“. <https://wbg-community.de/themen/gastbeitrag-digital-humanities-geistes-kulturwissenschaften-im-fokus-digitalisierung-von-katja-alexandrakis-dr-daniel-walther-fzi-forschungszentrum-informatik>.
- Alves, Daniel** (2014): „Digital Methods and Tools for Historical Research“ in: *International Journal of Humanities and Arts Computing* 8 (1): 1–12. <https://doi.org/10.3366/ijhac.2014.0116>.
- Braake, Serge ter / Fokkens, Antske / Ockeloen, Niels / Son, Chantal van** (2016): „Digital History: Towards New Methodologies“ in: Božić, Bojan / Mendel-Gleason, Gavin / Debruyne, Christophe / O’Sullivan, Declan (eds): *Computational History and Data-Driven Humanities*. Cham: Springer International Publishing 23–32. https://doi.org/10.1007/978-3-319-46224-0_3.
- Fickers, Andreas** (2020): „Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?“ in: *Zeithistorische Forschungen / Studies in Contemporary History*, Nr. 17 (1) 157–68. <https://doi.org/10.14765/ZZF.DOK-1765>.
- Föhr, Pascal** (2019): *Historische Quellenkritik im Digitalen Zeitalter*. Glückstadt: Werner Hülsbusch.
- Gibbs, Frederick W. / Trevor J. Owens** (2013). „Hermeneutics of Data and Historical Writing“ in: Dougherty, Jack / Nawrotzki, Kristen (eds): *Writing History in the Digital Age*. University of Michigan Press. <https://doi.org/10.3998/dh.12230987.0001.001>.
- Guldi, Jo / Armitage, David** (2014). *The History Manifesto*. Cambridge University Press.
- Hiltmann, Torsten / Keupp, Jan / Althage, Melanie / Schneider, Philipp** (2021). „Digital Methods in Practice“ in: *Geschichte und Gesellschaft* 47 (1): 122–56.
- Krämer, Sybille** (2018). „Der ‚Stachel des Digitalen‘ – ein Anreiz zur Selbstreflexion in den Geisteswissenschaften? Ein philosophischer Kommentar zu den Digital Humanities in neun Thesen“ in: *Digital Classics Online* 4 (1): 5–11. <https://doi.org/10.11588/dco.2017.0.48490>.
- Moretti, Franco** (2013). *Distant Reading*. London: Verso, 2013.
- Röhle, Theo** (2014). „Big Data – Big Humanities?: Eine historische Perspektive“ in: Ramón Reichert (ed.): *Digitale Gesellschaft*, Bielefeld: transcript 157–72. <https://doi.org/10.14361/transcript.9783839425923.157>.

Posterpräsentationen

Aktualität und Gedächtnis

Zur korpusanalytischen

Untersuchung von

Gegenwartsliteratur auf Twitter

Meier-Vieracker, Simon

simon.meier-vieracker@tu-dresden.de
TU Dresden, Germany

Kreuzmair, Elias

elias.kreuzmair@uni-greifswald.de
Universität Greifswald, Germany

Twitter gilt als ein besonders präsentisches Medium. Mit der in der deutschen Fassung in der Eingabemaske gestellten Frage „Was gibt's Neues?“ und überhaupt der Möglichkeit, sich gleichsam mit einem an eine potenziell unbegrenzte Öffentlichkeit gerichteten Instant Messenger in Echtzeit an öffentlichen Diskursen zu beteiligen, lädt Twitter in besonderer Weise zu gegenwartsorientiertem Schreiben ein. Auch aus einer Rezipierendenperspektive kann der primäre Zugriffsmodus der Timeline als Echtzeit-Nachrichtenticker beschrieben werden (Hermes 2021). Zugleich machen die medialen Affordanzen der Persistenz und der Durchsuchbarkeit (boyd 2014: 11) der auf Twitter platzierten Inhalte die Plattform zu einem umfassenden und weitgehend frei zugänglichen Archiv und mithin zu einem breit und vielfältig genutzten Medium des digitalen Gedächtnisses. Dass Twitter sich inzwischen selbst nicht nur als auf Aktualität gerichtetes Medium, sondern auch als Gedächtnismedium versteht, zeigt die Öffnung des Twitter-Archivs mit der neuen API für Forscher:innen (Torres/Trujillo 2021).

Dieses Spannungsverhältnis zwischen gegenwartsbezogenem Schreiben und digitalem Archiv wird nicht nur in zeitdiagnostischen Texten (Osten 2004, Renouard 2018), sondern auch im gegenwartsliterarischen Diskurs auf Twitter selbst immer wieder thematisch. Schriftsteller:innen nutzen Twitter – oder werden dort überhaupt erst zu solchen – und die damit einhergehenden Publikations- und Vernetzungsmöglichkeiten, wobei Kurztexte mit eigenem literarischem Wert, Metadiskurse über Literatur- und den Literaturbetrieb und schließlich auch interaktionsorientierte Kommunikation ineinander übergehen. Bei dieser manchmal als „Twitteratur“ bezeichneten Domäne (Kreuzmair 2016) handelt es sich also um Gegenwartsliteratur und -reflexion, die flüchtig und momentbezogen ist und sich dennoch gleichsam selbst archiviert. Durch die hohe Selbstreflexivität des Diskurses im Sinne ständiger Selbstbeobachtung werden dabei die Gegenwärtigkeit, aber auch Erinnerbarkeit des eigenen Schreibens immer wieder reflektiert.

Für das DFG-Projekt „Schreibweisen der Gegenwart. Zeitreflexion und literarische Verfahren nach der Digitalisierung“ (2020–2022), das nach Wechselwirkungen von Zeitreflexion und literarischen Verfahren unter den Bedingungen der Digitalisierung fragt, haben wir im Februar 2020 ein Twitter-Korpus mit den Timelines von 117 öffentlichen Accounts erstellt, die sich in einem erweiterten Sinne der deutschsprachigen Literaturszene zurechnen lassen. Der Datenerhebung war eine teilnehmende Beobachtung über zwei Monate vorausgegangen, die neben der Identifizierung relevanter Accounts auch ergeben hat, dass

eine starke Trennung Autor*innen/Literaturbetrieb nicht sinnvoll ist. Unter Nutzung der API über die Software rtweet (Kearney 2018) konnten den Beschränkungen der API entsprechend pro Account bis zu 3000 Tweets erhoben werden, das Korpus umfasst insgesamt 219.450 Tweets aus dem Zeitraum 2009–2020 im Umfang von 3.552.773 Wörtern. Für korpuslinguistische Untersuchungen wurden die Daten in einem XML-Format mit umfangreichen Metadaten aufbereitet, das über den Text hinaus auch interaktive Aspekte wie Reply-Strukturen (Hoppe et al. 2018) und Social Media-charakteristische Metadaten wie Anzahl der Likes und Retweets erfasst. Die Entscheidung für ein eigenes Datenmodell begründet sich durch die fehlenden Standards zur Encodierung von Social Media-Texten (etwa nach TEI), welche die für unsere Fragestellung relevante Interaktivität erfassen. Die Daten wurden mithilfe der auf Social Media-Daten trainierten Python-Module SoMaJo und SoMeWeTa (Proisl & Uhrig 2016; Proisl 2018) tokenisiert, nach Wortarten annotiert und lemmatisiert. Über die webbasierte Korpusanalyseplattform CQPweb (Hardie 2012), das äußerst flexible Abfragen der annotierten Daten und der Metadaten erlaubt, wird das Korpus den Projektbeteiligten zur Verfügung gestellt. Formulierungsmuster, semantische Profile, aber auch Interaktions- und Vernetzungsstrukturen können so computergestützt untersucht werden. Dafür können die in die Software implementierten korpuslinguistischen Methoden wie Keywords in Context, Kollokationsanalysen, Ngramm-Analysen, Distributionsanalysen und Keyword-Berechnungen genutzt werden. Darüber hinaus sind im Projekt auch andere digitale Textanalysemethoden wie etwa Topic Modeling zum Einsatz gekommen (Schöch 2017).

Das so erstellte und als Grundlage für die weiteren Analysen herangezogene Korpus ist somit eine Momentaufnahme, eine Fixierung und mithin auch ein Stillstellungs-Artefakt (Jäger 2011: 315) eines eigentlich äußerst fluiden Diskurses. Er wird so der auch quantifizierbaren und reproduzierbaren Analyse zugänglich, gleichsam als digital prozessiertes Datenerbe, und büßt dadurch aber zugleich jene Offenheit und Flüchtigkeit ein, die Social Media-Kommunikation in besonderem Maße auszeichnet.

Auf dem Poster stellen wir zum einen unsere Pipeline für die Korpuserstellung und -aufbereitung sowie unser Datenmodell für die Korpusrepräsentation vor. Zum anderen präsentieren wir exemplarische Analyseergebnisse zu Zeitreflexionen auf Twitter wie etwa die auffallende Häufigkeit der Trigramme „das erste Mal“ und „den ganzen Tag“, die einerseits die Zeitbezogenheit des Schreibens auf Twitter anzeigt und andererseits seine tagebuchartige Funktion als Archiv für bemerkenswerte Ereignisse vor Augen führt. Davon ausgehend werden wir die methodologischen Schwierigkeiten diskutieren, die sich aus der analytisch unumgänglichen Fixierung eines eigentlich fluiden Medienformates ergeben.

Da der Urheberrechtsstatus von Tweets bislang ungeklärt ist, ist eine vollständige Publikation des Datensatzes leider nicht möglich. Interessierten Forschenden Zugriff auf das Korpus in der verwendeten Analyseplattform CQPweb gewährt werden, wo im Korpus recherchiert werden kann.

Bibliographie

boyd, danah (2014): *It's complicated: the social lives of networked teens*. New Haven: Yale University Press.

Hardie, Andrew (2012): "CQPweb — combining power, flexibility and usability in a corpus analysis tool", in: *International Journal of Corpus Linguistics* 17(3): 380–409. 10.1075/ijcl.17.3.04har.

Hermes, Jürgen (2021): "Chirpy Humanities". Billet *Public Humanities* <https://publicdh.hypotheses.org/42> (letzter Zugriff 14.07.2021).

Hoppe, Imke, Lörcher, Ines / Neverla, Irene / Kießling, Bastian (2018): „Gespräch zwischen vielen oder Monologe von einzelnen? Das Konzept „Interaktivität“ und seine Eignung für die inhaltsanalytische Erfassung der Komplexität von Online-Kommentaren“, in: *Neue Komplexitäten für Kommunikationsforschung und Medienanalyse: Analytische Zugänge und empirische Studien (Digital Communication Research Band 4)*. doi:10.17174/dcr.v4.9.

Jäger, Ludwig (2011): „Intermedialität – Intramedialität – Transkriptivität. Überlegungen zu einigen Prinzipien der kulturellen Semiosis“, in: Arnulf Deppermann / Angelika Linke (eds.), *Sprache intermedial. Stimme und Schrift, Bild und Ton*, 301–323. Berlin, Boston: De Gruyter. 10.1515/9783110223613.299.

Kearney, Michael W. (2018): *rtweet: Collecting Twitter data*. Zenodo. 10.5281/zenodo.2528481.

Kreuzmair, Elias (2016): „Was war Twitteratur?“, in: *Merkur*. <https://www.merkur-zeitschrift.de/2016/02/04/was-war-twitteratur/> (letzter Zugriff 10.06.2021).

Osten, Manfred (2004): *Das geraubte Gedächtnis: Digitale Systeme und die Zerstörung der Erinnerungskultur. Eine kleine Geschichte des Vergessens*. Frankfurt am Main: Insel.

Proisl, Thomas (2018): „SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts“, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* 665–670.

Proisl, Thomas / Uhrig, Peter (2016): „SoMaJo: State-of-the-art tokenization for German web and social media texts“, in: *Proceedings of the 10th Web as Corpus Workshop*. Berlin: Association for Computational Linguistics, 57–62. 10.18653/v1/W16-2607.

Renouard, Maël (2018): *Fragmente eines unendlichen Gedächtnisses*. Zürich: Diaphanes.

Schöch, Christof (2017): „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“, in: *Digital Humanities Quarterly* 11(2).

Tornes, Adam / Trujillo, Leanne (2021): *Enabling the future of academic research with the Twitter API*. https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api (letzter Zugriff 15.07.2021).

Analyse der Rezeption von Telenovelas und Serien über lateinamerikanische Geschichte durch Algorithmen

Meding, Holle Ameriga

holle.meding@fu-berlin.de

Geschichtsvermittlung durch Unterhaltungsmedien in Lateinamerika. Labor für Erinnerungsforschung und digitale Methoden (GUMELAB, FU Berlin), Germany

Contreras Saiz, Mónica

m.contreras.saiz@fu-berlin.de

Geschichtsvermittlung durch Unterhaltungsmedien in Lateinamerika. Labor für Erinnerungsforschung und digitale Methoden (GUMELAB, FU Berlin), Germany

Muessemann, Hannah

hannah.muessemann@fu-berlin.de

Geschichtsvermittlung durch Unterhaltungsmedien in Lateinamerika. Labor für Erinnerungsforschung und digitale Methoden (GUMELAB, FU Berlin), Germany

In Lateinamerika hat die Produktion von Telenovelas und Serien, die die jüngere Vergangenheit der Region thematisieren, in den letzten zwanzig Jahren zugenommen. Ereignisse, wie die Militärdiktaturen in Chile, Argentinien und Brasilien oder der bewaffnete Konflikt in Kolumbien dienen als Grundlage für Geschichten, die durch audiovisuellen Unterhaltungsmedien im Fernsehen ein breites Publikum erreichen. Die Darstellungen prägen dadurch die kollektiven Erinnerungen über Ländergrenzen hinweg. Auch der Einfluss der sozialen Medien hat in den letzten zehn Jahren im Feld des kulturellen Gedächtnisses erheblich zugenommen. Soziale Medien werden dabei aufgrund von Kommentaren und Hashtags zu einem Raum für die Vermittlung und Diskussion der Telenovelas und Serien.

Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Projekt *Geschichtsvermittlung durch Unterhaltungsmedien in Lateinamerika. Labor für Erinnerungsforschung und digitale Methoden* (GUMELAB) des Lateinamerika-Instituts der Freien Universität Berlin untersucht die trans-/nationale Rezeption lateinamerikanischer Geschichte anhand von Telenovelas und Serien. Dabei wird der Einfluss dieser Unterhaltungsmedien auf die Erinnerungsbilder, das Geschichtsbewusstsein und die politische Bildung der Zuschauer:innen analysiert.

Durch die Diskussionen über Telenovelas und Serien in sozialen Netzwerken entstehen große Datenmengen, die sich als stetig wandelnde historische Quellen betrachtet werden können. Im interdisziplinären Projekt GUMELAB werden sie mit Methoden der Natural Language Processing ausgewertet. Dafür werden Algorithmen zur Textklassifizierung und Named Entity Recognition verwendet, um die Hauptaussagen von Interaktionen auf Twitter, Facebook, YouTube und Online-Zeitungen zu identifizieren und zu kategorisieren. Dies erfolgt durch die Programmierung eines „Suchalgorithmus“ für die Auswahl und Charakterisierung für die forschungsrelevanten Informationsquellen. Analysiert werden spanischsprachige Daten, die seit dem Ausstrahlungsdatum der jeweiligen chilenischen oder kolumbianischen Produktion bis Ende Juli 2021 im Netz auftauchten.

Die nötigen Kriterien beruhen auf der Relevanz, dem Umfang und der Verfügbarkeit der Informationen in der jeweiligen Analyseinheit (Tweet, Post, Kommentar, Absatz usw.). Eine repräsentative Teilmenge wird anschließend manuell in relevante und nicht relevante Einheiten sortiert. Diese bilden die Grundlage und den Input für die Konstruktion dreier Modelle kognitiver Agenten (KA; Evidence-based Reasoning, EBR) (Koeman 2019, Tecuci et al. 2019, ITU 2019), die für die Hauptanalysekategorien (Erinnerungsbilder, Geschichtsbewusstsein und politische Bildung) entworfen wurden. Nach der Zuordnung einzelner Daten zu den Kategorien werden die KAs mit Hilfe des GPT-3-Modells (Generative Pre-trained Transformer 3) und der Theorie komplexer Netzwerke (Barabási 2016) trainiert. Sie ermöglichen es, das logische Denken zu trainieren, das die Hauptkategorien der For-

schungsanalyse in der historischen Quelle interpretiert. Außerdem helfen sie bei der quantitativen Auswertung des Quellenkorpus und der Visualisierung der Verhaltensdynamik der sozialen Netzwerke. Die erhobenen Daten werden in einer internen Datenbank gespeichert und im Anschluss mit den Ergebnissen der qualitativen Interviews mit Zuschauer:innen aus Chile, Kolumbien und der USA abgeglichen.

Ein tieferes Verständnis von den untersuchten Prozessen ist von wissenschaftlicher und gesellschaftspolitischer Relevanz, da es ein neues Licht auf die Rezeption politischer und historischer Zusammenhänge durch Unterhaltungsmedien wirft. Die Vermittlung eines kulturellen Gedächtnisses durch fiktionale Darstellungen, die auf geschichtlichen Ereignissen beruhen, erlauben uns, Geschichtsvermittlung als eine gesamtgesellschaftliche Aufgabe zu verstehen.

Innerhalb der Erinnerungsforschung (Assmann 1999; Halbwachs 1939; Jelin 2002) wurde der Einfluss audiovisueller Produktionen auf das kollektive Gedächtnis (Erll 2008; Landsberg 2004) und deren Rezeption wenig untersucht. Eine vergleichende Untersuchung, die die Rezeption von Unterhaltungsmedien in sozialen Netzwerken mit digitalen Methoden (Botero 2018; Jensen 2012) untersucht, wurde bisher noch nicht durchgeführt. Die angestrebten Ergebnisse des Projektes GUMELAB sollen aufzeigen, wie Unterhaltungsmedien, in denen die Vergangenheit interpretiert und inszeniert wird, zu Zwecken der politischen Bildung genutzt werden können. Der methodische Einsatz von kognitiven Agenten wird durch die stetig anwachsenden Datenmengen auch in Zukunft für die Geschichtswissenschaften immer wichtiger werden, insbesondere wenn es sich dabei um dynamische, digital entstehende Quellen unterschiedlicher Art handelt, die im Netz kursieren.

Bibliographie

- Assmann, Aleida** (1999): *Erinnerungsräume. Formen und Wandlung des kulturellen Gedächtnisses*. München: Beck.
- Barabási, Albert-László** (2016): *Network Science*. Unter Mitarbeit von Márton Pósfai. Cambridge: Cambridge University Press.
- Botero, J.** (2018): *Description and Prediction of Collective Human Behavior: A Complex Systems Perspective*. Doctoral Thesis, Universidad de Antioquia.
- Erll, Astrid** (Hg.) (2008): *Film und kulturelle Erinnerung. Plurimediale Konstellationen*. Berlin: De Gruyter.
- Halbwachs, Maurice** [1939] (1991): *Das kollektive Gedächtnis*. Frankfurt am Main: Fischer-Taschenbuch-Verl.
- ITU** (2019): *United Nations Activities on Artificial Intelligence (AI)*.
- Jelin, Elizabeth** (2002): *Los trabajos de la memoria*. Madrid: Siglo XXI de España; SSRG.
- Jensen, K.** (Hg.) (2012): *A Handbook of Media and Communication Research. Qualitative and Quantitative Methodologies*. London: Routledge.
- Koeman, V.** (2019): *Tools for Developing Cognitive Agents*. <https://doi.org/10.4233/uui-d:f80750ee-db68-480e-8c58-2c167bd24ee5>
- Landsberg, Alison** (2004): *Prosthetic memory. The transformation of American remembrance in the age of mass culture*. New York: Columbia University Press.
- Tecuci, G., Marcu, D., Mihai B., David S.** (2019): *Toward a Computational Theory of Evidence-Based Reasoning for Instructable Cognitive Agents*. arXiv:1910.03990 [cs.AI]

Anpassungen von LERA zum Vergleich hebräischer Textzeugen des kabbalistischen Traktats Keter Shem Tov

Pöckelmann, Marcus

marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Germany

Rebiger, Bill

bill.rebiger@judaistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Germany

LERA, ein Tool zum Vergleich von Textzeugen

LERA ist ein interaktives digitales Werkzeug zur Analyse der Differenzen und Gemeinsamkeiten zwischen mehreren Fassungen oder Zeugen, eines Textes¹. Die Berechnung der Differenzen und Gemeinsamkeiten erfolgt in zwei Schritten. In einem ersten Schritt wird eine automatische Alignierung der Segmente, in der Regel der Absätze, der zu vergleichenden Textzeugen auf Basis des Jaccard-Indexes berechnet. Der/Die Fachwissenschaftler:in kann anschließend durch Aufschneiden und Zusammenfügen von Segmenten bzw. durch Bewegen einzelner Segmente die Segmentalignierung manuell nach bearbeiten. In der zweiten Phase erfolgt ein detaillierter, auf der Levenshtein-Distanz basierender Vergleich der alignierten Segmente (Pöckelmann et al., 2021). Die gewonnenen Daten werden in einer synoptischen Gegenüberstellung der einander zugeordneten Segmente samt der farblich hervorgehobenen Textvarianten dargestellt. Ein in LERA integriertes Tool zum *Distant Reading* weist den/die Fachwissenschaftler:in auf Stellen hin, an denen sich die Textzeugen stark (oder weniger stark) unterscheiden (Pöckelmann et al. 2015). Neben der direkten Analyse der Textvarianten in LERAs webbasierter Nutzeroberfläche, stehen verschiedene Exportformate² für die externe Weiterverarbeitung zur Verfügung.

LERA wurde ursprünglich, von 2012-2016, zur Untersuchung der Genese der *Histoire philosophique et politique des établissements et du commerce des Européens dans les deux Indes* von Guillaume-Thomas Raynal im Rahmen eines BMBF-Projektes entwickelt, einem Werk der französischen Aufklärung, welches in vier Druckauflagen (1770, 1774, 1780, 1820) erschien. Die vier Druckauflagen unterscheiden sich insbesondere in der sich ändernden Wortwahl des Autors, bedingt durch den jeweils aktuellen Wissensstand des Autors in Bezug auf die durch die Europäer kolonisierten Gebiete in Mittel- und Südamerika sowie Indien und anderen Teilen Asiens (Bremer et al., 2015). Seit dem wird bzw. wurde LERA in mehreren Editionsprojekten eingesetzt, beispielsweise im Rahmen der Studien der arabischen Manuskripte zu Kalila and Dimna (Gründler und Pöckelmann, 2018), zur Erstellung

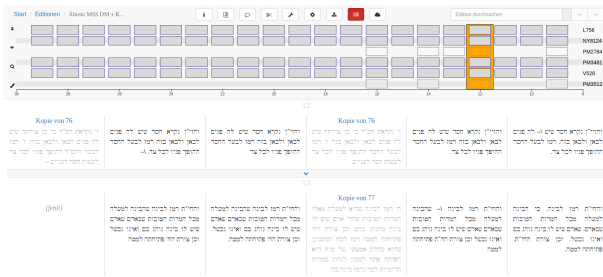


Abb. 3: Darstellung transponierter Segmente als Kopien in LERAs Übersichtsleiste und in der Spaltensynopse.

Judaistische Ergebnisse, die mit LERA erzielt wurden

Der anonyme kabbalistische Traktat *Keter Shem Tov* („Krone des guten Namens“; um 1260), der häufig auf der Basis eines Responsums von Salomo ben Adret (1235–1310) einem aschkenasischen Prediger, Abraham ben Axelrad aus Köln, zugeschrieben wird, bezeugt die auch in didaktischer Hinsicht für kabbalistische Lehren zentrale Synthese von Spekulationen zum vierbuchstaben Gottesnamen mit dem Konzept der zehn innergöttlichen Kräfte (*Sefirot*). Die sechzig wichtigsten Handschriften, d.h. über die Hälfte der bekannten Textzeugen, wurden transkribiert und für die digitale Weiterverarbeitung in LERA mit Metadaten aufbereitet, wobei die Auswahl durch Kriterien wie textkritische Relevanz, Datierung, Umfang, Zustand und Verfügbarkeit bestimmt wurde. Neben der Standardversion, die in der Tradition und Forschung bereits durch Druckausgaben bekannt ist – allerdings in deutlich schlechterer Lesart als in der nunmehr vorliegenden Edition –, konnten eine Mischversion und eine Kurzversion unterschieden werden. Darüber hinaus enthalten einige Textzeugen auch Sondergut, das in die jeweiligen Versionen integriert wurde. Eine spezielle Rezeption liegt in MS Jerusalem, NLI, 541, im Kontext der Überlieferung von Schriften vor, die der Schule des Abraham Abulafia (1240–ca.1292) nahestehen. Mithilfe der digitalen Analyse in LERA wurden die stemmatologischen Beziehungen zwischen den Textzeugen untersucht sowie deren Zugehörigkeit zu einzelnen Versionen ermittelt. Die Ergebnisse dieser Forschungen führten zu einer überraschenden Neubewertung: die Versionen von *Keter Shem Tov* unterscheiden sich nicht einfach nur darin, dass seine Mikroformen unterschiedlich konfiguriert sind, sondern dass diese Mikroformen gleichzeitig Bestandteile anderer Makroformen sind.⁸ So ist *Keter Shem Tov* mit anderen Makroformen wie z.B. den verschiedenen Versionen der sog. *Divre Menahem* („Worte des Menahem“) oder *Perush 'Eser Sefirot* („Auslegung der zehn Sefirot“) sowie des Hohelied-Kommentars des Ezra von Gerona (13. Jahrhundert) überlieferungs- und traditions-geschichtlich eng verbunden und teilt mit diesen einige Mikroformen.

Entsprechend des in Abschnitt 2 beschriebenen Vorhabens wurden sechs repräsentative Textzeugen für die Spaltensynopse der Printausgabe identifiziert und in LERA digital kollationiert. Aus dem Werkzeug kann eine Datei exportiert werden, die als Grundlage für den manuell verfeinerten Drucksatz genutzt wird. Erweitert um eine ausführliche Kommentierung und Übersetzung entsteht so derzeit die Printausgabe.

Die entstehende digitale Edition auf Basis der LERA-Arbeitsumgebung ergänzt diese Inhalte um die Möglichkeit, die vollständige

Transkription aller Textzeugen einzusehen und automatisch eine eigene Spaltensynopse mit den hervorgehobenen Varianten von bis zu acht frei gewählten Textzeugen generieren zu lassen. Auch hierbei stehen die verschiedenen Visualisierungen zur Analyse sowie diverse Exportformate zur Verfügung.

Förderung

Die Arbeiten wurden durch die Deutsche Forschungsgemeinschaft im Rahmen des Projekts *Synoptische Edition des kabbalistischen Traktats Keter Shem Tov mit englischer Übersetzung, Stellenkommentar und rezeptionsgeschichtlichen Studien* unter der Leitung von apl. Prof. Dr. Gerold Necker, Seminar für Judaistik /Jüdische Studien, und Prof. Dr. Paul Molitor, Institut für Informatik, beide Martin-Luther-Universität Halle-Wittenberg, gefördert. Siehe <https://gepris.dfg.de/gepris/projekt/414786977>.

Fußnoten

1. <https://lera.uzi.uni-halle.de>
2. Dazu gehören Formate für die Presentations als auch für die maschinenlesbare Weiterverarbeitung der Vergleichsergebnisse, wie HTML, LaTeX, PDF, JSON oder XML; letzteres nach den Richtlinien der Text Encoding Initiative (TEI).
3. <https://edinburgh-conan-doyle.org>
4. <https://ispp.zrcsazu.si/izdaje-srece-v-nesreci>
5. <https://www.arendteditionprojekt.de>
6. <https://kabbalaheditions.org>
7. Die Grenze von acht Textzeugen ist dabei eine Empfehlung für eine gute Lesbarkeit und keine fixe Beschränkung, welche der klassischen, spaltensynoptischen Darstellung von LERA Rechnung trägt.
8. Die von Peter Schäfer in der judaistischen Forschung etablierten Kategorien „Makroform“ und „Mikroform“ zur Beschreibung des Phänomens fluktuierender Textüberlieferung und -redaktion in der frühjüdischen Mystik werden im Folgenden aus heuristisch-pragmatischen Gründen übernommen, ohne damit bereits einer noch nicht vorliegenden Texttheorie mittelalterlicher kabbalistischer Texte vorzugreifen (Schäfer, 1981, 15f. und 199-201).

Bibliographie

Bremer Thomas / Molitor, Paul / Pöckelmann, Marcus / Ritter, Jörg / Schütz, Susanne (2015): „Zum Einsatz digitaler Methoden bei der Erstellung und Nutzung genetischer Editionen gedruckter Texte mit verschiedenen Fassungen – Das Fallbeispiel der *Histoire philosophique des deux Indes* von Guillaume Thomas Raynal“. *Internationales Jahrbuch für Editions-wissenschaften*, Hrsg. Rüdiger Nutt-Kofoth und Bodo Pächta, *editio* 29(1):29-51, de Gruyter. DOI: 10.1515/editio-2015-004

Gründler, Beatrice / Pöckelmann, Marcus (2018): „Adjusting LERA for the comparison of Arabic manuscripts of *kalila wa-dimna*“. In: *Proceedings of the 2018 Digital Humanities Conference (DH2018)*, Mexico City, Mexico.

Molitor, Paul / Necker, Gerold / Pöckelmann, Marcus / Reibiger, Bill / Ritter, Jörg (2020): „Keter Shem Tov – Prozessualisierung eines Editionsprojekts mit 100 Textzeugen“. In: *DHd-2020 Book of Abstracts*, Hrsg. Christof Schöch und Patrick Helling. DOI: 105281/zenodo.4621883

Pöckelmann, Marcus / Medek, André / Molitor, Paul / Ritter, Jörg (2015): „CATview – Supporting the investigation of text geneis of large manuscripts by an overall interactive visualization tool“. In: *Proceedings of the 2015 Digital Humanities Conference (DH2015)*, Sydney, Australia.

Pöckelmann, Marcus / Medek, André / Ritter, Jörg / Molitor, Paul (2021): „A user-friendly platform for synoptical representations of multiple text witnesses“. Submitted for publication in *Digital Scholarship in the Humanities*, April 2021.

Roeder, Torsten (2020): „Juxta Web Service, LERA, and Variance Viewer. Web based collation tools for TEI“. *RIDE*, Issue 11: *Tools and Environments*, (Eds.) Anna-Maria Sichani and Elena Spadini. DOI: 10.18716/ride.a.11.5.

Schäfer, Peter (1981): *Hekhalot-Studien*, Tübingen: Mohr Siebeck.

„Arbeitskulturen“ im Wandel Erfahrungen und Entwicklungen in 20 Jahren DH-Praxis

Czmiel, Alexander

czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften,
Germany

Neuber, Frederike

neuber@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften,
Germany

Einleitung

Im ‚Zeitalter des Druckes‘ war die Entstehung des kulturellen Gedächtnisses von Individuen und Autoritäten geprägt, das Buch Ziel und Ergebnis der (scheinbar) finalen Forschung (u.a. Burdick et al. 2012). Im Gegensatz zum traditionellen Forschungsprozess stehen die DH für ‚Teamwork‘ und ‚Kollaboration‘ sowie für eine Art und Weise zu forschen und zu publizieren, die eher ‚prozessorientiert‘ als ‚produktorientiert‘ ist (Griffin u. Haylor 2018: §1, Tabak 2017: §6). Um die skizzierten Paradigmen des digitalen Forschungsprozesses in der Praxis zu verinnerlichen und zu adaptieren, hat TELOTA (kurz für „The Electronic Life Of The Academy“),¹ die DH-Abteilung der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW), in seiner 20-jährigen Geschichte mehrfach die eigenen Arbeitsstrukturen evaluiert und überarbeitet. Ausgehend davon wird der Vortrag eine seit etwa 2020 etablierte Organisationsstruktur, die Team, Projekte und Forschungssoftwareentwicklung umfasst, vorstellen sowie Erfahrungen in der Umsetzung reflektieren.

Da die DH nicht nur für geisteswissenschaftliche Forschung mit digitalen Methoden, sondern auch für einen Wandel des Forschungsprozesses selbst stehen, zählt der Themenbereich ‚Organisation und Workflows‘ zu Kontext und Bedingungsrahmen, in dem das kulturelle Erbe digital erschlossen wird, und bildet rund um die Frage nach DH-spezifischen ‚Arbeitskulturen‘ ein span-

nendes Diskussionsfeld zu „Kulturen des digitalen Gedächtnisses“.

Kontext

In den letzten Jahren gab es auf den DHd-Konferenzen vermehrt Beiträge zu verschiedenen Formen von Arbeitsorganisation in den DH, die vor allem den Aufbau neuer DH-Standorte thematisiert haben (Roeder et al. 2020, 2019a u. 2019b). Entgegen der Annahme, vor allem Standorte, an denen ‚DH from Scratch‘ betrieben wird, sähen sich mit „institutionellen, organisatorischen, personellen und technischen Anforderungen“ (Roeder et al. 2019b) konfrontiert, gilt dies auch für langjährig gewachsene DH-Abteilungen, die zum Teil auf über zwei Dekaden des Bestehens und Arbeitens zurückblicken.² Aus der Phase des Experimentierens herausgewachsen stehen sie vor ganz anderen Herausforderungen: So hat die Etablierung der DH als Methode und Fach im Allgemeinen vor allem im letzten Jahrzehnt einen regelrechten ‚Boom‘ an digitalen Forschungsprojekten ausgelöst. Abgesehen von stetig neuen Projekten haben die Standorte ‚mit Vergangenheit‘ bereits eine ganze Generation von nicht mehr finanzierten Legacy-Projekten in Betrieb zu halten. Das Wachstum an Projekten hat aus ehemals kleinen Standorten mit geringer personeller Dichte und Lab-Charakter in relativ kurzer Zeit große Teams mit vielen und vielfältigen Mitarbeiter/innen gemacht. In der Konsequenz funktionieren althergebrachte, auf wenige Projekte und auf ein kleines Team ausgerichtete Workflows und Arbeitsstrukturen nicht mehr und müssen neu gedacht werden. Orientierungspunkte einer Restrukturierung von Arbeitsprozessen in den DH können dabei u.a. Methoden aus der Softwareentwicklung (Usher et al. 2020, Agile Business Consortium 2014, Rubin 2012) sowie - teilweise davon abgeleitete - DH-spezifische Ansätze aus dem internationalen Kontext bilden (Smithies u. Ciula 2020, Smithies et al. 2019, Ferraro u. Sichani 2018, Tabak 2017, Reed 2014).

Hintergrund: 20 Jahre TELOTA

2021 feiert TELOTA seinen zwanzigsten Geburtstag. Anfänglich eine eher strategische Initiative, begann die Arbeit im Bereich der Softwareentwicklung etwa 2005. TELOTA bestand damals aus einer Leitung, zwei Mitarbeiter/innen und zwei studentischen Hilfskräften und hatte einen stark experimentellen Charakter. Die Phase der technischen Professionalisierung setzte ab etwa 2008 ein, als die Fokussierung auf einzelne Projekte um die Perspektive der projektübergreifenden Standardisierung erweitert wurde (Grötschel u. Neumann 2011). Ab Mitte der 2010er Jahre führte das steigende Interesse der Vorhaben der BBAW an der Umsetzung digitaler Projekte sowie von externen Projektpartnern an der Nachnutzung der von TELOTA entwickelten Werkzeuge zu einem starken Anstieg an neuen Projekten, wodurch sich auch das TELOTA-Team vergrößerte: 2015 bestand das Team noch aus einer Leitung, fünf wissenschaftlichen Mitarbeiter/innen und vier studentischen Hilfskräften. Im Juli 2021 umfasst TELOTA eine Leitung, eine Koordination (beide Autor/innen dieses Beitrags), achtzehn wissenschaftliche Mitarbeiter/innen (die allerdings nicht alle Vollzeitstellen besetzen) sowie drei studentische Hilfskräfte. Zum TELOTA-Portfolio, das Forschungssoftware wie digitale Editionen, Objektsammlungen, Webservices, Frameworks und Tools umfasst, zählen rund 30 aktuell laufende bzw. in der Entwicklung befindliche Projekte sowie ca. 30 Legacy-Projekte, die

verfügbar gehalten und in regelmäßigen Abständen überarbeitet und migriert werden.

Die Arbeitsorganisation bei TELOTA war und ist von jeher vom ressourcentechnischen Umstand geprägt, dass es mehr Projekte als Mitarbeiter/innen gibt, wobei die Differenz kontinuierlich steigt, denn jedes nicht mehr finanzierte Projekt bleibt ein Projekt bzw. eine Publikation, die weiterhin gewartet und verfügbar gehalten werden muss. Dieser Tatsache ist es maßgeblich geschuldet, dass sich die Teamstruktur und Projektbetreuung bei TELOTA im Laufe der Zeit wie folgt entwickelte: Jede/r Mitarbeiter/in arbeitete kontinuierlich und meistens alleine an ein bis drei Akademienvorhaben und ggf. mit weiteren Stellenprozenten in Drittmittelprojekten. Diese Arbeitsweise war für die Entwickler/innen auf Dauer belastend, da die Entwicklungszeit und -konzentration sowohl durch die über mehrere Projekte verteilte Aufmerksamkeit als auch durch die gleichzeitige Übernahme von koordinatorischen Aufgaben beeinträchtigt wurde. Erschwerend kam hinzu, dass Wissen über Projekte oft bei einzelnen Entwickler/innen lag, wodurch der ‚Bus-Faktor‘ (oder auch ‚Truck Factor‘; Williams u. Kessler 2002: 41) gleich „1“ war, was bedeutet, dass jedes Teammitglied über Spezialwissen verfügt und ein Ausfall (z.B. durch Kündigung) das Projekt zumindest für eine gewisse Zeit lahmlegen könnte.

Herausforderungen

Die allgemeinen Herausforderungen der Arbeitsorganisation kann anhand einer vertikalen und einer horizontalen Skalierungsebene skizzieren. Die vertikale Ebene beschreibt die wachsende Komplexität der einzelnen digitalen Projekte von der Datenmodellierung über Datenspeicherung, Programmierschnittstellen und Visualisierungen bis hin zur Publikation, Langzeitverfügbarkeit und -archivierung. Die Kompetenzen, um dieses Aufgabenspektrum abzudecken finden sich nur in den seltensten Fällen in einer Person, so dass es in einem Projekt nicht mehr ausreicht, auf eine/n Digitalspezialisten/in zurückzugreifen, sondern auch hier die Aufgaben auf ein Team verteilt werden müssen.

Die zweite Ebene ist die horizontale, die schlicht die Menge der in den Einrichtungen parallel umzusetzenden digitalen Projekte beschreibt, sowie die Bedarfe und Einzelanforderungen dieser Projekte, die wiederum Einfluss auf die vertikale Ebene haben. Betreut ein DH-Team viele Projekte parallel, können in diesen Projekten nicht alle Anforderungen in einer kurzen Entwicklungszeit umgesetzt werden. Umgekehrt bedeutet dies, dass wenn komplexe Projekte bearbeitet werden, wenig bis keine parallelen Projekte betreut werden können.

Da die Schnittmenge beider Ebenen in den letzten Jahren deutlich gestiegen ist, konnte die oben beschriebene Arbeitsweise (siehe „Hintergrund“) nicht mehr fortgeführt werden. Nachdem verschiedene Modelle der Restrukturierung der Arbeitsorganisation bei TELOTA evaluiert und erprobt wurden, ist seit 2020 ist eine Organisationsstruktur im Einsatz, die ganzheitlich gedacht auf Personen, Projekte und Software ausgerichtet ist.

Teamstruktur

Wie bereits erwähnt waren zentrale Probleme der Teamstruktur, dass es eigentlich kaum Entwicklung in Teams gab, und, dass Personen mit mehreren Rollen belegt waren, zum Beispiel als Entwickler/in und DH-Koordinationsstelle eines Projekts. Die Restrukturierung von TELOTA umfasste daher zunächst einmal die

Einführung einer Koordinationsstelle, deren Aufgaben – in Absprache mit der Leitung – Kommunikation sowie die mittel- und langfristige Entwicklungsplanung umfassen. Die Koordinationsstelle entwickelt und begleitet den neu etablierten Projektworkflow, unterstützt und berät die Projektpartner/innen bei der Planung der DH-spezifischen Ziele, koordiniert die Absprachen mit den Entwicklungsteams und schreitet vermittelnd ein, wenn sich Sackgassen auftun.

Eine weitere Maßnahme der Restrukturierung war die Zuordnung der rund achtzehn Mitarbeiter/innen zu kleineren Teams, die wiederum gemeinsam Cluster aus thematisch und technologisch ähnlichen Projekten bearbeiten, um möglichst große Synergieeffekte zwischen Projekten zu kreieren. Die sich dadurch ergebenden Teams bestehen aus rund 6 bis 8 Personen (mit unterschiedlichen Stellenanteilen), die jeweils zu zweit oder dritt ein Projekt bearbeiten. Während der Entwicklungsphasen (siehe „Projektworkflow“) sind die Entwickler/innen ebenfalls an Planungsaufgaben für die direkt anstehenden Softwareentwicklungsaufgaben beteiligt, werden aber von größeren Aufgaben (z.B. Mitwirkung bei Antragstellungen, strategische Sitzungen u.Ä.) durch die Koordinationsstelle entlastet.

Projektworkflow

Die Entwicklungsarbeiten der oben erwähnten Teams erfolgt seit Anfang 2020 nach dem Modell der ‚Entwicklungsblöcke‘ (Abb. 1). Angelehnt an das Konzept von „Sprints“ aus der Scrum-Methodik (Rubin 2012), handelt es sich dabei um fest definierte Zeiträume, in denen ein Team aus Entwickler/innen ein zuvor geplantes Zwischenziel umsetzt.³ Die Zwischenziele werden iterativ, jeweils vor einem Entwicklungsblock, definiert, stehen mit der grundsätzlichen Roadmap eines Projekts im Einklang, bleiben aber nicht in starren und ausführlichen Vorab-Planungen verhaftet. Im Gegensatz zum klassischen Scrum-Sprint, der nach Features und möglichst kleinteilig konzipiert ist, handelt es sich bei den Entwicklungsblöcken um Arbeitspakete, die eine Menge an Features bzw. Tasks zusammenfassen, welche nach einer gemeinsam bestimmten Priorisierung bearbeitet werden. Durch dieses Vorgehen dauern die Entwicklungsblöcke mit 4-12 Wochen Laufzeit auch wesentlich länger als klassische Scrum-Sprints von etwa 2-4 Wochen.

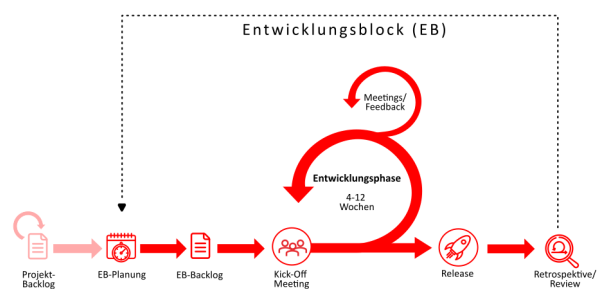


Abb. 1: Ablauf eines Entwicklungsblocks bei TELOTA.

Alle Teilziele, Tasks, Fehler und Featurewünsche zu einem Projekt fließen in ein von allen Beteiligten gepflegtes Projekt-Backlog ein (z.B. ein Issues in GitLab, Tickets in Redmine, oder Zeilen in einem Spreadsheet), aus dem eine Teilmenge im Vorfeld einer

Entwicklungseinheit mit Blick auf die Roadmap des Projekts, die Aufgabenpriorisierung durch die Projektpartner sowie entwicklungstechnische Empfehlungen TELOTAs als Milestone definiert wird. Die eigentliche Entwicklungsphase startet mit einem Kick-Off-Treffen mit allen Mitarbeiter/innen des Projekts, bei dem Ablauf und Aufgaben noch einmal durchgegangen werden. In der Entwicklungsphase kann sich das Entwickler/innenteam voll und ganz auf die definierten Zwischenziele konzentrieren, neue Featurewünsche müssen zunächst mit der Koordinationsstelle, die in etwa der Rolle des „Product Owners“ in Scrum entspricht, abgestimmt werden. Ist die Softwareentwicklung gestartet, so wird regelmäßig Feedback eingeholt, u.a. in fest etablierten *face-to-face*-Meetings.

Die Entwicklungsphase endet mit einem Release der Entwicklungen, der gesamte Entwicklungsblock mit Reviewgesprächen innerhalb des TELOTA-Teams und mit den Projektpartnern. Die Demonstration der neu entstandenen Entwicklungen, die Evaluation der Zusammenarbeit sowie der Blick auf die Projektroadmap bilden die Grundlage für den nächsten Entwicklungsblock.

Auch wenn das klassische Scrum-Modell der „Sprints“ aufgrund der personellen Ressourcen nicht in aller Konsequenz bei TELOTA durchführbar ist,⁴ kann die Forschungssoftwareentwicklung und die parallel erfolgende geisteswissenschaftliche Forschung von einigen Prinzipien der Methode profitieren. Das Konzept von ‚Time-Boxes‘, die intensive Bearbeitungen von Teilzielen und unmittelbare Releases wirken sich beispielsweise motivationssteigernd aus, weil man von Beginn an auf konkrete Resultate bzw. ‚sichtbare‘ Ergebnisse zusteuert. Gleichzeitig werden bei häufigen Releases frühzeitig Probleme erkannt und Kurskorrekturen möglich. Schließlich generieren die regelmäßigen Meetings, der Erfolgsmoment beim Release und die Feedbackkultur ein Gefühl der Zusammengehörigkeit aller Projektbeteiligten, der das „wir“ und „ihr“, das es oftmals in der Zusammenarbeit zwischen Forschungssoftwareentwickler/innen und Geisteswissenschaftlern gibt, auflöst.

Werkzeugkasten

Zentral bei der Neuorganisation der Softwareentwicklung ist die Verständigung auf einen gemeinsamen technologischen Werkzeugkasten, was u.a. durch die aufwendige Wartung vieler technologisch komplexer Legacy-Projekte dringlich wurde. Leitprinzipien des Werkzeugkastens sind u.a. möglichst nachhaltige und keine in ihrer Funktion redundanten Technologien zu verwenden. Unterschieden wird dabei zwischen ‚Primär-‘ und ‚Sekundärtechnologien‘, wobei erstere beispielsweise Programmiersprachen, Datenbanksysteme und umfassende Framework-Plattformen, deren erstmalige Verwendung einen erhöhten Einarbeitungsaufwand bedeutet, umfassen. Die Neueinführung einer Primärtechnologie bedarf daher der Abstimmung mit Leitung, Koordination und Systemadministrator/innen sowie, wenn der Einführung stattgegeben wird, die Durchführung von Schulungen im Team. Unter Sekundärtechnologien fallen Libraries oder kleinere Frameworks, die ggf. auf einer bereits verwendeten Primärtechnologie basieren, leicht zu lernen sind, und deren Verwendung keiner Abstimmung bedarf.

Der Aufbau des Werkzeugkastens ist ‚produktorientiert‘, d.h. die Technologie-Stacks werden ergebnisorientiert vorgeschlagen.⁵ Mittels eines Flow-Charts bzw. Entscheidungsbaums (Abb. 2) wird veranschaulicht, wie man als Nutzer/in an den Werkzeugkasten herantritt. Für jedes Ergebnis ist eine User-Story definiert, die, abgesehen von User Story 1 („der Werkzeugkasten beinhaltet

bereits eine passende Technologie bzw. einen passenden Technologie-Stack“), Szenarien und Prozeduren für die Neueinführung einer Produktkategorie oder Technologie definiert.

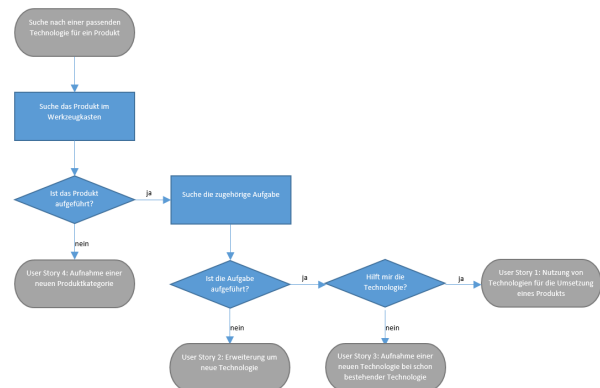


Abb. 2: Der Weg durch den TELOTA-Werkzeugkasten

Ausblick

Jeder DH-Standort agiert und arbeitet unter spezifischen resourcentechnischen, personellen und finanziellen Bedingungen, was es erschwert, von den vorgestellten Überlegungen und Ergebnissen allgemeine Empfehlungen für die Strukturierung von Teams, Projekten und Software zu abzuleiten. Grundsätzlich wird aber deutlich, dass die stärkere Reflexion und Konzeptualisierung der ‚Arbeitskultur‘ nicht nur bei sich im Aufbau befindlichen DH-Zentren, sondern auch bei Standorten mit langer Tradition und eher ‚eingefahrenen‘ Strukturen sinnvoll sein kann. Neben den hier vorgestellten Verfahren gibt es eine Reihe weiterer Punkte, die eine regelmäßige Evaluation und Überarbeitung der eigenen Arbeitsstrukturen sinnvoll erscheinen lassen und die zukünftig in den Blick genommen werden. Dazu gehört u.a. eine weitere Professionalisierung und Spezialisierung der einzelnen Teammitglieder auf bestimmte Technologien bzw. Technologie-Cluster, was neben einer Reduzierung des Risikos von Ausfällen auch zu einer stärkeren Fokussierung der Kompetenzen führt. Darüber hinaus sichert dies die Qualität von Forschungssoftware und damit die Langlebigkeit unseres „digitalen Gedächtnisses“.

Fußnoten

- <https://www.bbaw.de/en/bbaw-digital/telota>
- Zu nennen wären hier beispielsweise das Grazer „Institut Zentrum für Informationsmodellierung“ (<https://informationsmodellierung.uni-graz.at>), das „Cologne Center for e-Humanities“ (<https://ceeh.uni-koeln.de/>) und das Trierer „Kompetenzzentrum - Trier Center for Digital Humanities“ (<https://tcdh.uni-trier.de>).
- Auch Tabak 2017 integriert Scrum-Elemente in das von ihm entwickelte „Hybrid Model for Managing DH Projects“ (2017).
- Beispielsweise sind in den Entwicklungsblöcken TELOTAs die Entwicklungszeiträume länger und die Teams kleiner, wodurch auch die Rolle des „Scrum Masters“ nicht besetzt ist.
- z.B. zur Realisierung einer digitalen Edition, eines Eingabe-tools für strukturierte Daten oder eines Bilderalbums.

Bibliographie

Agile Business Consortium (2014): *The DSDM Agile Project Framework (2014 Onwards)*, <https://www.agilebusiness.org/page/TheDSDMAgileProjectFramework> [letzter Zugriff: 15.7.2021]

Baars, Wouter / Harmsen, Henk / Kramer, Rutger / Sink, Laurents / Zundert, Joris van (2006): *Project management handbook*. Data Archiving and Networked Services, The Hague, <https://www.projectmanagement-training.net/articles-tools/book/> [letzter Zugriff: 15.7.2021]

Burdick, Anne / Drucker, Johanna / Lunenfeld, Peter / Presner, Todd / Schnapp, Jeffrey (2012): *Digital Humanities*. Cambridge: MIT Press.

Ferraro, Ginestra / Sichani, Anna-Maria (2018). „Design as Part of the Plan: Introducing Agile Methodology in Digital Editing Projects“, in: Bleier, Roman / Bürgermeister, Martina / Klug, Helmut / Neuber, Frederike / Schneider Gerlinde (eds.): *Digital Scholarly Editions as Interfaces*, 83-105. Norderstedt: BoD. <https://kups.ub.uni-koeln.de/9113/> [letzter Zugriff: 15.7.2021]

Grötschel, Martin / Neumann, Gerald (2011): „10 Jahre TELOTA“, in: *Jahrbuch 2011 der Berlin-Brandenburgischen Akademie der Wissenschaften*, 202-215. https://edoc.bbaw.de/files/2007/BBAW_Jahrbuch_2011.pdf [letzter Zugriff: 15.7.2021]

Reed, Ashley (2014): „Managing an Established Digital Humanities Project: Principles and Practices from the Twentieth Year of the William Blake Archive“, in: *Digital Humanities Quarterly*, Vol. 8, 1. <http://www.digitalhumanities.org/dhq/vol/8/1/000174/000174.html> [letzter Zugriff: 15.7.2021]

Roeder, Torsten / Cremer, Fabian / Dogunke, Swantje / Elwert, Frederik / Lordick, Harald / Ott, Katrin / Söring, Sibylle / Wübbena, Thorsten (2020): „Digital Humanities from Scratch“ (Workshop), in: Schöch, Christoph (ed.): *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*. 27-29. <http://doi.org/10.5281/zenodo.3666690> [letzter Zugriff: 15.7.2021]

Roeder, Torsten / Söring, Sibylle / Dogunke, Swantje / Elwert, Frederik / Wübbena, Thorsten / Lordick, Harald / Cremer, Fabian / Klammt, Anne (2019a): „Digital Humanities ‚from Scratch‘. Herausforderungen der DH-Koordination zwischen Querschnittsaufgaben und ‚one-(wo)man-show‘“ (Panel), in: Sahle, Patrick (ed.): *DHd 2019 Digital Humanities: multi-medial & multimodal. Konferenzabstracts*. 68-71. <https://doi.org/10.5281/zenodo.2596095> [letzter Zugriff: 15.7.2021]

Roeder, Torsten / Söring, Sibylle / Dogunke, Swantje / Elwert, Frederik / Wübbena, Thorsten / Lordick, Harald / Cremer, Fabian / Klammt, Anne (2019b): „Digital Humanities ‚from Scratch‘. Ein Panel-Bericht zur DHd 2019“ (Blogpost), in: *DHd-Blog*, 3. Juli 2019, <https://dhd-blog.org/?p=11804> [letzter Zugriff: 15.7.2021]

Rubin, Kenneth S. (2012): *Essential Scrum: A Practical Guide to the Most Popular Agile Process*. Addison-Wesley Professional.

Smithies, James / Ciula, Ariana (2020): „Humans in the Loop. Epistemology & Method in King’s Digital Lab“, in: Schuster, Kristen / Dunn, Stuart. (eds.): *Routledge international handbook of research methods in digital humanities*. London: Routledge, 155-172.

Smithies, James / Westling, Carina / Sichani, Anna-Maria / Mellen, Pam / Ciula, Ariana (2019): „Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King’s Digital Lab“, in: *Digital Humanities Quarterly*, 13, 1. <http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html> [letzter Zugriff: 15.7.2021]

[gitalhumanities.org/dhq/vol/13/1/000411/000411.html](http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html) [letzter Zugriff: 15.7.2021]

Spiro, Lisa (2012): „‘This Is Why We Fight’: Defining the Values of the Digital Humanities“, in: Gold, Matthew K.: *Debates in the Digital Humanities*. Minneapolis 2012, S. 16-35.

Tabak, Edin (2017): „A Hybrid Model for Managing DH Projects“, in: *Digital Humanities Quarterly*, 11, 1. <http://www.digitalhumanities.org/dhq/vol/11/1/000284/000284.html> [letzter Zugriff: 15.7.2021]

Usher, Will / Chue Hong, Neil / Darst, Richard / Gonzalez-Beltran, Alejandra / Katz, Daniel S. / Löffler, Frank / Pronk, Thomas / Richmond, Paul / Shad, Mahmood / Stadler, Konstantin / van den Bergh, Erik / van Werkhoven, Ben (2020): „How do RSE groups work?“ A blog post from the 2nd International RSE Leaders Workshop 2020. <https://researchsoftware.org/2020/11/19/how-do-rse-groups-work.html> [letzter Zugriff: 15.7.2021]

Williams, Laurie / Kessler, Robert (2002): *Pair Programming Illuminated*. Boston u. a.: Addison-Wesley Professional.

Aufbau eines Referenzkorpus “Erste Sätze in der deutschsprachigen Literatur”

Busch, Anna

annabusch@uni-potsdam.de

Theodor-Fontane-Archiv, Universität Potsdam

Roeder, Torsten

dh@torstenroeder.de

Bergische Universität Wuppertal, Germany

Stand der Forschung und Problem- aufriss

In literatur- und sprachwissenschaftlichen Untersuchungen ist der erste Satz eines narratologischen Zusammenhangs ein regelmäßig untersuchter Gegenstand (dazu u.a. Alt 2020, Haubrichs 1995, Hirdt 1974, Queng 2019, Miller 1965, Neuhaus 2019, Raulff 2019, Retsch 2000, Selbmann 2019). Das verwundert insofern wenig, gilt doch der erste Satz seit Wolfgang Iser’s Studie *Der Akt des Lesens* als Eingang in den Text durch die Lektüre, als Schlüsselstelle der Interaktion zwischen Text und Leser (1976: 38). Der erste Satz ist Verdichtungspunkt, sinnstiftender Ort für den Fortgang der Erzählung. Er unterliegt Moden, Bedingungen und Abhängigkeiten, bewegt sich in literarischen Traditionen, folgt Wirkungsabsichten, führt Reminiszenzen mit, steht für sich und erzeugt Kontext. Der erste Satz offenbart im Reichtum seiner unterschiedlichen Formen “die Schätze der Literatur in nuce” (Alt 2020: 18) und es ließe sich mit Alain Robbe-Grillet die These aufstellen, dass Literaturgeschichte aus der Untersuchung ihrer Anfangssätze zu schreiben ist (1992: 38).

Eine systematische, digital gestützte Untersuchung von “ersten Sätzen” steht bislang aus. Vereinzelt wurden händisch Korpora erster Sätze zusammengetragen (Beck 1992, Beck 1993, Wolkersdorf 1994) und Versuche unternommen, eine Typologie des ersten Satzes in der Literatur anhand ausgewählter Einzelanalysen zu entwerfen (zuletzt Alt 2020). Ergänzend dazu kann eine systematische Kategorisierung auf der Basis eines semiautomatisiert erstellten, größeren Untersuchungskorpus – wie sie hier projiziert wird – zielführend sein. Ähnlich gelagerte Untersuchungen, die nach der Quintessenz des Poetischen in der Literatur durch ihre Zählbarkeit fragen (vgl. beispielhaft Moretti 2009, auch Fischer/Strötgen 2015, Fischer/Jäschke 2018a/b), liegen vor, eine einzige sich dezidiert mit deutschsprachigen Erzählanfängen (nicht ersten Sätzen) beschäftigende quantifizierende Studie findet sich in der Arbeit von Herrmann 2018.

Der sämtlichen bisherigen Studien zu ersten Sätzen “mangelnden Gesamtsicht” (Alt 2020: 246) zu begegnen, ist Anliegen des Korpus “Erste Sätze in der deutschsprachigen Literatur”. Dazu wird ein Datenkorpus erstellt, publiziert und anschließend in einer Verzahnung von quantitativen und textanalytischen Herangehensweisen eine erste Auswertung unternommen.

Projekt, Vorgehen und Korpus

Als Ausgangsmaterial dienen mehrere Volltextkorpora (Deutsches Textarchiv, Zeno, u.a.), aus denen Texte nach Gattungen extrahiert wurden. Es ist deutlich, dass die vorhandenen Volltextangebote zwar unterschiedlich reichhaltige Strukturinformationen über das jeweilige Dokument bieten, aber die automatische Abgrenzung geschlossener Texteinheiten oft nicht trivial und ohne Einzelprüfung nicht zuverlässig möglich ist (z.B. bei Sammelbänden, Texten mit mehreren Kapiteln, Texte in mehreren Bänden). Dies bildet allerdings die Voraussetzung für das Extrahieren der ersten Sätze. Hinzu kommt, dass der Beginn des “poetischen Texts” durch z.B. vorangestellte Vorworte, Widmungstexte oder Einleitungen automatisiert nicht immer eindeutig zu lokalisieren ist.

Ferner ist die Abgrenzung von “ersten Sätzen” ein semantisches Problem. Sätze lassen sich als grammatisch-analytische Einheiten begreifen, die durch bestimmte Satzzeichen voneinander abgetrennt werden, was der maschinellen Verarbeitung entgegenkommt. Jedoch unterscheiden und verändern sich die zur Abgrenzung eines Satzes verwendeten Zeichen erheblich (man betrachte allein die Entwicklungen zwischen dem 17. und 18. Jahrhundert). Die absolute Trennschärfe mancher Satzzeichen steht zudem kontextabhängig infrage, weshalb Sätze teils auch als Sinneinheiten zu begreifen sind, in denen Satzzeichen eine strukturierende, aber nicht unterbrechende Funktion innewohnt (vgl. Abb. 2a/b). Sollte man also eher von einem fließenden “Beginn” oder “Anfang” sprechen? Bei der Bestimmung der “ersten Sätze” spielen somit Unschärfbereiche hinein, die sich wiederum auf Korpuskonsistenz und -vergleichbarkeit auswirken können.

Auswertung und Reflexion der Ergebnisse

Das derzeitig erstellte Korpus ist vollständig mitsamt Metadaten und Quellenangaben inkl. Positionsangaben in TEI codiert. Gattungsabhängig bewegt sich die Anzahl der Satzanfänge zwischen 100 und 1000 Einträgen. Mithilfe der manuell und automa-

tisch erstellten Annotationen lässt sich das Korpus nach verschiedenen Parametern analysieren und visualisieren, beispielsweise nach Veröffentlichungsdatum, Textgattung, Geschlecht von Verfasserin oder Verfasser, Personen-, Orts- oder Zeitbezüge im Text (vgl. Abb. 1c) oder Länge des Gesamttexts. Außerdem wird dokumentiert, welchen Auswahlkriterien die jeweiligen Datenquellen unterlagen und wie dies im Hinblick auf die Ausgewogenheit des Korpus bei der Auswertung berücksichtigt werden sollte (vgl. Hug/Boenig 2021). Zur Dissemination des Korpus wurde 2021 das Twitter-Projekt “@satzomat” gelauncht, das täglich zwei erste Sätze sendet (vgl. Abbildungen 1–3).

Ziel ist es, eine “Typologie des ersten Satzes” mithilfe computerphilologischer Auswertungsverfahren zu erstellen sowie zu fragen, inwieweit Gattungen im Verlaufe der Geschichte bestimmte Typen von ersten Sätzen determinierten (z.B. Landschaftsbild, Rahmenhandlung) und ob sich weitere Korrelationen mithilfe der Metadaten und Annotationen feststellen lassen.

Abbildungen

”

Gewiß feid ihr alle voll Unruhe, daß ich fo
lange – lange nicht gefchrieben.

*erster Satz aus »Der Sandmann«
von E. T. A. Hoffmann*

”

Es war Sommers-Frühe, die Nachtigallen fangen erft feit
einigen Tagen durch die Straßen, und verftummt heut in
einer kühlen Nacht, welche von fernen Gewittern zu uns
herwehte; der Nachtwächter rief die elfte Stunde an, da fah
ich, nach Haufe gehend, vor der Thür eines großen Gebäudes
einen Trupp von allerlei Gefellen, die vom Biere kamen, um
Jemand, der auf den Thürtufen faß, verflammt.

*erster Satz aus »Gefchichte vom braven Kasperl und dem schönen Annerl«
von Clemens Brentano*

”

Johann Heinrich Ludwig Hanemann wurde im Jahre 1803 in Hoya geboren, fiedelte in einem Alter von 5 Jahren nach dem hannóverfchen Stádtchen Wunstorf im Amt Blumenau, wo er bis zu seiner Konfirmation verblieb, und begab sich dann, als er das Bäckergeschäft erlernt, im Jahre 1819 nach Hamburg.

erster Satz aus »Vom heimatlofen Vaterland«
von Ernst Dronke

”

Weit hinaus im Meer ist das Wasser so blau, wie die Blätter der schönsten Kornblume, und so klar, wie das reinste Glas, aber es ist sehr tief, tiefer als irgend ein Ankertau reicht; viele Kirchtürme müßten auf einander gestellt werden, um vom Boden bis über das Wasser zu reichen.

erster Satz aus »Die kleine Seejungfrau«
von Hans Christian Andersen

Abb. 1a/b/c: Twitter-Grafiken mit Novellen-Anfängen.

”

Wie –? Was –? rief man von allen Seiten.

erster Satz aus »Die neuen Serapionsbrüder«
von Karl Gutzkow

”

Louise naschte gern.

erster Satz aus »Die Nüscherrinnen und das mäßige Kind«
von Karoline Stahl

”

ACh/ ich Unglückfeeliger! was fange ich doch nunmehr an?

erster Satz aus »Der Academische Roman«
von Eberhard Werner Happel

”

Es war einmal ein König, der hatte zwölf Töchter, eine immer schöner als die andere.

erster Satz aus »Die zertanzten Schuhe«
von Jacob und Wilhelm Grimm

Abb. 3a/b/c: Twitter-Grafiken mit Märchen-Anfängen.

”

Berlin fchließ noch, aber es lag in jenem leifen Schlummer, der dem Erwachen vorhergeht.

erster Satz aus »Meister Timpe«
von Max Kretzer

Abb. 2a/b/c: Twitter-Grafiken mit Roman-Anfängen.

Bibliographie

Alt, Peter-André (2020): 'Jemand musste Josef K. verleumdet haben ...' Erste Sätze der Weltliteratur und was sie uns verraten. München: Beck .

Beck, Harald (1992) : Roman-Anfänge. Rund 500 erste Sätze . Zürich: Haffmans.

Beck, Harald (1993) : Romanenden. Rund 500 letzte Sätze . Zürich: Haffmans.

Fischer, Frank / Strötgen, Jannik (2015): "Wann findet die deutsche Literatur statt? – Zur Untersuchung von Zeitausdrücken in großen Korpora." Presented at the DHd2015 Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation. 2. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd2015), Graz: Zenodo. <http://doi.org/10.5281/zenodo.4623384> [letzter Zugriff: 6. Juli 2021]

Fischer, Frank / Jäschke, Robert (2018a): "Liebe und Tod in der Deutschen Nationalbibliothek. Der DNB-Katalog als Forschungsobjekt der digitalen Literaturwissenschaft." Presented at the DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2018), Köln: Zenodo. <http://doi.org/10.5281/zenodo.4622376> [letzter Zugriff: 9. Juli 2021]

Fischer, Frank / Jäschke, Robert (2018b): "Ein Quantum Literatur. Empirische Daten zu einer Theorie des literarischen Textumfangs." DFG-Symposium "Digitale Literaturwissenschaft". Villa Vigoni, 9.–13. Oktober 2017. [noch unveröffentlicht]

Haubrichs, Wolfgang (1995): "Kleine Bibliographie zu "Anfang" und "Ende" in narrativen Texten (seit 1965)", in: *Zeitschrift für Literaturwissenschaft und Linguistik* 25, 99: 36-50.

Herrmann, Berenike (2018): "Anschaulichkeit messen. Eine quantitative Metaphernanalyse an deutschsprachigen Erzählanfängen zwischen 1880 und 1926", in: Köppe, Tilmann / Singer, Rüdiger (eds.): *Show, don't tell: Konzepte und Strategien anschaulichen Erzählens*. Bielefeld: Aisthesis 167-212.

Hirdt, Willi (1974): "Incipit. Zu einer Poetik des Romananfanges", in: *Romanische Forschungen* LXXXVI: 419-436.

Hug, Marius / Boenig, Matthias (2021): *Die Geschichte der Digitalen Bibliothek, oder: Aller guten Kurationen sind drei+* <https://sprache.hypotheses.org/2436> [letzter Zugriff: 6. Juli 2021]

Iser, Wolfgang (1976): *Der Akt des Lesens. Theorie ästhetischer Wirkung*. München: Fink.

Miller, Norbert (1965): *Romananfänge. Versuch zu einer Poetik des Romans*. Berlin: Verl. Literarisches Colloquium.

Moretti, Franco (2009): "Style, Inc Reflections on Seven Thousand Titles (British Novels, 1740-1850)", in: *Critical Inquiry* 36, I: 134-158.

Neuhaus, Stefan (2019): "'Aber wehe, wehe, wehe! Wenn ich auf das Ende sehe!'" Wie in Romanen und Erzählungen durch Anfang und Ende ein Rahmen erzeugt wird", in: Neuhaus, Stefan / Weber, Petra (eds.): *Anfangen und Aufhören*. Paderborn: Wilhelm Fink 141-157.

Queng, Jesse (2019): "Syntaktische Strukturen als poetologisches Mittel des Anfangens in der Prosa: Der erste Satz von Heinrich Bölls Irischem Tagebuch", in: Neuhaus, Stefan / Weber, Petra (eds.): *Anfangen und Aufhören*. Paderborn: Wilhelm Fink 89-101.

Raulff, Ulrich (2019): "Letzte Sätze", in: *Zeitschrift für Ideengeschichte* 13: 129-142.

Reutsch, Annette (2000): *Paratext und Textanfang*. Würzburg: Königshausen & Neumann.

Richardson, Brian (2008): *Narrative Beginnings: Theories and Practices*. University of Nebraska Press.

Robbe-Grillet, Alain (1992): "Warum und für wen schreibe ich", in: Bühler, Karl Alfred (ed.): *Robbe-Grillet zwischen Moderne und Postmoderne - "nouveau roman", "nouveau cinéma" und "nouvelle autobiographie"*. Tübingen: Narr.

Selbmann, Rolf (2019): "Lauter erste Sätze", in: Neuhaus, Stefan / Weber, Petra (eds.): *Anfangen und Aufhören*. Paderborn: Wilhelm Fink 67-87.

Wolkersdorfer, Andreas (1994): *Der erste Satz. Österreichische Romananfänge 1960-1980*. Wien: WUV Univ.-Verl.

Berlin's Australian Archive Eine virtuelle Forschungsumgebung für naturkundliche Sammlungen aus den australischen Kolonien

Bischoff, Eva

bischoff@uni-trier.de
Universität Trier, Germany

Schwarz, Anja

anja.schwarz@uni-potsdam.de
Universität Potsdam, Germany

Naturkundliches Sammeln ging spätestens seit dem 18. Jahrhundert Hand in Hand mit der Katalogisierung und Kategorisierung entlang wissenschaftlicher Taxonomien (Bennett 2004). Diese besondere Form der Informationsverarbeitung entwickelte unterschiedliche Strategien und bediente sich verschiedener Technologien zur „Inventarisierung der Natur“ (Nadim 2016) sowie der Organisation und Speicherung der 'im Feld' gewonnenen Informationen: Bestandslisten, Zettelkästen, Datenbanken. Lokalisiert in Forschungsinstitutionen und Museen des Globalen Nordens, war der Zugang zu den in ihnen bewahrten Informationen beschränkt auf diejenigen, die physisch Zugang zu diesen Räumen erlangen konnten. Die solcherart entstandenen Wissensbestände bilden bis heute einen wesentlichen Bestandteil des kulturellen, ökonomischen und politischen Kapitals von Sammlungsinstitutionen des Globalen Nordens. Gleichzeitig sind diese Informationsasymmetrien vielfach Gegenstand post- und dekolonialer Kritik (Bennett 2004; Schiebinger 2004, Subramaniam 2014). Oft wird dabei die Forderung nach einer umfassenden Digitalisierung der Verzeichnisse und der Veröffentlichung von Datenbeständen erhoben. Entsprechende Veränderungen der Informationsethik lassen sich u.a. in internationalen Regelwerken wie der Biodiversitätskonvention nachweisen, welche seit 1992 für genetische Ressourcen den rechtlich verbindlichen Rahmen für das sog. „Access and Benefit-Sharing“ setzt. Für das in naturkundlichen Sammlungen archivierte kulturelle Erbe hingegen steckt diese Entwicklung noch in den Kinderschuhen.

Ausgehend von diesen Überlegungen und basierend auf unserer aktuellen Entwicklungsarbeit an einer FUD-basierten virtuellen Forschungsumgebung zu naturkundlichen Sammlungen aus Australien am Berliner Museum für Naturkunde, MfN (Bischoff und Schwarz 2020a, 2020b) thematisiert das Poster sowohl die Potentiale als auch mögliche Schwierigkeiten digitaler Zugänge zum kulturellen Erbe. Wenn naturkundliche Sammlungen wie in unserem Fall in kolonialen Kontexten erworben wurden (Antonelli 2020; Das & Lowe 2018; NatSCA 2020) und kulturell sensible Sammlungsgegenstände umfassen (Berner et. al 2011; German Museums Association 2021), müssen Auf- und Ausbau, sowie die Nutzung digitaler Archive eine Reihe von Problemstellungen adressieren. Diese beinhalten u.a. epistemische Konflikte um Metadatenstandards, Aushandlungen über Möglichkeiten des Zugangs aber auch der Restriktion des Zugriffs für unterschiedlich verfasste Öffentlichkeiten, sowie die Interpretation und kulturell autarke Ergänzung von Forschungsdaten durch Mitglieder der Herkunftsgesellschaften bis hin zur Begleitung von Prozessen der virtuellen Repatriierung.

Unterstützt durch das Servicezentrum eSciences der Universität Trier und den Arbeitsbereich Humanities of Nature am MfN greift unsere Arbeit Impulse aus gegenwärtigen Diskussionen zu kolonialen Traditionen naturkundlicher Klassifikation und ihrer Fortschreibung in digitalen Metadaten auf (Agrawal 2002; Boamah & Liew 2017; Briggs et al. 2020; Sarkhel 2016; Stevens 2008; Van der Velden 2010); wir lernen von Bemühungen zum Schutz indigenen Wissens in digitalen Kontexten und um indigene Datensouveränität (Anderson & Christen 2013; Christen 2018; Geismar 2013a/b; Kapepiso et al. 2020; Walter & Suina 2019); und wir orientieren uns an Diskussionen zur Bedeutung von digitaler Autonomie für eine zukunftsorientierte kulturelle Selbstbestimmung (Genovese 2016; Liew et al. forthcoming; McKemmish et al. 2011; Roy 2015; Thorpe 2016). Unser Projekt nimmt dabei insbesondere Erfahrungen aus dem australischen Kontext auf (Christen 2008; Christie 2004; Janke 2018; McGinnis 2020) und adaptiert diese im engen Austausch mit indigenen Kurator:innen und anderen wissenschaftlichen Mitarbeiter:innen des Australian Museums in Sydney und den Museums Victoria aus Melbourne für unsere virtuelle Forschungsumgebung zu kolonialen australischen Sammlungen in Deutschland. Konkrete Richtlinien und Anregungen für dieses Vorhaben finden wir in aktuellen Guidelines und Protokollen australischer Fachgesellschaften (ATSILIRN 2012; AIATSI 2020; EGIM 2019; Janke et al 2019) und dem für uns vorbildlichen Datenbankprojekt "Return Reconcile Renew".

Am konkreten Beispiel der in der Forschungsumgebung hinterlegten Digitalisate der südost-australischen Sammlungsbestände des preußischen Naturkundlers Wilhelm Blandowski (1822-1872) diskutiert unser Poster die Problemstellungen, mit denen unser Projekt regelmäßig befasst ist, die Prozesse und Entscheidungen, die wir bisher durchlaufen haben sowie deren konkrete Umsetzung in der Forschungsumgebung.

Bibliographie

Aboriginal and Torres Strait Islander Library, Information and Resource Network (ATSILIRN) (2012): *Aboriginal Torres Strait Islander Protocols for Libraries, Archives and Information Services* <https://atsilirn.aiatsis.gov.au/protocols.php> [letzter Zugriff 10. Mai 2021].

Agrawal, Arun (2002): "Indigenous Knowledge and the Politics of Classification", in: *International Social Science Journal* 54: 287–297.

Anderson, Jane/Christen, Kim (2013): "Chuck a Copyright on it". Dilemmas of Digital Return and the Possibilities for Traditional Knowledge Licenses and Labels" in: *Museum Anthropology Review* 7: 105–126.

Antonelli, Alexandre (2020): "It's Time to Decolonise Botanical Collections", *Royal Botanic Gardens Kew* <https://www.kew.org/read-and-watch/time-to-decolonise-botanical-collections> [letzter Zugriff 24. Mai 2021].

Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSI) (2020): *AIATSI Code of Ethics for Aboriginal and Torres Strait Islander Research* <https://apo.org.au/node/308966> [letzter Zugriff 14. Juli 2021].

Bennett, Tony (2004): *Pasts beyond Memory: Evolution Museums Colonialism. Museum Meanings*. London: Routledge.

Berner, Margit, et al. (eds.) (2011): *Sensible Sammlungen. Aus dem anthropologischen Depot* (= Fundus 210), Hamburg: Philo & Philo Fine Arts.

Bischoff, Eva/ Schwarz, Anja (2020a): *Wissen um Welt — Umweltwissen: Deutsche Naturkunde in Australien* <https://www.chest.uni-trier.de/projekte/wissen-um-welt-um->

weltwissen-deutsche-naturkunde-in-australien [letzter Zugriff 13. Juli 2021].

Bischoff, Eva/ Schwarz, Anja (2020b): *Collecting a Continent Reconstructing the Australian Archive of Berlin's Natural History Museum* <https://collectingoz.hypotheses.org> [letzter Zugriff 13. Juli 2021].

Boamah, Eric / Li, Chern Liew (2017): "Conceptualising the Digitisation and Preservation of Indigenous Knowledge. The Importance of Attitudes" in: Choemprayong, Songphan/ Crestani, / Cunningham, Sally Jo (eds.): *Digital Libraries. Data, Information and Knowledge for Digital Lives*. Cham: Springer 65-80.

Briggs, Carolyn et al. (2020): "Bridging the Geospatial Gap. Data about Space and Indigenous Knowledge of Place", in: *Geography Compass* 14: 1-17.

Christen, Kimberly (2008): "Ara Irititja: Protecting the Past, Accessing the Future—Indigenous Memories in a Digital Age. A Digital Archive Project of the Pitjantjatjara Council", in: *Museum Anthropology* 29: 56-60.

Christen, Kimberly (2018): "Relationships, Not Records. Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online", in: Sayers, Jentery (ed.): *The Routledge Companion to Media Studies and Digital Humanities*. London: Routledge 403-412.

Christie, Michael (2004): "Computer Databases and Aboriginal Knowledge", in: *Learning Communities: International Journal of Learning in Social Contexts* 1: 4-12.

Das, Subhadra/ Lowe, Miranda (2018): "Nature Read in Black and White. Decolonial Approaches to Interpreting Natural History Collections", in: *Journal of Natural Science Collections* 6: 4-14.

Expert Group on Indigenous Matters (EGIM) (2019): "Tandanya – Adelaide Declaration", Declaration by the EGIM of the International Council on Archives (ICA) <https://www.ica.org/en/egim-tandanya-adelaide-declaration> [letzter Zugriff 14. Juli 2021].

Geismar, Haidy (2013): "Defining the Digital", in: *Museum Anthropology Review* 7: 254-263.

Geismar, Haidy (2013b): *Treasured Possessions: Indigenous Interventions into Cultural and Intellectual Property*. Durham: Duke University Press.

Genovese, Taylor R. (2016): "Decolonizing Archival Methodology. Combating Hegemony and Moving towards a Collaborative Archival Environment", in: *AlterNative* 12: 32–42.

German Museum Association (2021): *Guidelines for German Museums. Guidelines for the Care of Collections from Colonial Contexts* <https://www.museumbund.de/publikationen/guidelines-on-dealing-with-collections-from-colonial-contexts-2> [letzter Zugriff 14. Juli 2021].

Janke, Terri et al. (2018): *Indigenous Knowledge. Issues for Protection and Management*. Discussion Paper, Commissioned by IP Australia & the Department of Industry, Innovation and Science, 21-23, 65-70, 106 <https://www.ipaustralia.gov.au/about-us/news-and-community/news/indigenous-knowledge-issues-protection-and-management> [letzter Zugriff 14. Juli 2021].

Janke, Terri et al. (2019): *First Peoples. A Roadmap for Enhancing Indigenous Engagement in Museums and Galleries* <https://www.amaga.org.au/shop/first-peoples-roadmap-enhancing-indigenous-engagement-museums-and-galleries-hardcopy-version> [letzter Zugriff 14. Juli 2021].

Kapepiso, Fabian Simasiku/ Higgs, Richard (2020): "Tracing the Curation of Indigenous Knowledge in a Biopiracy Case", in: *AlterNative* 16: 38-44.

Liew, Chern Li et al. (forthcoming): “Digitized Indigenous Knowledge Collections. Impact on Cultural Knowledge Transmissions, Social Connections, and Cultural Identity”, in: *Journal of the Association for Information Science and Technology*.

McGinnis, Gabrielle et al. (2020): “Indigenous Knowledge Sharing in Northern Australia. Engaging Digital Technology for Cultural Interpretation”, in: *Tourism Planning & Development* 17: 96-125.

McKemmish, Sue et al. (2011): “Distrust in the Archive. Reconciling Records”, in: *Archival Science* 11: 211-239.

Nadim, Tahini (2016): “Biodiversität erfassen: von Suppen und Satelliten”, in: André Blum/ Nina Zschocke/ Hans-Jörg Rheinberger/ Vincent Barras (eds.): *Diversität: Geschichte und Aktualität eines Konzepts*. Würzburg: Königshausen & Neumann 61-84.

Natural Sciences Collections Association (NatSCA) (2020): “Decolonising Natural Science Collections” <https://www.natsca.org/natsca-decolonising> [letzter Zugriff 24. Mai 2021].

Return Reconcile Renew <https://returnreconcilerenew.info/ohrm/index.html> [letzter Zugriff 14. Juli 2021].

Roy, Lorienne (2015): “Indigenous Cultural Heritage Preservation. A Review Essay with Ideas for the Future”, in: *International Federation of Library Associations and Institutions* 41: 192-203.

Sarkhel, Juran Krishna (2016): “Strategies of Indigenous Knowledge Management in Libraries”, in: *Qualitative and Quantitative Methods in Libraries* 5: 427-439.

Schiebinger, Londa (2004): *Plants and Empire. Colonial Bio-prospecting in the Atlantic World*. Cambridge: Harvard University Press.

Stevens, Amanda (2008): “A Different Way of Knowing. Tools and Strategies for Managing Indigenous Knowledge”, in: *Libri* 58: 25-33.

Subramaniam, Banu (2014): *Ghost Stories for Darwin. The Science of Variation and the Politics of Diversity*. Urbana: University of Illinois Press.

Thorpe, Kirsten et al. (2016): “Discovering Indigenous Australian Culture: Building Trusted Engagements in Online Environments”, in: *Journal of Web Librarianship* 10: 343-363.

Van der Velden, Maja (2010): “Design for the Contact Zone”, in: Sudweeks, Fay/ Hrachovec, Herbert/ Ess, Charles (eds.): *Proceedings Cultural Attitudes toward Communication and Technology. Proceedings of the Fifth International Conference on Cultural Attitudes Towards Technology and Communication*, Tartu, Estonia (28 June-1 July 2006), Murdoch: Murdoch University 1-18.

Walter, Maggie/ Suina, Michele (2019): “Indigenous Data, Indigenous Methodologies and Indigenous Data Sovereignty”, in: *International Journal of Social Research Methodologies* 22: 233-243.

Beyond Budweiser Creating a Digital Archive of Popular German-American Newspaper Literature

Keck, Jana

keck@ghi-dc.org
German Historical Institute Washington DC

Blessing, Andre

andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle
Sprachverarbeitung

Who gets to be remembered and historicized by ways of – digital – record creation? For many German-Americans in the long nineteenth century, German-language newspapers were the primary source of both information and entertainment. So far, research on the German-American press has predominantly focused on the male editors, writers, or advertisers and their influence – and success stories – on U.S. politics and economy. These HISTORIES have entered into schoolbooks and popular culture in the U.S. and Germany alike. Idealized versions, for instance, of the nineteenth-century German-American migrant as a hard-working, bright, self-made man have been “reused” for marketing strategies as the 2017 Budweiser Super Bowl commercial illustrates. The commercial video ad titled “Born the Hard Way” tells the fictitious tale of Adolphus Busch, co-founder of the brewery dynasty Anheuser-Busch, who emigrated to the U.S. in 1857, and ends with the slogan: “when nothing stops your dream.”¹

Such representations offer limited access to histories about the everyday life of other historical actors that go beyond the elite. Digitized historic German-language newspapers in the *Chronicling America* database² seem to provide a fruitful platform to find unknown stories that shed more light onto the daily experiences of historical actors of migration such as, for instance, women, girls, mothers, or daughters. Even though, such digitization projects offer access to thousands of newspapers pages (cf. Soni et al. 2021), there are no innovate methods to search and analyze them (cf. Hausdewell et al. 2020). Which keywords to enter when one does not even know what they are looking for?

In order to systematically rewrite histories about representations of marginalized groups in the German-American press, we are creating an expanded version of *Chronicling America's* repository: using OMEKA S,³ a free, flexible, and open-source web-publishing platform, users will have the opportunity to access news content that was not only published once, but reprinted several times across states and decades (see fig 1). The dataset of reprinted texts was created with text reuse detection software (Smith et al. 2013).⁴

Viral Texts in C19 German-American Newspapers

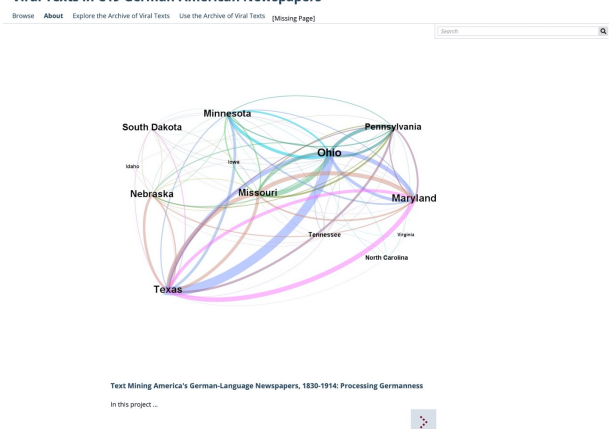


Fig. 1: Sample homepage (OMeka S) of the dataset of reprinted texts in nineteenth-century German-American newspapers (in progress).

With this method, we have uncovered approximately 500,000 viral texts, in Ryan Cordell's words, who uses the social media metaphor to describe reprinting practices in the industrial age.⁵ These viral texts were not only hard news. They range from advertisements, or factual texts to poems. However, these texts are not yet categorized into different genres. To add another layer that will make an expanded search possible by adding genre as metadata, we are using unsupervised methods (*topic models* , *clustering*) and supervised classification methods enabled by manual annotations that were integrated into a web-based interface, an adaptation of the DFR-Browser,⁶ and extended by an annotation module.



Fig. 2: DFR-Browser extended by an annotation module to tag genres

To annotate data, which is then being used as training data for the genre classifier, we have added a sample selection of documents. Such a step allows us to annotate texts, which have a high and a low likelihood of representing a specific topic.⁷ As figure 2 illustrates, computationally classifying genres does not mean that a text will be 100 % categorized as belonging to one type, but in varying degrees to several types. The genre which achieves the highest score will be decisive for the dataset published in OMEKA S. With this approach, we have identified 10 different newspaper genres. By linking the annotation interface with the OMEKA site, users will be able to gain insight into genre classification process and examine similarities and anomalies between texts and genres using a mixed-methods approach (Sá Pereira 2019). Additionally, scholars can always get redirected to *Chronicling America* and to examine, for instance, where the text was embedded in the newspaper page.

In nineteenth-century newspaper ads, women were not only used as marketing strategies for medical products to cure female weakness, but predominantly for products marketed to both sexes (Keck 2021). By linking different datasets and interfaces, our project shows how we can efficiently use data as “a check-in, (...) a resource to begin and continue dialogue”⁸ in gender studies. Only accessible datasets can be passed on to future generations. Adding the category of genre as metadata, provides a distinct approach to simply using keyword search, which requires prior – often biased – knowledge of the user. As Temi Odumosu proposes, we should see data and metadata as ways to rethink cataloguing spaces with the potential to alter historical imbalances of power (2020: 299). Machines can help in this way because they approach data differently: the algorithms used for text reuse detection and text classification do neither privilege specific writers, topics or groups.

Footnotes

1. <https://www.youtube.com/watch?v=IZaQQvflfPQ>
2. For details, see *Chronicling America*'s “About” page.
3. <https://omeka.org/s/>
4. See <https://github.com/dasmiq/passim.git>.
5. <https://viraltexts.org>
6. <https://agoldst.github.io/dfr-browser/>
7. For a critical reflection on the use of topic modeling in the humanities, see Shadrova (2021).
8. <https://www.manifestno.com>

Bibliography

Hauswedell, Tessa / Nyhan, Julianne / Beals, Melodee / Terras, Melissa / Bell, Emily (2020): "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers". In: *Archival Science* 20, 139-165. DOI: <https://doi.org/10.1007/s10502-020-09332-1>.

Keck, Jana (2021): "Let's talk data, bias, and menstrual cramps: Voicing Gerwomanness in the nineteenth century and today". In: *Bulletin of the German Historical Institute (Spring 2021)*. https://www.ghi-dc.org/fileadmin/publications/Bulletin/bu68/bu68_61.pdf

Odumosu, Temi (2020): "The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons". In: *Current Anthropology*, vol. 61. DOI: 10.1086/710062.

Sá Pereira, Moacir P. de (2019): "Mixed Methodological Digital Humanities". In: *Debates in the Digital Humanities*, chapter 34. DOI: <https://doi.org/10.5749/9781452963785>.

Shadrova, Anna (2021). "Topic models do not model topics: epistemological remarks and steps towards best practices". In: *Journal of Data Mining and Digital Humanities*, Episciences.org, 2021. DOI: <https://doi.org/10.46298/jdmdh.7595>.

Smith, David A. / Cordell, Ryan / Maddock Dillon, Elisabeth (2013). "Infectious Texts: Modelling Text Reuse in Nineteenth-Century Newspapers". In: *Proceedings of the Workshop on Big Humanities*, 86–94. Washington, DC: IEEE Computer Society Press.

Soni, Sandeep / Klein, Lauren F. / Eisenstein, Jacob (2021): "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers". In: *Journal of Cultural Analytics*, 43 (1). Doi: 10.22148/001c.18841.

Beyond the render silo
Semantically annotating 3D data
within an integrated knowledge
graph and 3D-rendering toolchain

Rossenova, Lozana

lozana.rossenova@tib.eu

TIB – Leibniz Information Centre for Science and Technology

Schubert, Zoe

Zoe.Schubert@sbb.spk-berlin.de
SPK – Stiftung Preussischer Kulturbesitz

Vock, Richard

vock@cs.uni-bonn.de
Bonn University

Blümel, Ina

Ina.Blumel@tib.eu
TIB – Leibniz Information Centre for Science and Technology,
Hannover University of Applied Sciences and Arts

Research problem

The proliferation of improved scanning technologies and digital record-keeping systems led to mass-digitisation efforts and the launch of numerous online archives and collections (Terras, 2011). Cultural heritage and research institutions have had to adapt their practices to account for shifts in what contemporary cultural stewardship and the study of cultural memory entails (Parry, 2010). Collections of images and texts need to be fully accessible to fulfil institutional missions (Kapsalis, 2016; Maher and Tallon, 2018). In the vast majority of cases, access implies viewing siloed resources, such as scanned text pages or photographs of physical objects and sometimes 3D renderings, alongside minimal descriptive metadata. But digital representations of cultural assets in the form of 3D models within disciplines such as architecture, art, archaeology, and 3D reconstruction are particularly heterogeneous in formats and structure, ergo standardized access and visualisation tools fail to meet new research objectives and institutional requirements. Especially as the result of state-of-the-art digitisation efforts, 3D datasets challenge renderers in terms of geometric complexity, memory and bandwidth requirements (Koller et al., 2009). What is more, cultural memory preservation is not guaranteed through digitisation activities alone, but instead requires the active participation of diverse audiences who can search, access and enrich datasets through annotation and critical interpretation.

Methodology

To address this knowledge gap, a suite of tools is being developed as part of the NFDI4Culture project across several partner organisations. Operating within *Task area 1: Data capture and enrichment*, the proposed toolchain focuses on the annotation of 3D data within a knowledge graph environment, so that 3D objects' geometry, attendant metadata, as well as annotations remain searchable, while data interconnections are not lost. The project builds on several existing FOSS tools:

- OpenRefine, a data cleaning, reconciliation and batch upload tool (Sternier 2019);
- Wikibase, a suite of services developed by Wikimedia Germany; Wikibase is the software behind Wikidata, the largest public knowledge graph on the web; it combines the ability to handle large volumes of data points with sophisticated data querying and extraction services via a dedicated SPARQL endpoint (Thornton et al. 2018);

- Kompakkt, a browser-based open-source 3D and multimedia viewer Kompakkt with built-in collaborative annotation features (Eide et al., 2019).

The toolchain implementation is taking place in an open, iterative process in close collaboration with the culture community to leverage transparency, reduce duplication of effort and ensure optimal usability. The first milestone of establishing a mechanism for automated deployment of reconciliation services between OpenRefine and an arbitrary, self-managed instance of Wikibase has been completed via Ansible and CI/CD pipelines on GitLab. The next milestone of gathering sample contextual metadata alongside images, videos and related 3D objects is being completed in collaboration with the Corpus der barocken Deckenmalerei in Deutschland (Bayerische Akademie der Wissenschaften, 2021) and research partners at TUM. Once all relevant data is modelled and uploaded in Wikibase's knowledge graph and linked to persistent storage of the media files, the next milestone is linking the graph database with Kompakkt's rendering and user interaction environment, which will act as end-user's entry point to exploring the datasets and annotating images and 3D models with high level of precision. In parallel to the work on the Wikibase-Kompakkt integration, we are extending Kompakkt's current capabilities, so that we can facilitate collaborative annotation of large-scale 3D pointclouds and meshes with the same ease and efficiency as the currently supported smaller mesh datasets by providing an alternate FOSS pointcloud rendering backend.

The integrated suite of tools follows FAIR principles, and adopts common standards like PIDs or the W3C annotation model. It facilitates linking 3D objects and annotations, and their cultural context (including historical people and places, geo-location and capture-technology metadata), to the broader semantic web and various national and international authority records (GND, Getty's AAT, VIAF and more).

Results

By the end of 2021, the toolchain will be developed as an MVP (minimum viable product) to be tested and refined further with more data partnerships. It will allow a wide range of users to interact with 3D and other types of multimedia objects and annotations, and ultimately open up new digital spaces for research, education and discourse around cultural stewardship and memory preservation without siloing knowledge. The proposed toolchain will extend the potentials of Wikimedia and Europeana GLAM initiatives, while remaining highly flexible and adaptable to individual institutional contexts and research needs. Results – in the form of open-source code repositories, workflow documentation and a public portal for requirements gathering (GitLab, 2021) – are being made available in alignment with community needs.

Bibliography

- Bayerische Akademie der Wissenschaften. (2021). *Corpus der barocken Deckenmalerei in Deutschland (CbDD)*. <https://deckenmalerei.badw.de/> (accessed 15 July 2021).
- Eide, Ø., Schubert, Z., Türkoğlu, E., Wieners, J.G. and Niebes, K. (2019). *The intangibility of tangible objects: re-telling artefact stories through spatial multimedia annotations and 3D objects*. Presented at the ICOM Kyoto 2019, 25th ICOM General Conference: Museums as Cultural Hubs: The Future of Tradition,

Kyoto. <http://doi.org/10.5281/zenodo.3878966> (accessed 15 July 2021).

Gitlab. (2021). *MVP requirements gathering portal*. https://gitlab.com/nfdi4culture-data-enrichment/kompakt-wikibase-integration/-/requirements_management/requirements (accessed 15 July 2021).

Kapsalis, E. (2016). *The impact of open access on galleries, libraries, museums, & archives*, http://siarchives.si.edu/sites/default/files/pdfs/2_016_03_10_OpenCollections_Public.pdf (accessed 15 July 2021).

Koller, D., Frischer, B. and Humphreys, G. (2009). *Research challenges for digital archives of 3D cultural heritage models*, *JOCCH* 2, (7): 10.1145/1658346.1658347.

Maher, K. and Tallon, L. (2018). *Wikimedia and The Met: A shared digital vision*, Wikimedia Blog. <https://blog.wikimedia.org/2018/04/19/wikimedia-the-met-shared-digital-vision/> (accessed 15 July 2021).

Parry, R. (ed). (2010). *Museums in a Digital Age*. Abingdon: Routledge.

Sterner, E. (2019). "Cleaning Collections Data Using OpenRefine", *Issues in Science and Technology Librarianship*, 92. <https://doi.org/10.29173/istl30> (accessed 15 July 2021).

Terras, M. (2011). "The rise of digitization". In: Rikowski, R. (ed.), *Digitisation Perspectives*. Rotterdam: Sense Publishers, pp. 3–20.

Thornton, K., Seals-nutt, K., Cochrane, E. and Wilson, C. (2018). "Wikidata for Digital Preservation". In: *Proceedings of iPRES'18, Cambridge, MA, USA, September 24–27, 2018*.

Brücken bauen für Buddha Das Projekt „Digitalisierung Gandharischer Artefakte“ (DiGA) und die Pelagios Working Group „Linked Data Methodologies in Gandharan Buddhist Art and Texts“

Elwert, Frederik

frederik.elwert@rub.de
Ruhr-Universität Bochum, Germany

Pons, Jessie

jessie.pons@rub.de
Ruhr-Universität Bochum, Germany

Das Projekt „Digitalisierung Gandharischer Artefakte“ (DiGA) digitalisiert und erschließt ein Korpus von 1.791 buddhistischen Skulpturen, die derzeit im Dir Museum in Chakdara und im Missionshaus der Missione Archeologica Italiana in Pakistan (MAIP) in Saidu Sharif (Provinz Khyber-Pakhtunkhwa, Pakistan) aufbewahrt werden.¹ Dabei handelt es sich um Statuen des Buddha, der Bodhisattvas, der Schutzgottheiten und der Stifter sowie um narrative Reliefs, die Ereignisse aus den vorherigen und dem letzten Leben des Buddha Siddhārtha Gautama darstellen.

Diese Sammlungen sind außergewöhnlich, weil der archäologische Kontext der Objekte dokumentiert ist. Damit unterscheiden sie sich von vielen anderen Sammlungen buddhistischer Kunst aus

Gandhara, deren Provenienz häufig unklar ist. Die in Chakdara und Saidu Sharif aufbewahrten Artefakte stammen von 13 alten buddhistischen Stätten, die sich im Gebiet des Flusses Swat befinden. Die Objekte, die DiGA digitalisiert, wurden bei wissenschaftlichen Ausgrabungen entdeckt, die von der pakistanischen Regierung, der Universität Peshawar und dem MAIP Ende der 1960er und in den 1990er Jahren durchgeführt wurden.

Die Objekte werden von allen Seiten mittels Digitalfotografie dokumentiert. Ausgewählte Objekte, die aufgrund ihrer Charakteristika oder ihrer Repräsentativität von einer räumlichen Erfassung besonders profitieren, werden zudem mittels Fotogrammetrie in 3D digitalisiert. Die Digitalisate und ihre Metadaten werden im Rahmen einer Kooperation mit dem FID Südasiens in der Mediendatenbank heidICON der Universitätsbibliothek Heidelberg erfasst und für die Nachnutzung bereitgestellt. Das Metadaten-Schema von heidICON folgt dabei dem LIDO-Standard, was den Export und die Einspeisung in nationale und internationale Nachweisportale erleichtert.

Die Digitalisierung dieser Sammlungen ist nicht nur eine technische und logistische Herausforderung. Um sicherzustellen, dass die entstehende digitale Sammlung kein isoliertes Silo darstellt, betrachten wir das Vorhaben ebenso als soziale Herausforderung, die die Einbeziehung möglichst vieler relevanter Akteure schon in der Konzeptionsphase erfordert. Daher haben wir parallel zur Erstellung des Digitalisierungskonzeptes eine Working Group im Rahmen des Pelagios-Netzwerkes initiiert, die relevante Stakeholder aus der internationalen Forschungsgemeinschaft umfasst und gemeinsam Leitlinien erarbeitet hat, die eine zukünftige Vernetzung der heterogenen Bestände bestehender und geplanter Projekte erleichtern soll (Elwert/Pons 2020).

Das im Februar 2021 gestartete DiGA-Projekt folgt diesen Leitlinien und will mit der Digitalisierung und Erschließung der Sammlungen zugleich in mehrfacher Hinsicht Brücken bauen:

Brücken zwischen traditioneller Forschung und Digital Humanities

Inhaltlich ist das Projekt an der Schnittstelle von südasiatischer Kunstgeschichte, Buddhismuskunde und dem spezialisierten Feld der Gandhara-Studien angesiedelt. Im Vergleich zu den relevanten Feldern der Digital Humanities, insbesondere der digitalen Kunstgeschichte, fällt eine doppelte Leerstelle auf: Die fachwissenschaftliche Forschung hat zwar für ihre Zwecke extensive Systematiken zur Beschreibung Gandharischer Kunst entwickelt (etwa Faccenna/Filigenzi 2007), diese sind aber nicht als nachnutzbare digitale Ressourcen verfügbar. Auf der anderen Seite weisen digitale Ressourcen wie der Getty Arts and Architecture Thesaurus (AAT) oder IconClass eklatante Lücken im Bereich der außereuropäischen Kunst und Ikonografie auf. Das DiGA-Projekt setzt hier an, indem es die gewachsenen Fachstandards als Linked-Data-Ressourcen verfügbar macht und zugleich ihre Vernetzung mit den etablierten digitalen Ressourcen vorantreibt. Als erstes Ergebnis dieser Bemühungen wird ein digitaler Thesaurus zur Beschreibung buddhistischer Kunst im SKOS-Format sowie ein Gazetteer archäologischer Grabungsstätten der Gandhara-Region vorgestellt.

Brücken zwischen digitalen Sammlungen

Im Sinne der Linked-Open-Data-Vision vernetzter Datenbestände will das Projekt nicht bei der Erstellung digitaler Thesauri für den eigenen Gebrauch stehen bleiben. Die eigene Sammlung dient vielmehr als Experimentierfeld für die Etablierung von best-practice-Ansätzen, die einerseits bestehende Konzepte aufgreifen² und andererseits als Leitbild für zukünftige Vorhaben dienen kann. Wir gehen dabei nicht davon aus, dass die Etablierung von Standards einseitig erfolgen kann. Vielmehr sehen wir dies als sozialen Prozess an. Die im Rahmen des Pelagios Networks gegründete Arbeitsgruppe „Linked Data Methodologies in Gandharan Buddhist Art and Texts“ dient dabei als Plattform für den Austausch zwischen verschiedenen Projekten mit dem Ziel, sich auf gemeinsame Beschreibungsstandards zu einigen und ihre Implementierung zu unterstützen.

Brücken zwischen Ländern und Kontinenten

Die Gandhara-Forschung ist ein internationales Feld, mit wichtigen Zentren in Italien, Frankreich, Großbritannien und Pakistan. Die Umsetzung des Digitalisierungsvorhabens kann sich daher nicht allein an deutschen Beschreibungsstandards (etwa GND) beschränken. Zugleich laufen Digitalisierungsprojekte unter europäischer Leitung Gefahr, Teil einer neo-kolonialen Wissensextraktion aus Ländern des globalen Südens zu werden, die lokale Ressourcen nur als Rohstoff für die (akademische wie finanzielle) Wertschöpfung in den Ländern des globalen Nordens begreifen (Rojas Castro 2020). Das DiGA-Projekt wird in enger Partnerschaft mit dem Direktorat für Archäologie und Museen der Provinz Khyber Pakhtunkhwa durchgeführt und berücksichtigt dabei auch die Rechte und Interessen der lokalen Akteure, etwa in Bezug auf die Speicherung der Digitalisate in lokalen Repositorien, die Wissensvermittlung und den Aufbau eigener Infrastrukturen.

Das Poster präsentiert das DiGA-Projekt, die Pelagios Working Group und die Ergebnisse der ersten Digitalisierungsphase: Die erstellten Vokabulare und Gazetteers sowie den Stand der Digitalisierungs- und Erschließungsarbeiten.

Fußnoten

1. Das Vorhaben wird mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01UG2048X gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin/beim Autor.
2. Der Abschlussbericht der Pelagios Working Group (Elwert/Pons 2020) listet eine Reihe von Vorhaben, an denen sich das DiGA-Projekt orientiert, etwa BUDA für die Buddhismuskunde (Roux 2018) oder perspektivisch ONAMA für die Modellierung von Narrativen (Zeppezauer-Wachauer et al. 2021).

Bibliographie

Elwert, Frederik/Pons, Jessie (2020): *Linked Data Methodologies in Gandhāran Buddhist Art and Texts: Pelagios Working Group Final Report*. Bochum: Ruhr-Universität Bochum.

Faccenna, Domenico/Filigenzi, Anna (2007): *Repertorio terminologico per la schedatura delle sculture dell'arte gandharica – Sulla base dei materiali provenienti dagli scavi della Missione Archeologica Italiana dell'IsIAO nello Swat, Pakistan*. Rome: IsIAO.

Rojas Castro, Antonio (2020): „#FAIR enough? Building DH Resources in an Unequal World.“ In: *Proyecto Humboldt Digital (ProHD)*. <https://habanaberlin.hypotheses.org/1730>.

Roux, Elie (2018): „The BUDA platform. LOD platform and model for Buddhist Studies: Bibliography, Prosopography, Geography.“ Präsentation auf der Linked Pasts IV, Mainz. <https://doi.org/10.17613/kqrz-hp97>.

Zeppezauer-Wachauer, Katharina et al. (2021): „Needful Things. Die Relationen der Dinge in einer Ontologie mittelalterlicher Narrative.“ In: *Medieval and Early Modern Material Culture Online* 8.

Building and Improving an OCR Classifier for Republican Chinese Newspaper Text

Arnold, Matthias

arnold@uni-heidelberg.de

Heidelberg Centre for Transcultural Studies, Universität Heidelberg, Germany

Henke, Konstantin

konstantin.henke@protonmail.ch

Institut für Computerlinguistik, Universität Heidelberg, Germany

For more than a decade, Republican magazines and newspapers have been collected by institutes and projects now joined in the Centre for Asian and Transcultural Studies (CATS) at Heidelberg University. Our platform “Early Chinese Periodicals Online” (ECPO, <https://uni-heidelberg.de/ecpo>), provides open access to more than 300.000 digital images and their metadata, cf. Arnold and Hessel (2020). Since the material consists mostly of image scans, the project ran a number of experiments to explore possible approaches towards full text generation (Arnold, 2021). For newspapers printed in Latin scripts much has changed since Rose Holley commented item “Use the ‘training’ facility (artificial intelligence) in the OCR software” with “Not viable for cost effective mass scale digitization” and noted “Do not pursue” in her list of “Potential methods of improving OCR accuracy suggested by ANDP team” (Holley, 2009, table 2, item 9). Today, when researchers write that “transforming [historical newspapers] into machine-readable data by means of OCR poses some major challenges” they do that while they introduce their own OCR pipeline (Holley, 2009).

Unfortunately, these approaches cannot just be adopted to historical Chinese newspapers. As we have shown, especially complex layout and resulting difficulties of reliable automatic page segmentation have so far prevented full text generation of these newspapers even within China (Arnold, 2021; Arnold, forthcoming; Arnold et al., forthcoming). In this long abstract we present the first results from a systematic approach towards full text extraction from a Republican China newspaper (1). Our basis is a small corpus for which text ground truth exists. We present our character segmentation method which produces about 90.000 images of characters. Based on the hypothesis that pre-training on extensive amounts of suitably augmented character images will increase the OCR accuracy for evaluation on real-life character image data, we generate synthetic training data. We then compare the OCR recognition results and show that a combination of synthetic and real characters produces best results. Finally, we propose a method that makes use of a masked language model for OCR error correction.

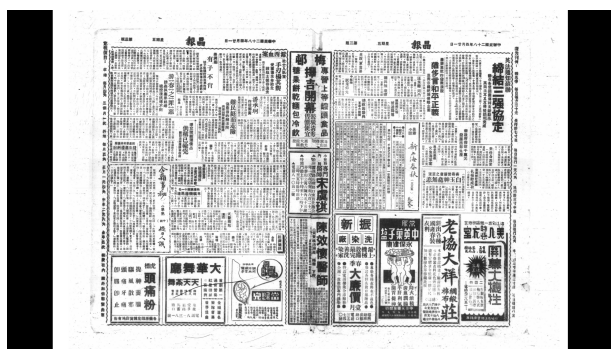


Fig. 1: An example fold from *Jing bao* 晶報 (*The Crystal*), April 21, 1939, pages 2-3.

Note: We will treat single rectangular text blocks (Fig. 2) as given and proceed from here to present effective methods for generating a data set later used to train an OCR model. We show that pre-training on artificially created training data can significantly improve OCR accuracy. Due to the limited scope of the presented experiments, this approach is still limited in terms of retrieved glyph size, image quality and font style, hence the model is not necessarily directly applicable to other historical Chinese documents.



Fig. 2: Manually cropped text blocks

The Corpus

Our corpus consists of 9.385 scanned folds from the entertainment newspaper *Jing bao* 晶報 (*The Crystal*), published 03.03.1919–23.05.1940 (Fig. 1). The double-keyed text ground truth comprises all April 1939 issues (40 folds, ~245.000 characters). Aside from text blocks and their headings, it also contains mastheads, advertisements and marginalia, however, the methods presented below will solely focus on “header-less” text blocks of uniform font-size.

Character Segmentation

Due to Chinese characters’ nearly squared appearance, it is common to find resulting text blocks implicitly displaying a grid layout (see Fig. 2). Deviation from the grid usually appears when additional characters had to be squeezed into one column or because of inaccurate printing. In order for the method described below to work, we manually sort out any text blocks that don’t adhere to the grid layout and then extract the corresponding ground truth section for every crop.

After adaptive binarization (kernel size: 125 px) we calculate horizontal and vertical projection profiles, cf. Fan et al. (1998). To perform deskewing, we find an angle α with $\alpha \in [-2.0, -1.5, \dots, 2.0]$ such that rotating the image by α maximizes

$$\sum_i^{w-1} (c_{i+1} - c_i)^2 + \sum_j^{h-1} (l_{j+1} - l_j)^2$$

where w and h are the width and height of the image, c_i is the number of black pixels in the i -th column (= the corresponding value of the vertical projection profile) and l_j in the j -th line.

After deskewing, we cut the gray-scale, non-binarized original text block image into single character images along separators defined by the following heuristic:

- (1) Use the valleys of the vertical projection profile to define separators between the columns.
- (2) Use the valleys of the horizontal (global) projection profile to define separators between the lines.
- (3) For every column, produce another (local) projection profile. If a local separator lies within 7px distance of a global separator defined by (2), discard the global separator and only use the local separator; else only use the global separator.

The positions of the valleys are obtained by `scipy.signal.find_peaks` using a minimum distance of (1) 22, (2) 20 and (3) 14.

For normalization and contrast enhancing the following method is used:

1. Globally (whole crop): Employ partial adaptive thresholding: Every pixel whose gray-scale value is larger (= brighter) than the average of a surrounding 7x7-kernel is set to 255 (white). Separately, every pixel whose value is greater than the median of the image (called threshold below) is assumed to be a background pixel and set to 255. Every other pixel keeps its gray-scale value. Choosing the median arises from the supposition that there are more background than content pixels.
2. Locally (after cropping rectangles containing one character each): Ignoring white pixels, linearly re-scale pixel values from $[\text{minval}, \text{threshold}]$ to $[0, 255]$, where minval refers to the

darkest pixel in the image. This allows even for very lightly printed characters to appear darker and have their decisive features more strongly separated from the background.

Finally, the resulting fields can be easily mapped to the ground truth text. Indentations have to be manually marked, and since the CNN (cf. Section 3.) requires squared images as input, we add white padding to transform the rectangular character images into square ones.

This method entirely relies on correct annotation. While we can easily detect errors like missing lines, this is harder for missing or extra characters within a line (checking the line length), and basically impossible for typos or swapped characters. To avoid such mistakes we can only double-check annotations, otherwise they lower recognition accuracy.

Character Image Generation

The method described in the section above yields a total of 92.039 character images (47.986 train + 21.676 dev + 22.377 test). Due to the Zipfian distribution, we additionally present the following table:

Tab. 1

x	number of characters with at least x samples
1	3045
2	2355
3	1995
...	...
10	1091
...	...
20	696
...	...
50	301
...	...
100	137

Motivated by the low quantity of training samples for higher x, we generate additional synthetic training data and propose the following research hypothesis:

Pre-training on extensive amounts of suitably augmented character images will increase the OCR accuracy for evaluation on real-life character image data.

With the goal of imitating the real-life character images with artificial training data, we apply the following, partly randomized (in b., e2.2, f., g., and h.) augmentations to glyph images extracted from various fonts:

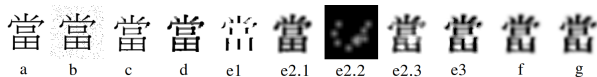


Fig. 4: Augmentations to glyph images

1. Extract PNG images of a predefined set of glyphs from a Song-Ti font (= the font-style used in the newspapers).
2. Add random noise (peppering).
3. Use morphological opening and then closing to enlarge noise pixels, grow them together with other close-by black pixels (other noise or the actual character) during erosion (= dilation of black contours on white background) and remove useless noise during dilation (= erosion of black pixels).

4. Use erosion to thicken lines.
5. Emphasize vertical lines while blurring and staining the remaining parts:
 1. Extract vertical elements of a certain minimum length using dilation with a vertical kernel.
 2. Separately apply the following:
 1. Further erode and blur the image.
 2. Generate random patches.
 3. Add the patches to the image.
 3. Join the result and the previously extracted vertical lines using bitwise AND.
6. Blur the image once more. Additionally, brightness can be randomly in-/decreased before. Afterwards, linearly rescale pixel values to cover the whole 0-255 range, like the real-life images.
7. Apply randomized elastic transformation.
8. Add padding and perform appropriate resizing.

Since ultimately, the classes used for OCR are Unicode points, the question arises which code points to synthesize additional training data from. We employ the simple heuristic of using all of the glyphs featured in the ground truth, and adding any missing ones from the 4000 most frequent characters of a representative corpus. Furthermore, inconsistencies caused by Han-unification have to be solved. For example, the image data features 青 instead of 青 and 清 instead of 清 (all different code points), however only one code point exists for every other character containing 青/ 青 as a component (請, 情, 靜, ...). While 值 and 值 (the latter being the variant used in our image data) have different code points, their right component itself (直) is Han-unified, etc. We decide to always use the most accurate code point as long as it's not part of the CJK Compatibility Ideographs block (U+F900...U+FAFF), so e.g. 令 (U+4EE4) is used instead of 令 (U+F9A8), even though the latter might appear more accurate, depending on the font. Generally, we find that the character variants printed in our image data to be visually closer to the Japanese standard (e.g. the components 𠂇 and 𠂇), so we choose several Japanese fonts for training data generation.

Character Recognition

We decide on using a GoogleNet CNN architecture (Szegedy et al., 2015), slightly modified to take 1-channel inputs instead of RGB-images. This has proven to be effective regarding both printed and handwritten Chinese character recognition, e.g. Zhong et al. (2015) and Xu et al. (2018). Training on different character image sets, we obtain the following top-k accuracies on the real-life validation set for $k \in \{1, \dots, 10\}$:

Tab. 2

k	1	2	3	4	5	6	7	8	9	10
only synthetic character images (4 different fonts)	69.73	78.3	81.68	83.65	84.99	86.06	86.87	87.49	87.97	88.46
only real character images	96.47	97.29	97.46	97.56	97.61	97.66	97.68	97.69	97.69	97.71
pretraining on synthetic; fine-tuning on real	97.63	98.57	98.78	98.91	98.98	99.01	99.07	99.1	99.12	99.13

We also find that the selection of the fonts by which variants (i.e. mainland Chinese, Taiwanese, Japanese, Korean) it was designed for is largely negligible, i.e. a Taiwanese font may score higher than a Japanese font, even though the latter features glyph variants closer to those found in our data. This is probably because the percentage of characters with regional variants is relatively small, and also implies that the characters' stroke length and distance as well as small variations in the size of single character components is more relevant to the OCR accuracy when evaluating on real-life character images.

Interestingly, while there is a huge difference in performance after training on synthetic vs. real data, the human eye is barely able to differentiate between even a big selection of synthetic and real character images if presented next to each other (cf. Fig. 5).

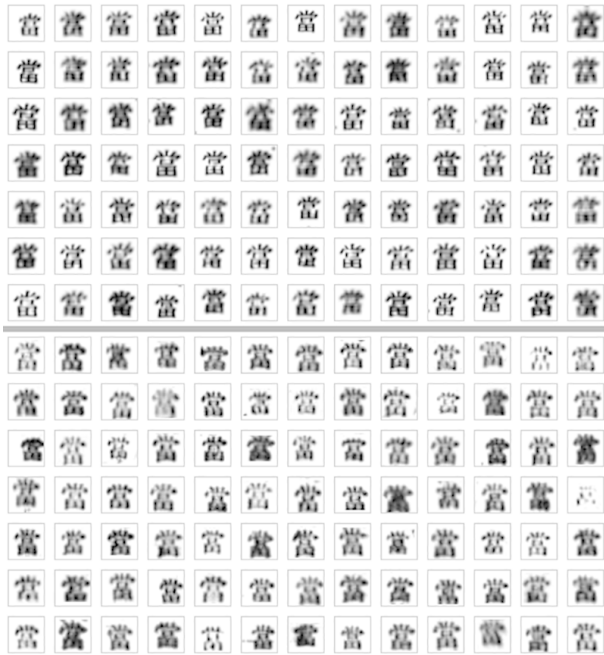


Fig. 5: Comparison between synthetic (top) and real (bottom) character images

OCR Error Correction

Finally, we aim to improve top-1 accuracy values by using language models to find the correct character among the second to k -th prediction. As can be seen in the table in Section 3, there is a significant jump from top-1 to top-2 accuracy, meaning that for wrong predictions the gold character is often predicted in the second position.

Inspired by Wang et al. (2019), we propose a method that identifies characters likely to be incorrect: Let x_1 and x_2 denote the logit scores of the top 2 candidates output by the OCR model. Now we set a threshold t for the difference between x_1 and x_2 . Any OCR prediction where $x_1 - x_2 < t$ is treated as likely to be incorrect and is passed on to the correction step. This step works by having a pre-trained BERT model re-predict the character from the top k OCR candidates. Systematically testing for different combinations of t and k (with $t \in [0, 0.5, \dots, 10]$ and $k \in [0, 1, \dots, 18]$), we settle with $t = 2.5$ and $k = 7$, where we attain the following final results:

Tab. 3

	Development set	Test set
Only OCR w/o pre-training	96.54	95.49
Only OCR w/ pre-training	97.63	96.95
OCR w/ pre-training + BERT-based correction	98.05	97.44

As becomes evident, the presented post-processing method reduces the error by 18.1% (dev. set) / 16.1 % (test set).

Bibliography

Arnold, Matthias (2021): *Ground Truth, Neural Networks, OCR: Towards Full Text of Republican China Newspapers*. <https://tinyurl.com/ecpo-intro> [letzter Zugriff 15. Juli 2021].

Arnold, Matthias (forthcoming): "Multilingual research projects: Challenges for making use of standards, authority files, and character recognition", in: *Digital Studies / Le champ numérique*.

Arnold, Matthias / Hessel, Lena (2020): "Transforming data silos into knowledge: Early Chinese Periodicals Online (ECPO)", in: Heuveline, Vincent / Gebhart, Fabian / Mohammadianbisheh, Nina (Hrsg.): *E-Science-Tage 2019: Data to Knowledge*. Heidelberg: heiBOOKS. S. 95–109. 10.11588/heibooks.598.c8420.

Arnold, Matthias / Paterson, Duncan / Xie, Jia (forthcoming): "Procedural Challenges: Machine Learning tasks for OCR of historical CJK newspapers", in: *International Journal of Digital Humanities*.

Fan, Kuo-Chin / Wang, Liang-Shen / Tu, Yin-Tien (1998): "Classification of Machine-Printed and Handwritten Texts Using Character Block Layout Variance", in: *Pattern Recognition* 31, S. 1275–1284.

Holley, Rose (2009): "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs", in: *D-Lib Magazine* 10.1045/march2009-holley.

Liebl, Bernhard / Burghardt, Manuel (2020): "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline", in: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*. Amsterdam, the Netherlands. S. 351–373. (= CEUR Workshop Proceedings). <http://ceur-ws.org/Vol-2723/long20.pdf> [letzter Zugriff 15. Juli 2021].

Sung, Doris / Sun, Liying / Arnold, Matthias (2014): "The Birth of a Database of Historical Periodicals: Chinese Women's Magazines in the Late Qing and Early Republican Period", in: *Tulsa Studies in Women's Literature* 33, S. 227–237.

Szegedy, Christian et al. (2015): "Going deeper with convolutions", in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA. S. 1–9. 10.1109/CVPR.2015.7298594.

Wang, Hsiang-An / Liu, Pin-Ting (2019): "Towards a Higher Accuracy of Optical Character Recognition of Chinese Rare Books in Making Use of Text Model", in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. New York, NY, USA: Association for Computing Machinery. S. 15–18. (= DATeCH2019). 10.1145/3322905.3322922.

Xu, Xin et al. (2018): "Chinese Characters Recognition from Screen-Rendered Images Using Inception Deep Learning Architecture", in: Zeng, Bing et al. (Hg.): *Advances in Multimedia Information Processing – PCM 2017*. Cham: Springer International Publishing. S. 722–732. (= Lecture Notes in Computer Science).

Zhong, Zhuoyao / Jin, Lianwen / Xie, Zecheng (2015): "High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature maps", in:

2015 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis, Tunisia. S. 846–850. 10.1109/ICDAR.2015.7333881.

Computational Literary Studies Data Landscape Review

Börner, Ingo

ingo.boerner@uni-potsdam.de
Universität Potsdam

Charvat, Vera Maria

VeraMaria.Charvat@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)

Đurčo, Matej

Matej.Durco@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)

Mrugalski, Michał

michal.mrugalski@hu-berlin.de
Humboldt-Universität zu Berlin

Odebrecht, Carolin

carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin

Literarische Werke und deren digitale Repräsentationen stellen auch in den Fachbereichen der Computational Literary Studies (CLS) das Fundament für epistemische Auseinandersetzungen und Diskurse. Die unterschiedlichen Prozessierungen und Visualisierungen wie digitale Editionen (Sahle 2013) oder Netzwerkanalysen (Trilcke 2013) von literarischen Werken aller Gattungen (Epik, Drama, Lyrik) erzeugen eine Vielzahl an heterogenen Daten, die immer flexibler und umfassender miteinander in Interaktion treten und kommunizieren. Diese Entwicklung stellt die Frage der Interoperabilität der Daten in den Mittelpunkt, wobei Linked Open Data (LOD; Heath & Bizer 2011) eine zentrale Rolle spielen.

Das übergeordnete Ziel von "Computational Literary Studies Infrastructure"¹ – ein von der EU finanziertes "Integrating Activities for Starting Communities (IASC)"-Projekt – ist die Schaffung eines einheitlichen und einfachen Zugangs zu den besten europäischen und nationalen Infrastrukturen für die CLS-Community. In unserem Arbeitspaket *Data Selection and Curation* möchten wir Informationen über literarische Werke systematisch für die CLS-Community kompilieren, aufbereiten und konsolidieren, um die Zugangsparadigmen für literarische Daten signifikant zu rekonfigurieren und die Einhaltung der FAIR-Prinzipien (findable, accessible, interoperable, reusable; Wilkinson et al. 2016) erheblich zu verbessern.

Um die Auffindbarkeit und den forschungsorientierten Zugang zu literarischen Daten für die CLS-Community zu ermöglichen, ist eine Inventarisierung der CLS-Datenlandschaft erforderlich, die forschungsrelevante Kriterien für die Datenauswahl sowie deren Erfassung und Beschreibung anwendet. Mit dieser Inventarisierung, die wir in Form einer *Data Landscape Review* durchführen, kann die vorhandene Datenlandschaft als digitales Erbe für CLS-Kontexte erst umfassend sichtbar und als Vorlage für weitere Forschungsvorhaben zugänglich gemacht werden.

Dabei stellen wir uns unter anderem folgende Fragen: Welche Beschreibungsmerkmale sind für die Daten als Kollektion im Sinne einer eigenen epistemischen Einheit wesentlich? Welche Beschreibungsmerkmale sind in Bezug auf die literarischen Vorlagen und deren Aufbereitungen wichtig? Wie kann nach Kollektionen oder einzelnen Datensätzen im Sinne der *programmable corpora* (Fischer et al. 2019) recherchiert werden?

Die Ergebnisse unserer *Data Landscape Review* werden wir als Posterpräsentation mit Fokus auf den technischen Bericht zur Kartierung und Kontextualisierung der CLS-Daten vorstellen.

Aufbauend auf der Review werden die Ergebnisse in Form eines stetig wachsenden, interaktiven Online-Katalogs literarischer Corpora für die CLS-Community bereitgestellt. Dieser wird eine umfassende Übersicht über die verfügbaren Ressourcen inklusive ausführlicher beschreibender Metadaten liefern und die üblichen Abfrage- und Erschließungsmöglichkeiten mittels verschiedener Such- und Filtermechanismen bieten. Konzeptueller Ausgangspunkt für die strukturierte Sammlung der Informationen ist das Metamodell für Korpusmetadaten (MKM; Odebrecht 2018) – ein, generisches erweiterbares Beschreibungsmodell, für die zentralen Entitäten *Korpus*, *Dokument* und *Annotation* sowie ihre Beziehungen untereinander.

Während das Modell selbst abstrakt definiert ist, erarbeiten wir eine kongruente/entsprechende Ontologie im OWL-Format (OWL, 2012), welche eine Repräsentation der Daten in RDF (Resource Description Framework)² ermöglicht. Die Formalisierung als OWL-Ontologie gestattet darüber hinaus auch, Äquivalenzen zu bereits bestehenden Ontologien und Schemata im Sinne des LOD-Paradigmas explizit zu machen. Hier sind insbesondere Ansätze zur Text- und Publikationseinordnung wie FRBR (IFLA, 1998) und Dublin Core (ISO standard 15836) zu nennen. Neben Äquivalenzen auf der Schema-Ebene wird der Datensatz um Verweise/Verlinkungen zu externen Referenzressourcen wie zum Beispiel die Normdateien GND (Gemeinsame Normdatei)³, VIAF (Virtual International Authority File)⁴, WikiData⁵ und GeoNames⁶ angereichert. Diese sind unabdingbar, um semantische Interoperabilität zwischen Datensätzen herzustellen. Die in RDF serialisierten Daten werden selbstverständlich regelmäßig als geschlossener Datensatz ("Dump"), sowie über einen SPARQL⁷-Endpoint verfügbar gemacht. Die Ontologie sowie eine erste proof-of-concept Version des Online-Katalogs werden wir bei der Tagung präsentieren.

Ebenso ist zu berücksichtigen, dass dieser Katalog Teil von einem komplexen Gefüge an Ressourcen, Providern und Disseminationskanälen bzw. Aggregatoren ist. Die Position des Katalogs in diesem Gefüge und seine Beziehung zu verwandten Aggregatoren wie CLARIN VLO (Virtual Language Observatory)⁸, Europeana⁹ oder OpenAIRE (Open Access Infrastructure for Research in Europe)¹⁰ müssen noch im Detail erarbeitet werden. Der grundlegende Ansatz wird dabei aber sein, die Information über mehrere Kanäle möglichst breit zu streuen/dissemिनieren und dafür auch Mappings der Metadaten in die erforderlichen Metadaten-Formate, wie CMDI (Component Metadata Initiative)¹¹ für

VLO bzw. EDM (Europeana Data Model)¹² für Europeana bereitstellen.

Die *Data Landscape Review* und der Online-Katalog werden den Forschenden Zugriff zu einer breiten Palette an Ressourcen, die über mehrere europäische Anbieter distribuiert sind, ermöglichen und mit Beschreibungen und Informationen auch einen umfassenden, domänenspezifischen Überblick über diese Ressourcen bieten.

Fußnoten

1. Website des Projekts CLS INFRA (No. 101004984): <https://clsinfra.io/> (letzter Zugriff 24.11.2021)
2. RDF W3C Recommendation: <https://www.w3.org/TR/rdf-primer/> (letzter Zugriff 24.11.2021)
3. GND: <https://gnd.network> (letzter Zugriff 24.11.2021)
4. VIAF: <http://viaf.org/> (letzter Zugriff 24.11.2021)
5. Wikidata: <https://www.wikidata.org/> (letzter Zugriff 24.11.2021)
6. GeoNames: <http://www.geonames.org/> (letzter Zugriff 24.11.2021)
7. SPARQL steht für SPARQL Protocol and RDF Query Language; SPARQL W3C Recommendation: <https://www.w3.org/TR/sparql11-overview/> (letzter Zugriff 24.11.2021)
8. CLARIN VLO: <https://vlo.clarin.eu/> (letzter Zugriff 24.11.2021)
9. Europeana: <https://www.europeana.eu/de> (letzter Zugriff 24.11.2021)
10. OpenAIRE: <https://www.openaire.eu/> (letzter Zugriff 24.11.2021)
11. CMDI: <https://www.clarin.eu/cmdl> (letzter Zugriff 24.11.2021)
12. gesammelte Dokumentationen zum EDM: <https://pro.europeana.eu/page/edm-documentation> (letzter Zugriff 24.11.2021)

Bibliographie

- Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hecht, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer** (2019): "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama", in: Fischer, Frank / Akimova, Marina / Orekhov, Boris (eds.): *Digital Humanities 2019. Conference Abstracts*, Utrecht University, Moscow, <https://dev.clarion.nl/files/dh2019/boa/0268.html>
- Heath, Tom / Bizer, Christian** (2011): "Linked Data: Evolving the Web into a Global Data Space", 1st edition, in: *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 1, No. 1 [San Rafael, Calif.]: Morgan & Claypool, S. 1-136, doi: <https://doi.org/10.2200/S00334ED1V01Y201102WBE001>
- IFLA** (1998): "Functional Requirements for Bibliographic Records: Final Report", in: *IFLA Series on Bibliographic Control* 19 (former UBCIM). München: K.G. Saur Verlag.
- ISO standard 15836** (2017): "The Dublin Core Metadata Element Set"
- Odebrecht, Carolin** (2018): "MKM – ein Metamodell für Korpusmetadaten. Dokumentation und Wiederverwendung historischer Korpora", Dissertation. Humboldt-Universität zu Berlin, Sprach- und literaturwissenschaftliche Fakultät, Berlin. doi: <https://doi.org/10.18452/19407>
- Sahle, Patrick** (2013): "Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels", 3 Bände, Norderstedt: Books on Demand, in: *Schriften des Instituts für Dokumentologie und Editorik*, Bände 7-9.

dels", 3 Bände, Norderstedt: Books on Demand, in: *Schriften des Instituts für Dokumentologie und Editorik*, Bände 7-9.

Trilcke, Peer (2013): "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft", in: Ajouri, Philip / Mellmann, Katja / Rauen, Christoph (eds): *Empirie in der Literaturwissenschaft*. Paderborn: Mentis 201–247.

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand J. / et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship". *Sci Data* 3, 160018. doi: <https://doi.org/10.1038/sdata.2016.18>

W3C OWL Working Group (2012): Web Ontology Language (OWL), <https://www.w3.org/OWL/>

Corpus Nummorum Eine digitale Forschungsinfrastruktur für antike Münzen

Köster, Jan

jan.koester@bbaw.de
Berlin-Brandenburg. Akademie der Wissenschaften, Germany

Franke, Claus

franke@bbaw.de
Berlin-Brandenburg. Akademie der Wissenschaften, Germany

Peter, Ulrike

peter@bbaw.de
Berlin-Brandenburg. Akademie der Wissenschaften, Germany

Geprägte Münzen sind Objekte von besonderem quellenkundlichen Wert. Als erstes allgemein akzeptiertes normiertes Tauschmittel waren sie einer der wichtigsten Schlüssel zur Entstehung übergreifender Handelsnetzwerke und damit letztendlich zu der Wirtschaft und Gesellschaft, wie wir sie heute kennen. Darüber hinaus wurden (und werden) Münzen sehr oft mit bildlichen Darstellungen und kurzen Textbotschaften versehen, sodass sie nicht nur einen finanziellen Wert darstellen, sondern auch ein Mittel der Kommunikation repräsentieren. Die geringe Größe erfordert jedoch eine spezielle, komprimierte, mithin chiffrierte Bildsprache, die zu entschlüsseln besondere wissenschaftliche Methoden und vor allem eine breite, systematisch erfasste Datengrundlage erfordert. Aufgrund ihrer großen Stückzahl, den sozioökonomischen Implikationen, der Symbolkraft und der in der Regel hervorragenden Datierbarkeit bilden Münzen eine der wichtigsten Materialgruppen in den klassischen Altertumswissenschaften.

Das an der Berlin-Brandenburgischen Akademie der Wissenschaften (Zentrum Grundlagenforschung Alte Welt) beheimatete Drittmittelprojekt Corpus Nummorum (<https://www.corpus-nummorum.eu/>) erschließt antike griechische Münzen geordnet nach Regionen und konzentriert sich auf Prägungen aus Thrakien, Moe-sien, Mysien und der Troas, die bereits in der Webpräsenz des CN verfügbar sind. Die Bestände werden in Kooperation mit dem Münzkabinett der Staatlichen Museen zu Berlin und internationalen Partnern sowie der direkten Einbeziehung der numismatischen Community (sowohl Forscher*innen als auch Laien) erschlossen. Die Sammlung der Objekte bildet die Grundlage für die Bestim-

mung und Erfassung der zur Herstellung der Münzen verwendeten Stempel sowie der Klassifizierung der Münzen in ausführlich beschriebenen Leittypen.

Das Corpus Nummorum bildet eine digitale Forschungsinfrastruktur, welche die Nachnutzung der Forschungsdaten und die uneingeschränkte Kollaboration mit anderen Personen bzw. Institutionen gestattet und fördert. Dabei kommt der Verwendung und Erweiterung numismatischer Normdaten eine große Bedeutung zu. Besonders hervorzuheben ist hier die enge Verzahnung mit dem Normdatenportal Nomisma (<http://nomisma.org/>). Des Weiteren werden zusammen mit dem Big Data Lab der Goethe-Universität zu Frankfurt a.M. die Potenziale des Natural Language Processing, der machine-learning-basierten Bilderkennung sowie der automatisierten Qualitätskontrolle eruiert und zur praktischen Anwendung gebracht. So soll unter anderem ein multilingualer ikonographischer Thesaurus entstehen, von dem alle ikonographisch arbeitenden Altertumswissenschaften profitieren.

Im Bereich des Research Software Engineering hat das Corpus Nummorum in Zusammenarbeit mit dem Arbeitsbereich TELOTA – IT/DH der BBAW den CN Editor, eine multifunktionale Web-App, entwickelt, welche den gesamten Workflow von der Anlage eines neuen Datensatzes, über den Upload und die Verknüpfung mit Bildern bzw. anderen Medien sowie die Anreicherung mit Normdaten bis hin zur Veröffentlichung händeln kann. Hinzu treten umfangreiche Such- und Filtermöglichkeiten einschließlich verschiedener Indices und einer Volltextsuche, welche Boolesche Operatoren, REGEX sowie CN-spezifische Ausdrücke erlaubt. Der CN Editor ist Open-Source-Software. Der Sourcecode ist seit 2021 vollumfänglich auf Github (<https://github.com/telota/corpus-nummorum-editor>) unter der GPL-3.0 License frei verfügbar ist. Er basiert auf dem PHP-Framework Laravel für das Backend sowie dem Javascript-Vue.js-Framework für das Frontend (Single-Page-Application). Hinzu tritt eine durchstrukturierte MySQL-Datenbank. Ein leichtgewichtiger, modularer Aufbau erlaubt eine schnelle Erweiterung des CN Editors um neue Funktionen oder die Anpassung an andere Objektgattungen, was ihn auch für Vorhaben jenseits der Numismatik interessant macht.

Das geplante Poster bzw. die Präsentation soll neben der Projektkonzeption vor allem die praktische digitale Arbeit des Corpus Nummorum veranschaulichen und gleichermaßen Herausforderungen wie Lösungsansätze vorstellen, die nicht nur für das Feld der Numismatik allein, sondern für alle Fächer, die mit ähnlich strukturierten Daten arbeiten, von Relevanz sind. Zu nennen wären hier etwa:

- Verknüpfung von Einzelobjekten mit entsprechenden Leitgruppen, wobei alle oder nur ausgewählte Werte der Leitgruppe dynamisch auf das Einzelobjekt übertragen werden können (und zwar sowohl manuell als auch automatisiert, etwa wenn die Leitgruppe aktualisiert wird)
- Versionierung und Edition. Aktuell beschäftigen wir uns intensiv mit einer eher technischen Versionierung, die jede direkt Änderung erfasst, und einer gezielten Neuedition, die z.B. nötig werden kann, wenn die Datierung einer Münze aufgrund neuer Erkenntnisse geändert werden muss. Diese Vorgänge transparent, nachvollziehbar und gleichzeitig technisch niedrigschwellig zu gestalten, ist eines unserer wichtigsten Anliegen.
- Erstellen eines ikonographischen Thesaurus basierend auf Natural Language Processing der eingegebenen Daten sowie einer entsprechenden Aufbereitung und Einordnung der Ergebnisse. Ziel ist ein innovatives standardisiertes multilinguales Normdatenportal für Münzikonographie.

Bibliographie

Ulrike Peter / Karsten Tolle (2019): "Corpus Nummorum – Coins, types and data quality control", *Proceedings of the 8th Joint Meeting of ECFN and nomisma.org 2019* (im Druck)

Patricia Klinger / Sebastian Gampe / Karsten Tolle / Ulrike Peter (2018): "Semantic Search based on Natural Language Processing – a Numismatic example", in: *Journal of Ancient History and Archaeology* 5, 3: 68-79

Ulrike Peter (2017): "Corpus Nummorum Thracorum – A Research Tool for Thracology and an Example of Digital Numismatic Collaboration", in: Maria Caccamo Caltabiano (eds.): *XV International Numismatic Congress Taormina 2015. Proceedings*, 1, Roma / Messina: 1306

Das DFG Schwerpunktprogramm Computational Literary Studies

Pielström, Steffen

pielstroem@biozentrum.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Jung, Kerstin

kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Germany

Die Computational Literary Studies (CLS) sind ein wachsendes, interdisziplinäres Forschungsfeld angesiedelt zwischen Literaturwissenschaft, Computerlinguistik und Informatik, in dem computergestützte Verfahren zur Analyse literaturwissenschaftlicher Fragestellungen zum Einsatz kommen. Insgesamt elf Einzelprojekte aus Deutschland und der Schweiz, die zur Zeit in diesem Emerging Field arbeiten, gehören dem seit 2020 aktiven Schwerpunktprogramm SPP 2207 "Computational Literary Studies" der Deutschen Forschungsgemeinschaft (DFG) an, davon erhalten 10 Projekte direkte Förderung aus dem Programm, ein weiteres Projekt ist mit dem Programm assoziiert. Hinzu kommt ein Zentralprojekt das, als Besonderheit neben der organisatorischen und inhaltlichen Koordination der Fortschrittsvorhaben, über eine eigens eingerichteten Personalstelle für das projektübergreifende Forschungsdatenmanagement verfügt. So bietet das Programm eine enge Begleitung und Abstimmung der Projekte in Fragen des Forschungsdatenmanagements über die gesamte Laufzeit. Für das kooperative Arbeiten wird vom Zentralprojekt u.a. eine Gitlab-Instanz zur Verfügung gestellt.

In den einzelnen Projekten kooperieren erfahrene Digital Humanists eng mit etablierten Literaturwissenschaftler*innen um an aktuell relevanten Fragen der Literaturwissenschaft zu arbeiten. Die Forschung im SPP 2207 konzentriert sich vor allem auf die deutschsprachige Literatur. Hier reicht das Spektrum der Forschungsgegenstände von Romanen über Dramen bis hin zur Poesie, die untersuchten Texte entstammen verschiedenen Epochen vom Mittelhochdeutschen bis ins späte 20. Jahrhundert. Hinzu kommen methodologische Untersuchungen die zum Ziel

haben, das methodische Repertoire der Computational Literary Studies für die spezifischen Anforderungen des Faches zu validieren und weiter zu entwickeln. So haben sich für Sentimentanalyse, Wordembeddings und Annotationen projektübergreifende Arbeitsgruppen etabliert und ein ganzes Projekt widmet sich der Methodenforschung im Bereich der kontrastiven Stilometrie.

Die Projekte in der ersten, dreijährigen Förderperiode sind im einzelnen:

- Anomaliebasierte quantitative Untersuchung von Stil und Gattung anhand des Stilmittelgebrauchs in mittelalterlicher Literatur (Joachim Denzler & Sophie Marshall)
- Die Anfänge der modernen Lyrik - Literaturgeschichte mit Textähnlichkeiten modellieren (Simone Winko & Fotis Jannidis)
- Advanced sentiment analysis for understanding affective-aesthetic responses to literary texts: A computational and experimental psychology approach to children's literature (Berenike Herrmann, Arthur Jacobs, Gerhard Lauer & Jana Lüdtko)
- Computergestützte Analyse von Unzuverlässigkeit und Wahrheit in Fiktion – Vernetzung und Operationalisieren der Narratologie CAUTION (Jonas Kuhn & Janina Jacke)
- Emotionen im Drama (Christian Wolff & Katrin Dennerlein)
- Evaluation von Events in der Narratologie EvENT (Evelyn Gius & Chris Biemann)
- Quantitative Drama Analytics: Tracking Character Knowledge Q:TRACK (Nils Reiter & Marcus Willand)
- Relating the Unread - Netzwerkmodelle in der Literaturgeschichte (Ulrik Brandes & Thomas Weitin)
- Literatur strukturieren - Varianten und Funktionen reflexiver Passagen in fiktionalen Erzähltexten (Anke Holler, Caroline Sporleder & Benjamin Gittel)
- Was ist wichtig? Schlüsselstellen in der Literatur (Robert Jäschke & Steffen Martus)
- Zeta und Konsorten - Distinktivitätsmaße für die Digitalen Literaturwissenschaften (Christof Schöch)

Angesichts der Vernetzung und Verankerung nahezu aller Programmbeteiligten in der nationalen wie internationalen Fachcommunity - so engagieren sich Mitglieder u.a. in der ADHO Special Interest Group "Digital Literary Stylistics", der EU COST Action "Distant Reading for European Literary History", dem EU-Programm "Computational Literary Studies Infrastructure" (CLSIInfra) und der ACL Special Interest Group on Humanities (SIGHUM) - sieht sich SPP 2207 nicht nur als Einrichtung für eine begrenzte Zahl geförderter Projekte sondern auch als Multiplikator und "Netzwerkknoten" für die gesamte CLS-Community, insbesondere im deutschsprachigen Raum. Veranstaltungen des Schwerpunktprogramms wie Meetings und Workshops sind daher in der Regel ebenso offen für Interessierte wie die projektübergreifenden Arbeitsgruppen, um die aktive Beteiligung weiterer Teile der Fachcommunity an den Aktivitäten von SPP 2207 zu fördern.

Mit dem vorliegenden Poster präsentiert sich SPP 2207 in seiner Gesamtheit und zeigt, wie die Vernetzung in einem solchen Programm Synergien und Gelegenheiten zur Zusammenarbeit schafft, die in der Zukunft auch in die weitere Forschungscommunity hinein wirken sollen.

Das optimale Datenmodell Eine Spurensuche im Möglichkeitsfeld der Kodierung

Saric, Sanja

sanja.saric@uni-graz.at

Institut Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich; Institut für Sprachwissenschaft, Universität Graz, Österreich

Steiner, Elisabeth

elisabeth.steiner@uni-graz.at

Institut Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich

Vogeltanz, Maximilian

maximilian.vogeltanz@uni-graz.at

Institut für Sprachwissenschaft, Universität Graz, Österreich

Einleitung

Zu Beginn jedes neuen Projektes in den Digitalen Geisteswissenschaften steht die Frage nach der adäquaten Modellierung der Forschungsdaten. Während im Förderansuchen häufig die Nennung von XML/TEI-Kodierung für Textquellen ausreichend ist, stellt sich die praktische Arbeit meist komplizierter dar: Schon die TEI (TEI Consortium 2021) bietet zahlreiche Möglichkeiten, ähnliche Sachverhalte zu annotieren und die gewählte Strategie muss dabei auf das Material, die Forschungsfrage sowie die Archivierung und Weiterverwendung der Daten Rücksicht nehmen.

Der vorliegende Beitrag stellt die Herausforderungen und Lösungsansätze anhand der Briefkorrespondenz von Hugo Schuchardt vor.

Herausforderung bestmögliche Kodierung

Ein Modellierungsansatz erfasst die untersuchten Merkmale möglichst genau und standardisiert in einem anerkannten Schema; berücksichtigt Referenzimplementationen und *best practice*-Guidelines; bezieht fachspezifische Vokabularien und Normdaten mit ein; versieht die Daten bereits im Entstehungsprozess mit Metadaten für die Weiterverwendung. Viele dieser Punkte sind ebenfalls Grundpfeiler der FAIR-Datenprinzipien (Wilkinson et al. 2016). In der Projektarbeit begrenzen jedoch oft verfügbare Zeit- und Personalressourcen die Umsetzbarkeit aller Aspekte, was notgedrungen zu Kompromissen führt. Zusätzlich konkurriert das Bedürfnis, projektspezifische Merkmale zu berücksichtigen, mit dem Anspruch an Vergleichbarkeit und Standardisierung. Trotzdem ist gerade die Interoperabilität zentral für die Nachhaltigkeit der Forschungsdaten.

Diese Herausforderungen stellten sich auch im *Hugo Schuchardt Archiv* (Hurch 2007-), einem langjährigen Vorhaben des Instituts für Sprachwissenschaft der Universität Graz. Im Mit-

telpunkt einer Kooperation mit dem Institut Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities (ZIM-ACDH) steht die Migration aller Ressourcen vom Institut für Sprachwissenschaft in das Repositorium GAMS, um die Korrespondenz und andere Dokumente aus dem Nachlass zu archivieren.¹

Daher tritt zu den genannten Faktoren ein weiterer hinzu: die Berücksichtigung von Legacy-Daten. Anpassungen an das neue TEI-Modell konnten zwar teilweise automatisch durchgeführt werden, stoßen aber an Grenzen. Manuelle Anpassungen sind jedoch aufgrund des Umfangs von mehreren Tausend Briefen nur eingeschränkt möglich. Dieses Kompatibilitätsproblem tritt einerseits auf technischer Ebene in Erscheinung, wo alle Daten gegen ein einheitliches ODD-Schema validiert werden sollen, andererseits aber ebenso auf inhaltlicher Ebene, was beispielsweise die Strukturierung des Schlagwortthesaurus betrifft.

Für die TEI-Annotation der Korrespondenzdaten drängen sich die Empfehlungen der *SIG Correspondence* und die entsprechende Weiterverarbeitung im CMI-Format (Dumont et al. 2019) auf, um den Zugriff über *correspSearch* (Dumont 2016 und Dumont/Grabsch/Müller-Laackman 2021) zu ermöglichen. Als Referenzimplementationen wurden die Briefeditionen der BBAW (insbesondere die beiden Humboldt-Editionen von Ette et al. 2020 und BBAW 2021) konsultiert. Für den Brieftext wurde Kompatibilität mit dem DTA-Basisformat (DTABf) angestrebt. Normdaten und Vokabulare wurden einerseits aus den in Vorarbeiten aufgebauten Thesauri bezogen, andererseits aus *authority files* wie VIAF, GND und *GeoNames*. Das ZIM-ACDH versucht durch interne Kodierungsrichtlinien für den TEI-Header einen Grundstock an Metadaten zu erzeugen, der in weiterer Folge für die Langzeitarchivierung in weitere Standards umgewandelt werden kann. Unter zusätzlicher Berücksichtigung der Legacy-Daten ergab die Zusammenwirkung dieser Anforderungen bereits mehrere Konflikte in der Annotation.

Lösungsansätze

Das Abstimmen verschiedener Anforderungen und der Abgleich mit den vorhandenen Daten nahm erhebliche Zeit in Anspruch. Danach wurde evaluiert, wo eine automatische Anpassung der Altdaten vertretbar ist und wo sich das ODD-Schema den Daten anpassen muss. Schließlich wurden die Konflikte in den unterschiedlichen Kodierungsvarianten besprochen und versucht, die passendste Lösung zu finden. Dieser Prozess mündete in einem Schema, das die gewünschten Eigenschaften vereinte, aber notgedrungen Abweichungen zu den Ausgangsschemata enthielt. Um diesen Schwachpunkt zu entschärfen, wurde versucht, sowohl die Abweichungen wie auch die Gründe dafür innerhalb und außerhalb des Schemas zu dokumentieren. Dies dient nicht nur in der Erfassung als Referenz für unterschiedliche BearbeiterInnen, sondern gewinnt vor allem in der erhofften Weiternutzung der Daten durch Dritte an Bedeutung.

Zusammenfassung und Ergebnisse

Der Prozess der Datenmodellierung muss zahlreiche Einflussgrößen berücksichtigen, wie am Beispiel des *Hugo Schuchardt Archivs* illustriert wurde. Dieser Prozess beinhaltet immer kritische Entscheidungen und Kompromisse, die sich aus dem Material, aber auch durch eingeschränkte Zeit- und Personalressourcen ergeben. Die Frage nach der bestmöglichen Kodierung kann da-

her nicht allgemeingültig beantwortet werden, vielmehr muss sie individualisiert betrachtet werden. Trotzdem können konstituierende Eigenschaften für eine *gute* Annotationspraxis beobachtet werden: sie sollte gut dokumentiert sein und unter der Berücksichtigung von FAIR-Data die Weiterverwendung erlauben. Gerade dem Aspekt der Metadaten, die für die Archivierung und Aggregation nach dem Ende des befristeten Projektes zentral sind, wird zu Beginn oft wenig Beachtung geschenkt. Für die nachhaltige Weiternutzung der Daten im wissenschaftlichen Kontext stellt dies jedoch einen essenziellen Bestandteil dar. In diesem Sinne sollte auch die Interoperabilität mit ähnlichen Ressourcen als Faktor bei Kodierungsentscheidungen in Betracht gezogen werden.

Fußnoten

1. Das Hugo Schuchardt Archiv wurde von zahlreichen Fördergebern berücksichtigt, hervorzuheben sind die letzten FWF-Projekte „Netzwerk des Wissens“ (P 24400, Bernhard Hurch 2012–2016) und „Philingk: Verlinktes Wissen zur Fachgeschichte“ (I 5076, Ursula Bähler und Bernhard Hurch 2021–2023).

Bibliographie

Berlin-Brandenburgischen Akademie der Wissenschaften (2011–2020): *DTABf. Deutsches Textarchiv – Basisformat*. <http://deustextarchiv.de/doku/basisformat> [letzter Zugriff 15. Juli 2021].

Berlin-Brandenburgischen Akademie der Wissenschaften (2021): *Wilhelm von Humboldt: Sprachwissenschaftliche Korrespondenz*. <https://wvh-briefe.bbaw.de> [letzter Zugriff 15. Juli 2021].

Dumont, Stefan (2016): "correspSearch – Connecting Scholarly Editions of Letters" in: *Journal of the Text Encoding Initiative* 10. <https://doi.org/10.4000/jtei.1742> [letzter Zugriff 15. Juli 2021].

Dumont, Stefan / Börner, Ingo / Müller-Laackmann, Jonas / Leipold, Dominik / Schneider, Gerlinde (2019): "Correspondence Metadata Interchange Format (CMIF)" in: Dumont, Stefan / Haaf, Susanne / Seifert Sabine (eds.): *Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf*. Berlin. <https://encoding-correspondence.bbaw.de/v1/CMIF.html> [letzter Zugriff 15. Juli 2021].

Dumont, Stefan / Grabsch, Sascha / Müller-Laackman, Jonas (2021): *correspSearch – Briefeditionen vernetzen*. Version 2.0.0. Berlin-Brandenburgische Akademie der Wissenschaften. <https://correspSearch.net> [letzter Zugriff 15. Juli 2021].

Ette, Ottmar et al. (2020): *edition humboldt digital*. Berlin-Brandenburgische Akademie der Wissenschaften. Version 6. <https://edition-humboldt.de> [letzter Zugriff 15. Juli 2021].

Hurch, Bernhard (2007–): *Hugo Schuchardt Archiv*. <http://schuchardt.uni-graz.at> [letzter Zugriff 15. Juli 2021].

Institut Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities (2021): *Geisteswissenschaftliches Asset Management System (GAMS)*. <https://gams.uni-graz.at> [letzter Zugriff 15. Juli 2021].

Steiner, Elisabeth / Stigler, Johannes (2018): "GAMS – Eine Infrastruktur zur Langzeitarchivierung und Publikation geisteswissenschaftlicher Forschungsdaten". In: *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 71: 207–216.

TEI Consortium (2021): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.2.2. <https://tei-c.org> [letzter Zugriff 15. Juli 2021].

Wilkinson, Mark D. et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship" in: *Sci. Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18> [letzter Zugriff 15. Juli 2021].

„Das Puzzle zusammensetzen“ Von analogen Dokumentensammlungen zu datenbankbasierten Biografien sowjetischer Kriegsgefangener des Zweiten Weltkriegs

Kindler, Sebastian

sebastian.kindler@dhi-moskau.org
Deutsches Historisches Institut Moskau, Germany

Wolf, Katrin

katrin.wolf@dhi-moskau.org
Deutsches Historisches Institut Moskau, Germany

Einleitung

Während des Zweiten Weltkriegs gerieten über fünf Mio. sowjetische Soldaten in deutsche Kriegsgefangenschaft. Charakterisiert war diese Gefangenschaft durch die Nichtbeachtung internationaler Regeln zum Umgang mit feindlichen Kriegsgefangenen und einer unmenschlichen Behandlung bis hin zur Ermordung der Internierten. Mehr als drei Millionen überlebten die Gefangenschaft nicht (Otto, Keller 2011; Wissenschaftliche Dienste 2010). Die Überlebenden wurden nach der Rückkehr in ihre Heimat des Landesverrats und der Kollaboration mit dem Feind verdächtigt und waren unterschiedlichen Repressionen ausgesetzt. Bis heute sind in vielen Fällen sowohl die Namen als auch die Schicksale dieser Personen nicht bekannt.

Im deutsch-russischen Gemeinschaftsprojekts „Sowjetische und deutsche Kriegsgefangene und Internierte“ ist es die Aufgabe des DHI Moskau, die heute noch verfügbaren Archivdokumente zu digitalisieren, die personenbezogenen Angaben zu den Kriegsgefangenen zu extrahieren und diese Information für Forschung und Schicksalsklärung öffentlich verfügbar zu machen. Das Projektziel besteht im Aufbau einer Datenbank, die die „Gefangenenbiografien“ der Kriegsgefangenen vom Zeitpunkt ihrer Gefangennahme, ihre Lagerstandorte bis hin zum Ende der Gefangenschaft durch Befreiung oder Tod sowie ggfs. eine Repression nach der Repatriierung und Filtration, d.h. der Suche nach tatsächlichen oder vermeintlichen „Staatsfeinden“, enthält. Voraussetzung für den Aufbau einer solchen digitalen Infrastruktur als Werkzeug für die zukünftige Forschung und Schicksalsklärung ist die Gewinnung, Organisation und Verarbeitung einer sehr großen Daten-

menge. Der dafür etablierte Workflow soll als „best practise“-Modell vorgestellt werden.¹

Der Workflow: Standards, wo möglich – Kompromisse, wo nötig

Die Digitalisierung der Archivdokumente bildet den ersten Schritt auf dem Weg vom Papierdokument zur digitalen Gefangenenbiografie und erfolgt nach Maßgabe der „DFG-Praxisregeln ‚Digitalisierung‘“ (DFG 2016). Aufgrund der Arbeit in einer Vielzahl internationaler Archive und daraus resultierender unterschiedlicher Rahmenbedingungen (technische Ausrüstung, Zustand der Dokumente, Kooperationsbereitschaft etc.) ist dieser „Goldstandard“ jedoch nicht immer umsetzbar.

Die Indexierung als Übertragen der auf dem Dokumentenscan enthaltenen Informationen in die digitale Form erfolgt sowohl projektintern als auch durch externe Dienstleister. Im ersten Schritt wird sie auf Basis des Dokuments ohne Änderung, Interpretation oder Änderung des Originaleintrags durchgeführt und verbindet die Informationen unmittelbar mit dem zugehörigen Digitalisat. Bei der Arbeit mit den teils sehr heterogenen Quellen (Provenienz, Entstehungszeit, verwendetes Alphabet, genutzte Sprache, enthaltene Informationen) bestehen für jeden Dokumententyp eigens angepasste Datenfelder. Die Auswahl der zu indexierenden Informationen orientiert sich an den Nutzungsbedürfnissen von Schicksalsklärung und historischer Forschung: Neben den verfügbaren Personenbasisdaten sind das Angaben zu Lageraufenthalt, Transporten, Arbeitskommandos sowie zum Ende der Gefangenschaft und der Repatriierung. Aufwand und Ertrag sind im Hinblick auf die einzelnen Dokumententypen elementar: Bei einer Vielzahl von Dokumenten sind nicht alle enthaltenen Informationen extrahierbar, eine zu starke Limitierung schränkt jedoch im Ergebnis auch die Nutzbarkeit der Gefangenenbiografien ein.

Die genutzte Datenbankinfrastruktur sind die „Memorial Archives“ der KZ-Gedenkstätte Flossenbürg.² Die im Rahmen des Projekts generierten Daten sind in mehrere Arbeitsebenen unterteilt: Die unterste Ebene bildet der aus indexierter Information und Digitalisat bestehende Dokumentendatensatz, der die jeweilige Archivale repräsentiert und die Basis aller weiteren Ebenen bildet. Die indexierten Informationen werden um eine Interpretation des originären Eintrags ergänzt, um Daten zu vereinheitlichen und vergleichbar zu machen. Der unveränderte Eintrag bleibt parallel als Referenz bestehen. Weitere Arbeitsebenen bestehen u. a. aus einer Aktenebene zur Repräsentation von Dokumentensammlungen, einer Ebene zu Transporten und Überstellungen in andere Lager und einer Ebene zu in den Dokumenten genannten dritten Personen. Die Teilebenen fließen auf der Ebene von Personendatensätzen zusammen, die die bereits mehrfach erwähnte Gefangenenbiografie darstellen: Alle zu einem Individuum gehörenden Informationen aus unterschiedlichen Dokumenten werden hier miteinander verbunden, sodass das auf Grundlage aller verfügbaren Informationen erstellte Gefangenen-schicksal sichtbar wird. Die Integration der Informationen zu „Transporten“ und „weiteren Personen“ löst diese Informationen von den individuellen Schicksalen und ermöglicht die Forschung zu gruppenbezogenen Erfahrungen nach geografischen und zeitlichen Kriterien ebenso wie nach familiären oder anderen persönlichen Netzwerken.

Fazit

Die Erstellung datenbankbasierter Gefangenenbiografien sowjetischer Kriegsgefangener des Zweiten Weltkriegs stellt auf mehreren Ebenen eine Herausforderung dar, bis aus einem Papierdokument eine digitalisierte Quelle mit Potenzial für Schicksalsklärung und Forschung wird. Bei dem vorgestellten Projekt sind es neben bilateraler deutsch-russischer Koordination der Arbeitsschwerpunkte auf politischer und inhaltlicher Ebene die technische Umsetzung der Digitalisierungs- und Indexierungsmaßnahmen sowie die detaillierte konzeptionelle Vorbereitung der einzelnen Arbeitsschritte. Darüber hinaus betreffen rechtliche und ethische Einschränkungen nicht nur die Nachnutzbarkeit der Digitalisate, sondern auch den Umgang mit den personenbezogenen Daten z.T. noch lebender Personen.

Diese und weitere Faktoren sind angesichts des immensen Quellenkorpus zu beachten, um eine Balance zwischen Qualität und Quantität der Verarbeitung zu gewährleisten und die Nutzbarkeit der Gefangenenbiografien für Wissenschaft und Erinnerungskultur zu maximieren.

Fußnoten

1. Aktuell (Stand 7/2021) verfügt das Projekt über mehr als 1,5 Mio. Digitalisate, aus denen die personenbezogenen Datensätze erstellt werden.
2. Abrufbar unter <https://memorial-archives.international>.

Bibliographie

Deutsche Forschungsgemeinschaft (2016): *DFG-Praxisregeln „Digitalisierung“* [12/16], https://www.dfg.de/formulare/12_151/12_151_de.pdf [letzter Zugriff 6. Juli 2021].

Otto, Reinhard / Keller, Rolf (2011): "Zur individuellen Erfassung von sowjetischen Kriegsgefangenen durch die Wehrmacht", in: *VfZ* 59: 563-577.

Overmans, Rüdiger / Hilger, Andreas / Poljan, Pavel [eds.] (2012): *Rotarmisten in deutscher Hand. Dokumente zu Gefangenschaft, Repatriierung und Rehabilitation sowjetischer Soldaten des Zweiten Weltkriegs*. Paderborn: Schöningh.

Wissenschaftliche Dienste des Bundestags [eds.]: *Sowjetische Kriegsgefangene in Deutschland 1941-1945*. <https://www.bundestag.de/resource/blob/414030/3224bdbbdaed8abc7b833e237a3cdc73/WD-1-036-10-pdf-data.pdf> [letzter Zugriff 6. Juli 2021].

Das zoroastrische Mittelpersische Digitale Corpus und Wörterbuch (MPCD)

Neuefeind, Claes

c.neuefeind@uni-koeln.de
Universität zu Köln

Mondaca, Francisco

f.mondaca@uni-koeln.de
Universität zu Köln

Eide, Øyvind

oeide@uni-koeln.de
Universität zu Köln

Colditz, Iris

Iris.Colditz@ruhr-uni-bochum.de
Ruhr-Universität Bochum

Jügel, Thomas

Thomas.Juegel@ruhr-uni-bochum.de
Ruhr-Universität Bochum

Rezania, Kianoosh

kianoosh.rezania@rub.de
Ruhr-Universität Bochum

Zeini, Arash

a.zeini@gmail.com
Freie Universität Berlin

Cantera, Alberto

alberto.cantera@fu-berlin.de
Freie Universität Berlin

Emanuel, Chagai

chagai17@gmail.com
Hebrew University Jerusalem

Shaked, Shaul

msshaul@mscc.huji.ac.il
Hebrew University Jerusalem

Einleitung

Mit diesem Beitrag möchten wir das Projekt "Das zoroastrische Mittelpersische - digitales Corpus und Wörterbuch (Middle Persian Corpus and Dictionary, MPCD)" vorstellen, das im April 2021 seine Arbeit aufgenommen hat. Das MPCD-Projekt wird von der DFG als Langfristvorhaben mit einer geplanten Laufzeit von insgesamt neun Jahren gefördert.¹ Das Vorhaben wird als Kooperationsprojekt der Universitäten Bochum, Berlin, Köln und Jerusalem durchgeführt. Während an den Standorten Bochum, Berlin und Jerusalem der Schwerpunkt auf der philologischen Erschließung des Korpus sowie des darauf aufbauenden Wörterbuchs liegt, ist auf Kölner Seite das Cologne Center for eHumanities (CCeH) für die technische Umsetzung einer kollaborativen Recherche- und Arbeitsumgebung zuständig, deren technischen Entwurf wir in diesem Poster-Beitrag thematisieren und zur Diskussion stellen wollen.

Gegenstand und Ziele des Projekts

Das Mittelpersische war als Amts- und Verkehrssprache insbesondere des Sasanidenreiches (3. – 7. Jhd.) von überkultureller und -religiöser Bedeutung. Von der Spätantike bis zur frühislamischen Zeit verbindet es sprachlich und kulturell die differierten Räume des iranischen Ostens und Westens. Das umfangreiche Korpus der mittelpersischen Texte ist bis heute jedoch nur partiell erschlossen. Ziel des MPCD-Projektes ist deshalb die Erstellung eines Korpus zoroastrisch-mittelpersischer Texte in Pahlavi-Schrift. Dieses mit Abstand größte mittelpersische Korpus (ca. 54 Texte, etwa 687.000 Wörter) wird in Transliteration und Transkription (vgl. dazu Rezania 2020) sowie in Handschriftenphotographien der 15 ältesten Codices, die zum Teil aus dem CAB-Projekt von Alberto Cantera (Corpus Avesticum Berolinense)² bezogen werden können, zugänglich sein. Die Texte werden morphosyntaktisch und lexikographisch annotiert und in TEI kodiert. Die morphosyntaktische Annotation der Texte folgt dem Standard *Universal Dependencies*³, der für die Annotation des Mittelpersischen angepasst wurde, indem das Subset der für die Annotation des Mittelpersischen notwendigen Tags bestimmt und die notwendigen pahlavi-spezifischen Tags hinzugefügt wurden.

Auf Grundlage des Korpus wird anschließend ein digitales Mittelpersisch-Englisch-Lexikon mit ca. 7000 Lemmata erstellt. Digitales Korpus und digitales Wörterbuch stellen im Projekt zwei eng verzahnte Analyseinstrumente dar, die mit unterschiedlichen Schwerpunkten – Syntax und Semantik – ineinandergreifen und auch im Arbeitsprozess eng miteinander verbunden sind. Hierbei kommt eine webbasierte Arbeitsumgebung zum Einsatz, die zum einen die kollaborative Bearbeitung von Korpus und Wörterbuch ermöglicht, zum anderen als Nutzer-Interface für Recherchen und Analysen der aufbereiteten Ressourcen dient.

Mit der Einarbeitung der 15 ältesten Codices in ein digitales Korpus mit darauf aufbauendem Wörterbuch schafft das Projekt einen methodisch neuen Zugang zum gesamten zoroastrisch-mittelpersischen Sprachmaterial, der die Voraussetzung für umfassende linguistische und begriffsgeschichtliche Fragestellungen eröffnet. Mit der engen Verzahnung von Text und Wörterbuch ergänzt das Vorhaben bestehende Textsammlungen zum Mittelpersischen wie bspw. TITUS⁴ (Thesaurus Indogermanischer Text- und Sprachmaterialien) und bildet eine wertvolle Aktualisierung gegenüber vorliegenden Wörterbüchern wie MacKenzie (1971) oder Nyberg (1964, 1974). Das Projekt bietet eine Grundlage dafür, das komplexe Gewebe der Texte der zoroastrisch-mittelpersischen Literatur in seinen internen und externen Bezügen zu identifizieren und damit einer (weithin ausstehenden) kultur- und religionshistorisch differenzierten Beschreibung zuzuführen. Es zielt damit darauf, die ‚horizontalen‘ (d.h. genrespezifischen) und die ‚vertikalen‘ (d.h. historischen) Differenzen der Texte und ihres Wortschatzes sichtbar werden zu lassen.

Systementwurf

Der Schwerpunkt des Posters liegt auf der Präsentation des Systementwurfs der kollaborativen Recherche- und Arbeitsumgebung sowie der dort eingesetzten Technologien. Dies umfasst zum einen eine Beschreibung der funktionalen Elemente der Nutzerschnittstelle, die dem Anwender als Forschungsumgebung dient, indem sie verschiedene Werkzeuge bereitstellt (z.B. Suche, Verknüpfung von Korpus und Wörterbuch, Export in TEI-Format).

Zum anderen werden die Systemarchitektur und die für deren Umsetzung verwendeten Technologien thematisiert.

Die Daten werden mithilfe des Python-Webframework Django⁵ modelliert und in PostgreSQL persistiert. Für die Suche in den Daten werden wir Elasticsearch⁶ einsetzen. Suche und CRUD-Operationen werden über eine REST-API verfügbar sein, die sich konzeptuell an Vorarbeiten aus dem Projekt VedaWeb⁷ orientiert (vgl. dazu Mondaca et al. 2019a, 2019b). Für das Frontend werden wir eine Single-Page-Applikation mit React.js⁸ entwickeln.

Wesentliche Funktionen des Frontend sind zum einen die Darstellung der digitalisierten Handschriften und der transliterierten und transkribierten Texte sowie die Möglichkeit zur differenzierten Suche nach linguistischen Parametern, die unter Verwendung von Konzepten aus dem VedaWeb-Projekt implementiert wird (vgl. Kiss et al. 2019). Zum anderen soll das Frontend im Sinne eines Redaktionssystems einen separaten Bereich für die Bearbeitung anbieten, der durch die Nutzerverwaltung nur angemeldeten Benutzern zugänglich ist. Ein solcher Bereich für Korpus und Wörterbuch kann unter Verwendung einer Nutzerverwaltung jeweils als separater View umgesetzt werden, in dem Korrekturen und (Neu-)Eingaben vorgenommen werden.

Mit der Fokussierung auf den Systementwurf möchten wir mit dem Poster vor allem einen kompakten Überblick über die Nutzungsmöglichkeiten sowie über die technische Architektur der geplanten Arbeitsumgebung geben.

Fußnoten

1. <https://gepris.dfg.de/gepris/projekt/452473565>
2. <https://www.geschkult.fu-berlin.de/e/iranistik/forschung/CAB/index.html>
3. <https://universaldependencies.org/>
4. <http://titus.uni-frankfurt.de/indexe.htm>
5. <https://www.djangoproject.com/>
6. <https://www.elastic.co/>
7. <https://vedaweb.uni-koeln.de/>
8. <https://reactjs.org/>

Bibliographie

Kiss, Börge / Kölligan, Daniel / Mondaca, Francisco / Neuefeind, Claes / Reinöhl, Uta / Sahle, Patrick (2019): "It Takes a Village: Co-developing VedaWeb, a Digital Research Platform for Old Indo-Aryan Texts." In: Steven Krauwer und Darja Fišer (Hg.), *TwinTalks at DHN 2019 – Understanding Collaboration in Digital Humanities*. Kopenhagen, 2019.

MacKenzie, David N. (1971): *A Concise Pahlavi Dictionary*. London: Oxford University Press.

Mondaca, Francisco / Rau, Felix / Neuefeind, Claes / Kiss, Börge / Kölligan, Daniel / Reinöhl, Uta / Sahle, Patrick (2019a): "C-SALT APIs - Connecting and Exposing Heterogeneous Language Resources." In: *Book of Abstracts of the Digital Humanities Conference 2019 (DH2019)* 09.07-12.07.2019. Utrecht, Netherlands.

Mondaca, Francisco / Schildkamp, Philip / Rau, Felix (2019b): "Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data". In: *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference, Sintra, Portugal*. Brno: Lexical Computing CZ, s.r.o., 907–921.

Nyberg, Henrik S. (1964): *A Manual of Pahlavi. Part I: Texts, Alphabets, Index, Paradigms, Notes and an Introduction*. Wiesbaden: Harrassowitz.

Nyberg, Henrik S. (1974): *A Manual of Pahlavi. Part II: Ideograms, Glossary, Abbreviations, Index, Grammatical Survey, Corrigenda to Part I*. Wiesbaden: Harrassowitz.

Rezania, Kianoosh (2020): "A Suggestion for the Transliteration of Middle Persian Texts in Zoroastrian Middle Persian: Digital Corpus and Dictionary (MPCD): A Three Layered Transliteration System". In: *Estudios Iranios y Turanios* 4: 153–73.

Datenbiographik im Literaturarchiv Konzept und Umsetzung digitaler Dienste am Theodor-Fontane- Archiv

Seifert, Sabine

sabine.seifert@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Busch, Anna

annabusch@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Trilcke, Peer

trilcke@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Genzel, Kristina

kristina.genzel@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Heilmann, Juliane

juliane.heilmann@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Möller, Klaus-Peter

klaus-peter.moeller@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Archiv und Biographik

Die Arbeit von Literaturarchiven steht seit deren ersten Konzeptualisierungen im 19. Jahrhundert (Dilthey 1970 [1889]; vgl. Thaler 2011, Schöttker 2016) in einem komplexen Wechselverhältnis zu den philologischen Tätigkeiten der Editorik und Biographik, die im 20. und 21. Jahrhundert noch ergänzt werden u.a. um Textgenetik und Material Media Studies. Während das Zusammenspiel von Archiv und Editorik dabei zuletzt vor dem Horizont der Digitalisierung intensiv diskutiert wird (vgl. exemplarisch Nutt-

Kofoth 2019), steht eine Neujustierung des Verhältnisses von Archiv und Biographik (vgl. Fetz 2009) im Zeichen der digitalen Transformation (Wettmann 2018) noch aus.

Im Zuge der digitalen Erweiterung seiner Dienste (Trilcke 2019; Trilcke, Busch, Seifert 2021) hat das Theodor-Fontane-Archiv in den vergangenen Jahren nicht nur bio- und bibliographische Datenbestände zu Fontane erstellt und offen im Web nutzbar gemacht, es hat auch an einem Konzept für eine digital-biographische Ressource und deren Umsetzung gearbeitet. Anders als für die Buch-Biographik typisch, wurde dabei kein narrativer Ansatz gewählt. Ziel war es vielmehr, eine Konzeption *datengetriebener Biographik* zu implementieren, die einem chronikalen Prinzip folgt.

Konzept der technischen Umsetzung

Orientierungspunkt für die Konzeption und (Daten-)Grundlage für die Umsetzung dieses datenbiographischen Dienstes war die fünfbandige Druckausgabe der *Theodor Fontane Chronik* (Bergbig 2010). Das Konzept wie auch das Vorgehen dieser *Fontane Chronik* wurde im Zuge der Umsetzung in zweifacher Hinsicht digitalisiert. Erstens folgt die Print-Chronik dem Prinzip der *Datenaggregation*: Biographische Informationen wurden aus zahlreichen Einzeldatenbeständen (Briefverzeichnisse und -ausgaben, Bibliographien etc.) systematisch zur Chronik zusammengetragen. Zweitens nimmt die Print-Chronik eine datumsbasierte *Datenkonsolidierung* vor, bei der jeder Datensatz einem übergeordneten Tages-, Monats- oder Jahresdatensatz zugewiesen wird. Der im April 2021 veröffentlichte Dienst *Fontane Chronik digital* übernimmt entsprechend nicht nur den Datenbestand der Print-Chronik, sondern implementiert auch diese beiden Prinzipien.

Die *Fontane Chronik digital* wird als frei nutzbarer Dienst für Ansicht, Recherche, Suche und Exploration verschiedener Datenbestände des Fontane-Archivs angeboten. Diese Datenbestände werden in der virtuellen Forschungsumgebung FuD, entwickelt vom Trier Center for Digital Humanities (TCDH), vorgehalten und gepflegt. Im Web sind sie als einzelne, vollwertige Dienste adressier- und recherchierbar. Die chronikalen Prinzipien sind in Form der algorithmischen Aggregation und Konsolidierung der unterschiedlichen Datenbestände in einer ElasticSearch-Instanz implementiert. Ein Frontend, das unterschiedliche Zugangsszenarien, Suchlogiken und Darstellungsformen berücksichtigt, präsentiert die Informationen usergerecht.

Einzeldatenbestände: Bibliographie, Briefdatenbank, Biographische Datenbank

Die Einzeldatenbestände, für die individuelle Dienste mit eigenen Interfaces entwickelt wurden, umfassen die wichtigsten Forschungsdaten zu Fontane. Als digitale Dienste werden dabei die *Fontane Briefdatenbank* und die *Fontane Bibliographie online* in Form offener Kulturdaten webbasiert und kostenfrei verfügbar gemacht. Ergänzt um die Daten der Biographischen Datenbank fließen die Datenbestände dieser Dienste in den übergreifenden Dienst *Fontane Chronik digital* ein.

Bibliographie

Die ursprünglich als Druck publizierte *Theodor Fontane Bibliographie* (Rasch 2006) enthält in systematischer Anordnung bibliographische Daten zu sämtlicher Primär- und Sekundärliteratur von und zu Fontane. 2019 konnte sie als Datenbank *Fontane Bibliographie online* veröffentlicht werden. Sie wird regelmäßig aktualisiert und enthält derzeit 17.641 Metadatensätze. Verknüpfungen gibt es intern u.a. zu den Digitalisaten der *Fontane Blätter retrodigital*. Zudem werden externe Ressourcen verlinkt (z.B. Digitalisate der *Vossischen Zeitung* in ZEFYS).

Briefdatenbank

Die *Fontane Briefdatenbank* verzeichnet sämtliche bekannte Briefe von Fontane. Informationen aus den archiveigenen Verzeichnisinstrumenten, dem 1988 publizierten 'Hanser Briefverzeichnis' und weiteren Quellen werden zusammengeführt, ergänzt und mit (inter-)nationalen Verbunddatenbanken verknüpft, z.B. correspSearch. Erfasst werden briefftypische Metadaten, Schlagworte zum Inhalt, Standortnachweise und Druckgeschichte. Mehr als 6.183 Datensätze sind über ein Interface durchsuchbar. Auf der Grundlage der Print-Chronik wurde die Briefdatenbank um 2.035 erschlossene Von-Briefe sowie 4.125 überlieferte und erschlossene An-Briefe ergänzt. Diese erweiterte Briefdatenbank bildet die Grundlage für die Datenweitergabe an die *Fontane Chronik digital*.

Biographische Datenbank

Aus XML-Druck-Daten der *Print-Chronik* wurden regelbasiert 19.183 Datensätze in das FuD-Datenbanksystem geparkt, wo sie aktualisiert und mit Norm- und Registerdaten angereichert werden. Die biographische Datenbank umfasst Datensätze zu folgenden Typen biographischer Daten: "Unternehmungen, Begegnungen, Ereignisse" (11.749 Datensätze), "Lektüren" (1.404), "Schriftstellerische und journalistische Arbeiten Fontanes" (2.263), "Druck von Publikationen Fontanes" (2.950), "Veröffentlichungen über Fontane" (692), "Wohnungen Fontanes" (125).

Fontane Chronik digital

Auf den drei Einzeldatenbeständen aufbauend operiert die *Fontane Chronik digital* (Abb.1) aggregierend und konsolidierend. Für diesen datenbiographischen Meta-Dienst wurde dabei ein Frontend entwickelt, das das chronikale Konsolidierungsprinzip in Form eines Kalender-Interfaces aufgreift.

Mit der Implementierung des chronikalen Prinzips in Form einer Datenbiographik ist ein entscheidender Entwicklungsschritt im digitalen Ausbau des Fontane-Archivs abgeschlossen. Auf der nun bestehenden Infrastruktur aufbauend, steht vor allem die Qualitätssteigerung der Daten (Normdaten, LOD) sowie die Anbindung und Öffnung qua APIs im Vordergrund der Entwicklungsarbeiten.

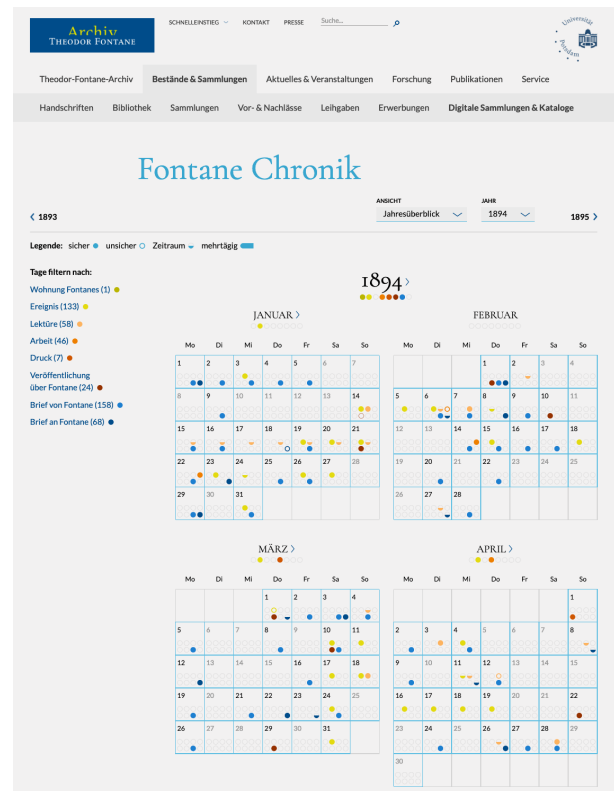


Abb. 1: Kalenderansicht der *Fontane Chronik digital*

Bibliographie

Digitale Dienste des Theodor-Fontane-Archivs

Berbig, Roland (Bearb.) / **Theodor-Fontane-Archiv** (ed.) (2021 ff.): *Theodor Fontane Chronik digital. Auf der Grundlage der "Theodor Fontane Chronik" (5 Bde., Berlin: De Gruyter 2010)*. Potsdam. <https://www.fontanearchiv.de/fontane-chronik> [letzter Zugriff 13. Juli 2021].

Rasch, Wolfgang (Bearb.) / **Theodor-Fontane-Archiv** (ed.) (2019 ff.): *Theodor Fontane Bibliographie online. Auf der Grundlage der "Theodor Fontane Bibliographie. Werk und Forschung" (3 Bde., Berlin: De Gruyter 2006)*. Potsdam. <https://www.fontanearchiv.de/fontane-bibliographie> [letzter Zugriff 13. Juli 2021].

Theodor-Fontane-Archiv (ed.) (2019 ff.): *Digitale Handschriftensammlung*. Potsdam. <https://www.fontanearchiv.de/fontane-handschriften> [letzter Zugriff 13. Juli 2021].

Theodor-Fontane-Archiv / Theodor Fontane Gesellschaft (eds.) (2019 ff.): *Fontane Blätter retrodigital*. Technische Betreuung: Universitätsbibliothek Potsdam. Potsdam. <https://www.fontanearchiv.de/fontane-blaetter> [letzter Zugriff 13. Juli 2021].

Theodor-Fontane-Archiv / UCLab Fachhochschule Potsdam (eds.) (2019): *Fontanes Handbibliothek visualisiert*. Design und Entwicklung: Mark-Jan Bludau. Projektkoordination: Anna Busch. Potsdam. <https://www.fontanearchiv.de/handbibliothek> [letzter Zugriff 13. Juli 2021].

Theodor-Fontane-Archiv (ed.) (2021 ff.): *Theodor Fontane Briefdatenbank. Unter Berücksichtigung von "Die Briefe Theodor Fontanes: Verzeichnis und Register" (München: Hanser*

1988). Potsdam. <https://www.fontanearchiv.de/fontane-briefdatenbank> [letzter Zugriff 13. Juli 2021].

Forschungsliteratur

Berbig, Roland (2010): *Theodor Fontane Chronik*. Projektarbeit 1999–2004: Josefine Kitzbichler. Berlin / New York: De Gruyter.

Dilthey, Wilhelm (1970 [1889]): “Archive der Litteratur in ihrer Bedeutung für das Studium der Geschichte der Philosophie”, in: *Archiv für Geschichte der Philosophie* II,3: 343–367.

Fetz, Bernhard (2009): “Der Stoff, aus dem das (Nach-)Leben ist. Zum Status biographischer Quellen”, in: Fetz, Bernhard (ed.): *Die Biographie. Zur Grundlegung ihrer Theorie*. Berlin / New York: De Gruyter 103–154.

Nutt-Kofoth, Rüdiger (2019): “Edition als Archiv? Zur Frage der Konvergenz zweier Wissensformationen im analogen und im digitalen Zeitalter”, in: Kastberger, Klaus / Maurer, Stefan / Neuhuber, Christian (eds.): *Schauplatz Archiv. Objekt – Narrativ – Performanz*. Unter Mitarbeit von Georg Hofer. Berlin / Boston: De Gruyter 107–123.

Rasch, Wolfgang (2006): *Theodor Fontane Bibliographie. Werk und Forschung*. Berlin: De Gruyter.

Schöttker, Detlev (2016): “Posthume Präsenz: Zur Ideengeschichte des literarischen Archivs”, in: Lepper, Marcel / Raulff, Ulrich (eds.): *Handbuch Archiv. Geschichte, Aufgaben, Perspektiven*. Stuttgart: J.B. Metzler Verlag 237–246.

Thaler, Jürgen (2011): “Zur Geschichte des Literaturarchivs. Wilhelm Diltheys Archive für Literatur im Kontext”, in: *Schiller-Jahrbuch* 55: 361–374.

Trilcke, Peer / Busch, Anna / Seifert, Sabine (2021): “Vom Ausbau des Digitalen Archivs. Neue digitale Dienste des Theodor-Fontane-Archivs. Chronik und Briefdatenbank”, in: *Fontane Blätter* 111: 173–183.

Trilcke, Peer (2019): “Auf dem Weg zu einem auch Digitalen Archiv. Digitale Dienste des Theodor-Fontane-Archivs”, in: *Fontane Blätter* 107: 98–103.

Wettmann, Andrea (2018): “Die Archive und der ‘Digital Turn’. Eine Standortbestimmung”, in: Bonte, Achim / Rehnolt, Julianne (eds.): *Kooperative Informationsinfrastrukturen als Chance und Herausforderung*. Berlin / Boston: De Gruyter 361–371. 10.1515/9783110587524-038.

Datenschutz in der wissenschaftlichen Praxis Der DARIAH-EU ELDAH Consent Form Wizard

Scholger, Walter

walter.scholger@uni-graz.at
Universität Graz; DARIAH-EU WG Ethics and Legality in Digital Arts and Humanities (ELDAH); CLARIN-ERIC Legal and Ethical Issues Committee (CLIC)

Hanneschläger, Vanessa

vanessa.hanneschlaeger@gmail.com
Österreichische Akademie der Wissenschaften (ÖAW); DARIAH-EU WG Ethics and Legality in Digital Arts and Humanities (ELDAH); CLARIN-ERIC Legal and Ethical Issues Committee (CLIC)

Kamocki, Pawel

pawel.kamocki@gmail.com
Institut für Deutsche Sprache (IDS); DARIAH-EU WG Ethics and Legality in Digital Arts and Humanities (ELDAH); CLARIN-ERIC Legal and Ethical Issues Committee (CLIC)

Kuzman-Šlogar, Koraljka

koraljkak@gmail.com
Universität Zagreb; DARIAH-EU WG Ethics and Legality in Digital Arts and Humanities (ELDAH)

Insbesondere im Bereich der Digital Humanities (DH) gilt der offene Zugang zu Wissen und den Ergebnissen wissenschaftlicher Forschung als Selbstverständlichkeit (vgl. Berliner Erklärung) und aufgrund der Vorgaben nationaler und europäischer Förderprogramme geradezu als Notwendigkeit (vgl. DFG, FWF, EU). Der Gedanke der Öffnung von Daten und Ergebnissen umfasst aber nicht nur das Ideal der Ermöglichung eines demokratischen Zugangs zu Forschungsergebnissen, sondern auch einen behutsamen und verantwortungsvollen Umgang mit den beforschten Quellen – besonders dann, wenn es sich dabei um lebende Menschen und deren Zeugnisse handelt. Vertraulichkeit und Datenschutz sind ethische Grundanforderungen für die Verarbeitung personenbezogener (Forschungs-)Daten.

Aus diesem Grund müssen sich Forscher*innen notwendigerweise mit komplexen Rechtsgrundlagen und ethischen Herausforderungen auseinandersetzen, wenn sie Daten über/von Menschen speichern, verwenden, verarbeiten, veröffentlichen und langzeitarchivieren wollen: Zunächst unüberwindbar scheint die Divergenz zwischen dem skizzierten Offenheitsparadigma und den rechtlichen Vorgaben sowie den wissenschaftsethischen Erfordernissen. Seit dem Inkrafttreten der EU-Datenschutzgrundverordnung (DS-GVO) gibt es einen europaweit anwendbaren, verbindlichen Rechtsrahmen für die Verarbeitung personenbezogener Daten, dessen Berücksichtigung unerlässlich für wissenschaftliche Tätigkeiten ist.

Wie eng dabei Ethik und Recht miteinander verwoben sind, zeigt sich etwa auch daran, dass jene Gruppen, die sich im Rahmen der großen europäischen DH-Forschungsinfrastrukturkonsortien DARIAH-EU und CLARIN-ERIC Rechtsfragen widmen, die Frage nach ethischen Dimensionen in ihr Programm (und ihre Bezeichnungen) aufgenommen haben: das CLARIN-ERIC Legal and Ethical Issues Committee (CLIC) und die DARIAH-EU Arbeitsgruppe Ethics and Legality in Digital Arts and Humanities (ELDAH).

Der Consent Form Wizard (<https://consent.dariah.eu/>) wurde von der DARIAH-EU Arbeitsgruppe ELDAH als Kooperationsprojekt von Jurist*innen, Entwickler*innen und Wissenschaftler*innen entwickelt. Dieses Werkzeug ermöglicht es, nach der Beantwortung einer Reihe einfacher Fragen eine standardisierte Einwilligungserklärung zu erstellen, die für das Einholen der Einwilligung von Studienteilnehmer*innen, Benutzer*innen, Veranstaltungs- oder Umfrageteilnehmer*innen etc. im Kontext wissenschaftlicher Datenerhebungen und -verarbeitungen verwendet

werden kann. Die so generierten Einwilligungserklärungen berücksichtigen die durch die DS-GVO geschaffenen Rahmenbedingungen und Verpflichtungen, aber auch die darin für Forschungs- und Archivierungskontexte definierten Ausnahmen (Kamocki, Ketzan, Wildgans 2018; Bergauer / Jahnel 2017). Sie können daher von der gesamten europäischen DH-Community - und auch in internationalen Kontexten, in denen sich die Verantwortlichen aus wissenschaftsethischen Gründen freiwillig den strengen europäischen Datenschutzvorschriften unterwerfen - verwendet werden. Die zum gegenwärtigen Zeitpunkt konzipierten Anwendungsszenarien des Consent Form Wizards reichen, basierend auf den Befragungen von TeilnehmerInnen mehrerer internationaler DH-Veranstaltungen und themenspezifischer Workshops, von Zustimmungserklärungen für die Datenverarbeitung im Rahmen der Veranstaltungsorganisation über Ton- und Videoaufzeichnungen, das Betreiben von Mailinglisten und Newsletters bis hin zur Verarbeitung personenbezogener Daten im Rahmen von Umfragen und Interviews in der Forschungspraxis.

Das englischsprachige Werkzeug ist auch als Quellcode auf GitHub frei verfügbar und wird gerade in mehrere Sprachen übersetzt: Die deutschen, kroatischen und französischen Übersetzungen sind bereits fertiggestellt. Außerdem wurden und werden eine Reihe von Disseminations-Workshops abgehalten und in Kooperation mit der Universität Jena im Rahmen der Digital4Humanities Reihe Tutorial Videos in deutscher und englischer Sprache produziert, die wie das Werkzeug zum Zeitpunkt der Konferenz auf mehreren Kanälen frei verfügbar sein werden und in die Materie der Datenschutzgrundverordnung und die Anwendung des Consent Form Wizards einführen. Diese Videos sind auch auf Youtube verfügbar und werden Teil eines auf der OER Plattform DARIAHcampus veröffentlichten Moduls sein.

Mittels des Posters bei der DHd2022 wollen wir den Consent Form Wizard der breiteren deutschsprachigen DH-Community jenseits von CLARIN-ERIC und DARIAH-EU vorstellen und Erfahrungsberichte weitergeben. Vor allem wollen wir zur Benutzung dieses Werkzeugs in der Praxis einladen und mögliche Erweiterungen - sei es durch weitere Übersetzungen, sei es durch die Umsetzung weiterer Anwendungsszenarien - mit den Teilnehmer*innen der Konferenz diskutieren.

Bibliographie

Bergauer, Christian / Jahnel, Dietmar (2017): *Das neue Datenschutzrecht DSGVO und DSGVO 2018*. Jan Sramek Verlag KG.

Europäisches Parlament (2016): *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG* (Datenschutz-Grundverordnung). <https://eur-lex.europa.eu/legal-content/DE/ALL/?uri=celex:32016R0679>.

Kamocki, Pawel / Erik Ketzan / Julia Wildgans (2018): *Language Resources and Research Under the General Data Protection Regulation*. <http://nbn-resolving.de/urn:nbn:de:bsz:mh39-97562>.

Der DHd Data Steward Maßnahmen zur Entwicklung einer nachhaltigen Datenstrategie für die Digital Humanities im deutschsprachigen Raum

Borges, Rebekka

RebekkaBorges@gmx.de

Rheinische Friedrich-Wilhelms-Universität Bonn, Deutschland

Debbeler, Anke

adebbell1@uni-koeln.de

Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Helling, Patrick

patrick.helling@uni-koeln.de

Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Der DHd Data Steward

Mit der Ernennung eines Data Stewards hat der Verband Digital Humanities im deutschsprachigen Raum e.V. (DHd) auf der Jahreskonferenz 2020 in Paderborn eine Funktion geschaffen, um eine umfassende Datenstrategie für alle Materialien, Publikationen und Ergebnisse, die im Kontext des DHd-Verbandes entstanden sind/entstehen werden, zu entwickeln.¹ Sie sollen im Sinne der FAIR-Prinzipien (Wilkinson et al. 2016) (1) langfristig gesichert und archiviert als auch (2) nach Möglichkeit nachhaltig publiziert und verfügbar gemacht werden.

Mit diesem Posterbeitrag soll die bisherige Arbeit des DHd Data Stewards präsentiert sowie ein Blick in die Datenzukunft des Verbandes gegeben werden.

Materialien der DHd-Konferenzen

Langfristige Verfügbarkeit von DHd-Konferenzwebsites

Einen zentralen Gegenstand und Zugangspunkt einer DHd-Jahreskonferenz stellt die Konferenzwebsite dar. Die Erstellung und das Hosting werden von den lokalen Organisator*innen übernommen. Nach einer Konferenz fehlt es häufig an notwendigen Mitteln für den langfristigen Betrieb dieser Konferenzwebsites.

Zur Gewährleistung einer dauerhaften Erreichbarkeit wurde damit begonnen, statische HTML-Versionen einzelner Konferenzwebsites zu erzeugen, um mögliche technische Abhängigkeiten aufzulösen und den Kurationsaufwand herunterzufahren. Mit einem Umzug auf die technische Infrastruktur der Alliance of Digital Humanities Organizations (ADHO) sollen die Konfe-

renzwebsites langfristig als Unterseite der Verbandswebsite bereitgestellt.²

Nachhaltige Publikation aller einzelnen DHd-Beiträge

Das Book of Abstracts einer Jahreskonferenz stellt ein wichtiges Publikationsinstrument dar. Seit 2016 wurden die Book of Abstracts via Zenodo publiziert (siehe Burr 2017; Stolz 2017; Vogeler 2018; Sahle 2019; Schöch 2020). Für die erste DHd-Jahreskonferenz 2014 wurde kein Book of Abstracts publiziert. Die Beiträge zur DHd-Jahreskonferenz 2015 sind als PDF-Datei über die Konferenzwebsite verfügbar (siehe Stiegler 2015).

Wenngleich die Books of Abstracts ein wichtiges Schaufenster der deutschsprachigen Digital Humanities (Sahle 2019; Schöch 2020) sind, ermöglichen sie weder eine eindeutige Zitierbarkeit einzelner Beiträge noch ihre Erfassung in digitalen Katalogen. Potentiale und Lösungsansätze zum Umgang mit DHd-Beiträgen wurden bereits in verschiedenen Formaten community-getrieben adressiert (Cremer 2018; Andofer 2019; Andorfer et al. 2019; Lordick 2020; Steyer et al. 2020).

Um die DHd-Beiträge einzeln zitierbar zu veröffentlichen, wurden alle Beiträge gesammelt:

- 2016, 2018-2020: TEI-Dateien via GitHub³
- 2017: TEI-Dateien von lokalen Organisator*innen
- 2015: gesammelte PDF-Datei von lokalen Organisator*innen
- 2014: gesammelte, unvollständige PDF-Datei von lokalen Organisator*innen + weitere Beiträge durch Autor*innen als PDF- und Word-Dateien

Zur Erstellung einzelner PDF-Dateien zu jedem Abstract wurde für die Jahrgänge 2016-2020 auf bestehende Transformationskripte zur Erstellung von Book of Abstracts aus TEI-Dateien zurückgegriffen.⁴ Die Skripte wurden angepasst, um für jedes Abstract einzeln eine PDF-Datei zu generieren.⁵

Zur persistenten Publikation der DHd-Beiträge wurde das generische Online-Repository Zenodo gewählt.⁶ Hier wird bereits seit 2019 eine DHd-Community als zentraler Publikationsort kuratiert.⁷

Mit Hilfe weiterer XSL-Transformationsskripte wurden für die Jahrgänge 2016-2020 jeweils eine Konferenz-Metadatendatei generiert, die dem DataCite-Schema und den Anforderungen von Zenodo entspricht (siehe Abb. 1).^{8,9}

```
<?xml version="1.0" encoding="UTF-8"?>
<metadaten:schema>
  <upload_type>publication</upload_type>
  <publication_type>conferencepaper</publication_type>
  <publication_date>2022-03-07</publication_date>
  <title>Der DHd Data Steward - Maßnahmen zur Entwicklung einer nachhaltigen Datenstrategie für die Digital Humanities im deutschsprachigen Raum</title>
</metadaten:schema>

<creators>
  <creator>
    <name>Borges, Rebekka</name>
    <affiliation>Rheinische Friedrich-Wilhelms-Universität Bonn, Deutschland</affiliation>
  </creator>
  <creator>
    <name>Debbeler, Anke</name>
    <affiliation>Data Center for the Humanities, Universität zu Köln, Deutschland</affiliation>
  </creator>
  <creator>
    <name>Helling, Patrick</name>
    <affiliation>Data Center for the Humanities, Universität zu Köln, Deutschland</affiliation>
  </creator>
</creators>

<description>A single abstract from the DHd-2022 Book of Abstracts.</description>
<access_right>open</access_right>
<license>cc-by</license>
<doi>
  <keywords>DHd2022, Forschungsdatenmanagement, FDM, Data Steward, Datenstrategie, Publikation, Archivierung, Veröffentlichung</keywords>
</doi>
<contributors>
  <contributor>
    <name>Gierkes, Michaela</name>
    <affiliation>Universität der Bundeswehr, München</affiliation>
    <type>editor</type>
  </contributor>
  <contributor>
    <name>Helling, Patrick</name>
    <affiliation>Digital Humanities im deutschsprachigen Raum e.V.</affiliation>
    <type>editor</type>
  </contributor>
</contributors>

<conference_title>DHd 2022 - Kulturen des digitalen Gedächtnisses</conference_title>
<conference_acronym>DHd 2022</conference_acronym>
<conference_dates>07.03.2022-11.03.2022</conference_dates>
<conference_place>Potsdam</conference_place>
<conference_url>https://www.dhd2022.de/</conference_url>

<committees>
  <committee>
    <name>https://doi.org/45.2353/beispieldoi.3409948</name>
    <notes>Sofern eine autorisierte Arbeit an dieser Publikation stattgefunden hat, dann bestand diese aus der Einleinerung von Bindestrichen in Überschriften, die aufgrund fehlerhafter Silbentrennung entstanden sind, der Voreinstellung von Namen der Autor*innen in das Schema "Nachname, Vorname" und/oder der Trennung von Überschrift und Unterüberschrift durch die Setzung eines Punktes, sofern notwendig.</notes>
  </committee>
</committees>
</metadaten:schema>
```

Abb. 1: Metadatenchema zu den DHd-Beiträgen mit Beispieleinträgen.

Für den Upload der Abstracts auf Zenodo wurde automatisiert für jedes Abstract ein Ordner (Bundle) mit PDF-Datei, TEI-Datei (sofern vorhanden) und Metadatenatz im JSON-Format generiert. Die Publikation dieser Bundles erfolgte automatisiert durch einen eigenen Publikationsworkflow (siehe Abb. 2) über die Zenodo REST API-Schnittstelle in die DHd-Zenodo Community.¹⁰ Die DHd-Beiträge verfügen über Digital Object Identifier (DOI), sind via OpenAIRE auffindbar und wurden durch die computer science bibliography (dblp) katalogisiert.^{11,12}

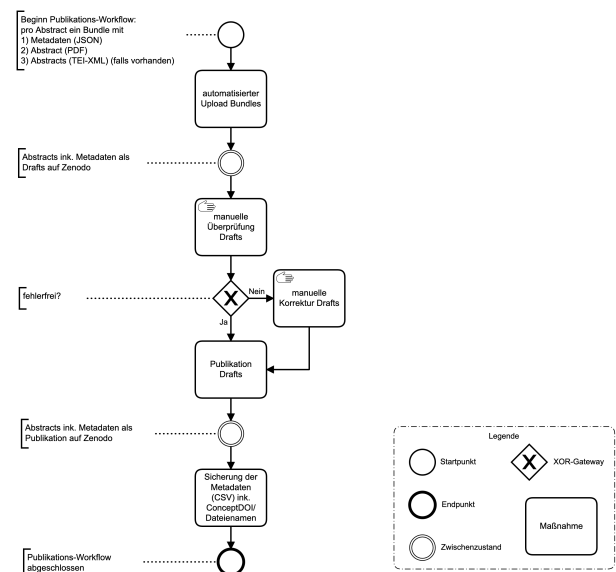


Abb. 2: Der Publikationsworkflow.

Ausblick und nächste Schritte

Der Umzug von alten Konferenzwebsites auf die ADHO-Infrastruktur soll vorangetrieben werden. Empfehlungen für die technische Umsetzung zukünftiger Konferenzwebsites soll deren Umzug auf die ADHO-Infrastruktur vereinfachen. Zusätzlich bedarf es einer Policy zur Speicherung, Archivierung und ggf. Publika-

tion von DHd-Materialien die außerhalb von der Jahreskonferenzen entstanden sind. Zuletzt sollen die entwickelten Workflows und die technische Realisation zur Sicherung und Publikation der einzelnen DHd-Abstracts im Rahmen einer durch die Community getragenen TaskForce zur kontinuierlichen Unterstützung der wechselnden Konferenz-Organisator*innen weiterentwickelt und optimiert werden. Die Einrichtung einer solchen TaskForce ist im Kontext der DHd 2022 Konferenz in Potsdam geplant. Ein Aufruf zur Beteiligung der Community wird im Vorfeld kommuniziert.¹³ Ein wichtiger Ansatz ist hier die Aufbereitung der TEI-Dateien, damit diese auch im Index of Digital Humanities Conferences (Weingart, Eichmann-Kalwara und Lincoln 2020) katalogisiert werden. Eine entsprechende Initiativgruppe hat sich bereits in einem Workshop auf der vDHd-Konferenz 2021 (Andorfer, Busch, Cremer et al. 2021) formiert.¹⁴

Fußnoten

1. <https://dig-hum.de/> [letzter Zugriff 10. November 2021].
2. Erste Website, die auf die ADHO-Infrastruktur umgezogen wurde: DHd-Jahreskonferenzwebsite 2020, Online: <https://dhd2020.dig-hum.de/> [letzter Zugriff 10. November 2021].
3. <https://github.com/peertrilcke/dhd2016-boa>; <https://github.com/GVogeler/DHd2018> [letzter Zugriff 10. November 2021]; Das GitHub-Repositorium mit den TEI-Dateien zur DHd-Jahreskonferenz 2019 ist nicht mehr verfügbar; <https://github.com/NinaSeemann/DHd2020-BoA> [letzter Zugriff 10. November 2021].
4. <https://github.com/karindalziel/TEI-to-PDF> [letzter Zugriff 10. November 2021].
5. <https://github.com/araborn/DHd2018> [letzter Zugriff 10. November 2021].
6. <https://zenodo.org/> [letzter Zugriff 10. November 2021].
7. <https://zenodo.org/communities/dhd> [letzter Zugriff 10. November 2021].
8. <https://github.com/reborg789/zenodup> [letzter Zugriff 10. November 2021].
9. <https://schema.datacite.org/> [letzter Zugriff 10. November 2021].
10. <https://github.com/cceh/zenodup> [letzter Zugriff 10. November 2021].
11. <https://www.openaire.eu/> [letzter Zugriff 10. November 2021].
12. <https://dblp.org/db/conf/dhd/index.html> [letzter Zugriff 10. November 2021].
13. Weitere Kontaktmöglichkeit zum Data Steward: <https://dig-hum.de/dhd-data-steward> [letzter Zugriff 17. November 2021].
14. Dieses Poster ist komplementär zum Posterbeitrag „Strukturen und Impulse zur Weiterentwicklung der DHd-Abstracts“ der Initiativgruppe auf der DHd 2022.

Bibliographie

- Andorfer, Peter** (2019): *dhd-boas-app*, Online: <https://dhd-boas-app.acdh-dev.oeaw.ac.at/> (letzter Zugriff: 14. Juli 2021).
- Andorfer, Peter / Busch, Anna / Cremer, Fabian / Henrich, Andreas / Helling, Patrick / Lordick, Harald / Mischke, Dennis / Steyer, Timo** (2021): "Bericht zur vDHd2021-Veranstaltung: Zukunftslabor DHd-Abstracts". DHd-Blog, Online: <https://dhd-blog.org/?p=15980> (letzter Zugriff: 14. Juli 2021).

Andorfer, Peter / Cremer, Fabian / Steyer, Timo (2019): "DHd 2019 Book of Abstracts Hackathon", Beitrag auf der *DHd 2019 Digital Humanities multimedial und multimodal*. 6. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2019). Frankfurt am Main und Mainz, Online: <http://doi.org/10.5281/zenodo.4622102>.

Burr, Elisabeth (ed.) (2017). *DHd 2016 Modellierung - Vernetzung - Visualisierung. Die Digital Humanities als Fächerübergreifendes Forschungsparadigma. Konferenzabstracts*. Leipzig, Online: <http://doi.org/10.5281/zenodo.3679331>.

Cremer, Fabian (2018): „Nun sag, wie hältst Du es mit dem Digitalen Publizieren, Digital Humanities?“. Digitale Redaktion Blog, Online: <https://editorial.hypotheses.org/113> (letzter Zugriff: 14. Juli 2021).

Lordick, Harald (2020): *DH(d) Konferenzbeiträge*, Online: <http://www.steinheim-institut.de/dhd/> (letzter Zugriff: 14. Juli 2021).

Sahle, Patrick (ed.) (2019): *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*. Frankfurt am Main, Online: <https://doi.org/10.5281/zenodo.2596095>.

Schöch, Christof (ed.) (2020): *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*. Paderborn, Online: <https://doi.org/10.5281/zenodo.3666690>.

Steyer, Timo / Andorfer, Peter / Cremer, Fabian (2020): „Abstract Enhancement. Potentiale der DHd-Konferenzabstracts als Daten/Publikation“, Beitrag auf der *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. 7. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2020). Paderborn, Online: <http://doi.org/10.5281/zenodo.4621706>.

Stiegler, Johannes (ed.) (2015): *DHd 2015 Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation. Konferenzabstracts*. Graz, Online: <https://dhd2015.uni-graz.at/de/nachlese/book-of-abstracts/> (letzter Zugriff: 14. Juli 2021).

Stolz, Michael (ed.) (2017): *DHd 2017 Digitale Nachhaltigkeit. Konferenzabstracts*. Bern, Online: <http://doi.org/10.5281/zenodo.3684825>.

Vogeler, Georg (ed.) (2018): *DHd 2018 Kritik der digitalen Vernunft. Konferenzabstracts*. Köln, Online: <http://doi.org/10.5281/zenodo.3684897>.

Weingart, Scott B. / Eichmann-Kalwara, Nickoal / Lincoln, Matthew (2020): *The Index of Digital Humanities Conferences*. Carnegie Mellon University, Online: <https://dh-abstracts.library.cmu.edu/> (letzter Zugriff: 14. Juli 2021).

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand Jan / Appleton, Gabrielle / Axton, Myles / Baak, Arie / Blomberg, Niklas / Boiten, Jan-Willem / da Silva Santos, Luiz Bonino / Bourne, Philip E. / Bouwman, Jildau / Brookes, Antony J. / Clark, Tim / Crosas, Mercè / Dillo, Ingrid / Dumon, Oliver / Edmunds, Scott / Evelo, Chris T. / Finkers, Richard / Gonzalez-Beltran, Alejandra / Gray, Alasdair J.G. / Groth, Paul, Goble, Carole / Grethe, Jeffrey S. / Heringa, Jaap / A.C't Hoen, Peter / Hooft, Rob / Kuhn, Tobias / Kok, Ruben / Kok, Joost / Lusher, Scott J. / Martone, Maryann E. / Mons, Albert / Packer, Abel L. / Persson, Bengt / Rocca-Serra, Philippe / Roos, Marco / van Schaik, Rene / Sansone, Susanna-Assunta / Schultes, Erik / Sengstag, Thierry / Slater, Ted / Strawn, George / Swertz, Morris A. / Thompson, Mark / van der Lei, Johan / van Mulligen, Erik / Velterop, Jan / Waagmeester, Andrea / Wittenburg, Peter / Wolstencroft, Katherine / Zhao, Jun / Mons Barend (2016): "The FAIR Guiding Principles for scientific data management and stewardship" in:

Scientific Data 3, Article number: 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>.

Der Dienstekatalog der AG Datenzentren

Ein digitales Verzeichnis für Forschungsdatenmanagement-Services in den Geisteswissenschaften

Rau, Felix

f.rau@uni-koeln.de

Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Helling, Patrick

patrick.helling@uni-koeln.de

Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Ausgangssituation

Die langfristige Sicherung, Verfügbarkeit und Nachnutzbarkeit von Forschungsdaten im Sinne der FAIR-Prinzipien (Wilkinson et al. 2016) ist ein wesentlicher Bestandteil guter wissenschaftlicher Praxis (DFG 2019) und nicht erst seit den Bestrebungen hin zu einer Nationalen Forschungsdateninfrastruktur (NFDI) (RfII 2016, 2017) ein wichtiger Motor wissenschaftlichen Fortschritts (Bryant, Lavoie & Maipas 2017).¹ An vielen Universitäten und außeruniversitären Einrichtungen wurden zur Unterstützung von Forschenden bei Fragen des Forschungsdatenmanagements (FDM) entsprechende, i.d.R. generisch ausgerichteten, Kompetenzzentren aufgebaut.²

Die Bedienung fachspezifischer FDM-Bedarfe in den Geisteswissenschaften übernehmen u.a. die Mitgliedsinstitutionen der DHd-AG Datenzentren. Die 2014 gegründete AG versteht sich dabei als offenes Forum, um Herausforderungen im geisteswissenschaftlichen Forschungsdatenmanagement gemeinsam zu adressieren.³ Die Arbeitsgruppe besteht aus insgesamt 28 Datenzentren und Interessensvertreter*innen, von denen 16 Einrichtungen über FDM-Servicestrukturen verfügen. Sie ergänzen das Forschungsdatenmanagement an ihren Standorten und darüber hinaus um ein geisteswissenschaftliches Profil und erarbeiten passgenaue Lösungsstrategien für die Forschungs- und Datenlandschaft der Geisteswissenschaften, die sich durch eine starke Heterogenität auszeichnet (Pempe 2012).

Um die Sichtbarkeit und Erreichbarkeit einzelner Datenzentren der AG und ihrer FDM-Services zu verbessern, hat die Arbeitsgruppe einen gemeinsamen Dienstekatalog (Helling, Moeller und Mathiak 2018) entwickelt, der als durchsuchbare Website verfügbar gemacht wurde.⁴

Vorgehensweise

Für die Erstellung des Katalogs wurden ca. einstündige Telefon-/Skypeinterviews mit Vertreter*innen aller AG-Mitgliedsinstitutionen geführt, die über FDM-Servicestrukturen verfügen. Die Interviews wurden nach einem Gesprächsleitfaden strukturiert, der sich wiederum an einer zuvor durchgeführten Online-Selbstauskunft innerhalb der AG orientiert. Der Leitfaden wurde den Befragten im Vorfeld der Gespräche zur Verfügung gestellt und ist auch online nachnutzbar.⁵ Er besteht aus 57 Fragen, die in zwei Kernbereiche unterteilt sind: Der erste Fragenkomplex zum *Profil der Datenzentren* behandelt unter anderem die institutionelle Anbindung der Datenzentren sowie Kooperationen und Zukunftsperspektiven. Der zweite Fragenkomplex *Dienstleistungen der Datenzentren* bezieht sich auf konkrete Services und Angebotsstrukturen. Insgesamt wurden dabei acht Servicebereiche behandelt:

- Allgemeines Beratungsangebot
- Bereitstellung/Vermittlung von technischen Infrastrukturen
- Konsolidierung von Services
- Speicherung und Archivierung
- Repositoriums-Lösungen
- Datenkuratierung
- Softwarekuratierung
- Softwareentwicklung

Quantitative Auswertung der Ergebnisse

Allgemeines Beratungsangebot

Nahezu alle Datenzentren beraten geisteswissenschaftliche Forscher*innen bei Fragen zum Management von Forschungsdaten in der Breite (siehe Abb. 1). Dabei erfolgt allerdings nicht jede Beratung zwangsläufig durch das jeweilige Datenzentrum. Insbesondere bei der Bedienung rechtlicher Fragestellungen gab die Mehrheit der Zentren an mit anderen Kompetenzstellen an ihren Einrichtungen zusammenzuarbeiten, bzw. an diese zu vermitteln.

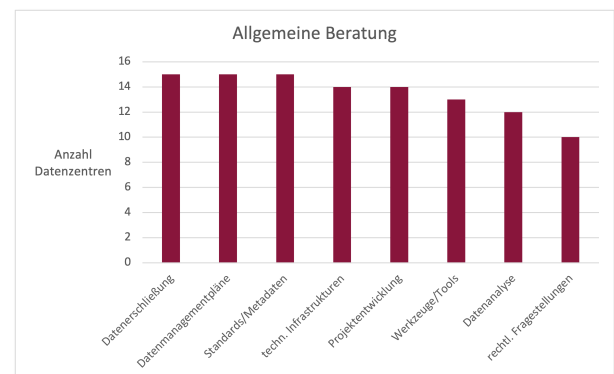


Abb. 1: Allgemeine Beratungskompetenzen der Datenzentren.

Bereitstellung/Vermittlung von technischen Infrastrukturen

Technische Infrastrukturen von zentralen Einrichtungen wie bspw. lokalen IT- oder Rechenzentren werden i.d.R. nachgenutzt und vermittelt. Auf diese Weise sind alle Datenzentren in der Lage mittelbar Speicherbedarfe zu bedienen und virtuelle Maschinen, Server sowie Netzwerke zur Verfügung zu stellen (siehe Abb. 2).

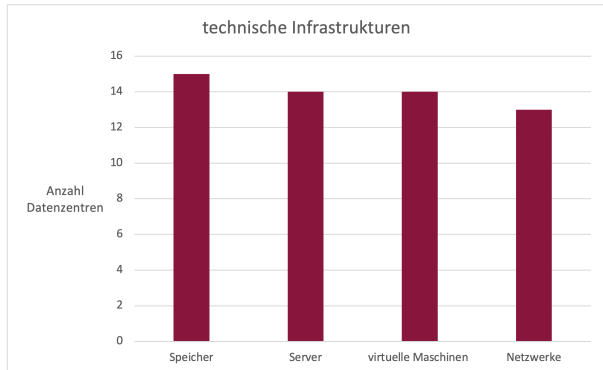


Abb. 2: Vermittlungskompetenz bei technischen Infrastrukturangeboten.

Konsolidierung von Services

Spezifische Übernahme- und Dokumentationsprozesse sowie konkrete Geschäftsmodelle in Form von Service Level Agreements (SLA) oder Verträgen befinden sich bei vielen Datenzentren noch in einer Entwicklungsphase (siehe Tab. 1).

Übernahmeprozesse	Anzahl Datenzentren
Dokumentationsprozesse	10
Verträge	8
Kurationsplanung	8
Service Level Agreements	7

Tab. 1: Übernahmeprozesse und Dokumentationsstandards.

Speicherung und Archivierung/Repositorien

Alle Datenzentren unterstützen aktiv bei der Speicherung von Forschungsdaten. Die meisten von ihnen helfen Forscher*innen zusätzlich auch bei der Langzeitarchivierung oder übernehmen diese direkt selbst. Insgesamt werden 21 generische und fachspezifische Repositorien von den Mitgliedsinstitutionen der AG Datenzentren unterhalten (siehe Tab. 2).

Speicherung/Archivierung	Anzahl Datenzentren
Speicherung	16
Langzeitarchivierung	13
Generische Repositorien	12
Fachspezifische Repositorien	9

Tab. 2: Speicherungs- und Archivierungsservices der Datenzentren.

Datenkuratierung

Während die langfristige Kuratierung von Forschungsdaten grundsätzlich einen Kernbereich aller befragten Datenzentren darstellt, unterscheiden sich die konkreten Services und Dienste zur Datenkuratierung zwischen den einzelnen Datenzentren untereinander (siehe Abb. 3).

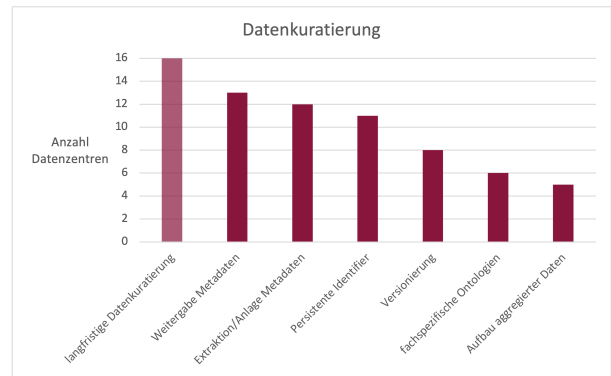


Abb. 3: Services zur Kuration von Forschungsdaten.

Softwarekuratierung und Softwareentwicklung

Noch wenige Datenzentren verfügen über Expertise beim Hosting und Betrieb lebender Systeme wie bspw. Websites, Visualisierungen, dynamischen Datenbanken und sonstigen Anwendungssystemen (siehe Tab. 3). Unterstützung bei der Entwicklung von unterschiedlicher Software bieten hingegen viele Datenzentren an (siehe Tab. 4).

Softwarekuratierung	Anzahl Datenzentren
Hosting lebender Systeme	5
Dauerbetrieb lebender Systeme	5

Tab. 3: Kompetenzverteilung beim Hosting und Betrieb lebender Systeme.

Softwareentwicklung	Anzahl Datenzentren
Tools	13
Portale	10
Schnittstellen	10
Repositorien	2

Tab. 4: Unterstützung bei der Entwicklung von Software.

Bereitstellung des Dienstekatalogs der AG Datenzentren

Um die gewonnenen quantitativen Ergebnisse, die einen Überblick über die aktuelle, fachspezifische FDM-Versorgungslandschaft geisteswissenschaftlicher Forscher*innen im deutschsprachigen Raum liefern, der Forschungscommunity sinnvoll verfügbar zu machen, wurden die sichtbar gemachten Servicestrukturen der einzelnen Datenzentren in eine durchsuchbare Wordpress-Website überführt. Den Kern dieser Ergebnispräsentation stellen einzelne Profilseiten aller Datenzentren dar, in denen die Servicestrukturen aufbereitet in Tabellenform adressierbar gemacht wurden. Neben der Zugänglichmachung der Services über die einzelnen Datenzentren kann die Website auch gezielt nach einzelnen Services durchsucht werden.

Die Wordpress-Website wurde mittlerweile in eine statische HTML-Version überführt und via GitHub publiziert, um den Aufwand für Betrieb und technischer Kuration möglichst gering zu halten. Die inhaltlich-redaktionelle Kuration des Dienstekatalogs obliegt der AG Datenzentren.

Fußnoten

1. Nationale Forschungsdateninfrastruktur (NFDI), Online: <https://www.nfdi.de/> (letzter Zugriff: 14. Juli 2021); Gemeinsame Wissenschaftskonferenz (GWK): Informationsinfrastrukturen/NFDI, Online: <https://www.gwk-bonn.de/themen/weitere-arbeitsgebiete/informationsinfrastrukturen-nfdi/> (letzter Zugriff: 14. Juli 2021).
2. Forschungsdaten.org: Sammlung „FDM-Kontakte“, Online: <https://www.forschungsdaten.org/index.php/FDM-Kontakte> (letzter Zugriff: 14. Juli 2021).
3. Arbeitsgruppe Datenzentren des Verbands „Digital Humanities im deutschsprachigen Raum e.V.“, Online: <https://dhd-ag-datenzentren.github.io/> (letzter Zugriff: 14. Juli 2021).
4. Dienstekatalog der Arbeitsgruppe Datenzentren, Online: <https://dhd-ag-datenzentren-dienstekatalog.github.io/> (letzter Zugriff: 14. Juli 2021).
5. Fragebogen zur Entwicklung eines Dienstekatalogs der AG Datenzentren im Verband Digital Humanities im deutschsprachigen Raum e.V., Online: <http://doi.org/10.5281/zenodo.5101280>.

Bibliographie

- Bryant, Rebecca / Lavoie, Brian / Malpas, Constance** (2017): *A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1*. Dublin, Ohio: OCLC Research. DOI: <https://doi.org/10.25333/C3PG8J>.
- DFG - Deutsche Forschungsgemeinschaft** (2019): *Guidelines for Safeguarding Good Research Practice. Code of Conduct*. Zenodo: <http://doi.org/10.5281/zenodo.3923602>.
- Helling, Patrick / Moeller, Katrin / Mathiak, Brigitte** (2018): „Forschungsdatenmanagement in den Geisteswissenschaften – der Dienstekatalog der AG-Datenzentren des Verbands Digital Humanities im deutschsprachigen Raum“ in: *ABI Technik*, Band 38, Heft 3, Seiten 251–261, ISSN (Online) 2191-4664, ISSN (Print) 0720-6763, DOI: <https://doi.org/10.1515/abitech-2018-3006>.
- Pempe, Wolfgang** (2012): „Geisteswissenschaften“ in: Neuroth, Heike / Strathmann, Stefan / Oßwald, Achim / Scheffel, Regine / Klump, Jens / Ludwig, Jens (eds.): *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Boizenburg: Verlag Werner Hülsbusch 137-160.
- RfII - Rat für Informationsinfrastrukturen** (2016): *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen. Online: <https://d-nb.info/1104292440/34> (letzter Zugriff: 14. Juli 2021).
- RfII - Rat für Informationsinfrastrukturen** (2017): *Schritt für Schritt - oder: Was bringt wer mit? Ein Diskussionsimpuls für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI)*. Göttingen, Online: <https://d-nb.info/1131083113/34> (letzter Zugriff: 14. Juli 2021).
- Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand Jan / Appleton, Gabrielle / Axton, Myles / Baak, Arie / Blomberg, Niklas / Boiten, Jan-Willem / da Silva Santos, Luiz Bonino / Bourne, Philip E. / Bouwman, Jildau / Brookes, Antony J. / Clark, Tim / Crosas, Mercè / Dillo, Ingrid / Dumon, Oliver / Edmunds, Scott / Evelo, Chris T. / Finkers, Richard / Gonzalez-Beltran, Alejandra / Gray, Alasdair J.G. / Groth, Paul, Goble, Carole / Grethe, Jeffrey S. / Heringa, Jaap / A.C't Hoen, Peter / Hooft, Rob / Kuhn, Tobias / Kok, Ruben / Kok, Joost / Lusher, Scott J. / Martone, Maryann**

E. / Mons, Albert / Packer, Abel L. / Persson, Bengt / Rocca-Serra, Philippe / Roos, Marco / van Schaik, Rene / Sansone, Susanna-Assunta / Schultes, Erik / Sengstag, Thierry / Slater, Ted / Strawn, George / Swertz, Morris A. / Thompson, Mark / van der Lei, Johan / van Mulligen, Erik / Velterop, Jan / Waagmeester, Andrea / Wittenburg, Peter / Wolstencroft, Katherine / Zhao, Jun / Mons Barend (2016): “The FAIR Guiding Principles for scientific data management and stewardship” in: *Scientific Data* 3, Article number: 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>.

DH2go

Lehr- und Lernumgebung für die Digital Humanities

Heckelen, Malte

malte.heckelen@ilw.uni-stuttgart.de
Universität Stuttgart, Germany

Schlesinger, Claus-Michael

claus-michael.schlesinger@ilw.uni-stuttgart.de
Universität Stuttgart, Germany

Burkhard, Fabienne

st170328@stud.uni-stuttgart.de
Universität Stuttgart, Germany

Mit unserem Poster möchten wir unsere Lehr- und Lernumgebung DH2go vorstellen. DH2go ist eine Remote-Desktop-Umgebung, die Strukturen für Kurse und Workshops sowie gängige DH-Technologien bietet. Gestartet als Experiment für unsere eigene DH-Lehre an der Universität Stuttgart, entwickelten wir die Umgebung über drei Jahre weiter und konnten ihre Features in der Lehre erproben.

DH2go nutzt die Remote-Desktop-Lösung X2Go, die das Einwählen auf einem Server mit eigener Benutzeroberfläche ermöglicht. Kursteilnehmer*innen wählen sich von ihrem eigenen Rechner aus mithilfe eines Clients auf auf einem Server ein und können dort mit einer grafischen Benutzeroberfläche arbeiten. Shell-Zugriff via SSH ist ebenfalls möglich. Es werden nahezu ausschließlich Open Source - Technologien eingesetzt und unsere Dokumentation erlaubt die Replikation auf eigenen Servern.

Nebst gängigen, vorinstallierten DH-Technologien, etwa für die Arbeit mit XML/TEI oder Stilometrie und Topic Modeling, bietet DH2go spezielle Tools für die Kursorganisation: Ordner für Kursmaterialien erlauben es Kursleiter*innen unkompliziert, die Teilnehmer*innen mit Code, Daten oder Medien zu versorgen. Teilnehmer*innen können darüber hinaus über einen Tauschordner für Gruppenarbeiten Dateien hin- und herschieben.

Wir entwickelten DH2go zunächst für ein spezifisches Lehrformat: eine einführende Übung zu grundlegenden Methoden und Techniken in den Digital Humanities.. Die Einführung neuer Methoden und Tools alle 2-3 Wochen führte zu Problemen: Variierende Betriebssysteme, kryptische Fehlermeldungen und Installationsprobleme verhinderten die effektive Vermittlung von Inhalten und der häufige Support am Heimrechner der Teilnehmer*innen war für uns in punkto Privacy problematisch. Insbe-

sondere im Zuge der Corona-Krise haben sich diese Probleme aufgrund der Remote-Lehre potenziert.

Als einheitliche Arbeitsumgebung konnten wir mit DH2go also Probleme lösen, die auch für andere Lehr- und Lernszenarien in den Digital Humanities relevant sind und die sich durch das Ziel zusammenfassen lassen, den Einstieg in die computergestützte Datenverarbeitung zu ebnen und einen reflektierten Umgang mit Methoden und Softwaretools zu vermitteln. DH2go als Arbeitsumgebung mit allen nötigen Tools ready to go unterstützt dieses Ziel in drei Aspekten: 1. reproduzierbare Abläufe, dadurch vereinfachte und wiederholbare Vermittlung von Arbeitsschritten und verbesserte Interaktion der Teilnehmer*innen; 2. besserer Support durch a) strukturierte Anleitungen, b) Supportsystem (First- und Second-Level-Support), c) vereinfachte gegenseitige Unterstützung der Teilnehmer*innen; 3. Schutz der Privatsphäre durch Entkopplung der Arbeitsumgebung von den Privatrechnern der Lernenden (und Lehrenden).

In verschiedenen Anwendungsszenarien konnten wir feststellen, dass DH2go insbesondere für workshopartige Lehr- und Lernformen geeignet ist, wo Inhalte nicht "von Anfang an" vermittelt werden sollen, aber Anpassungen durch Teilnehmer*innen, etwa bei Python-Skripts, dennoch möglich sein müssen. Dort, wo Interaktionen zwischen Teilnehmer*innen Teil des didaktischen Konzepts sind, zeigen sich die Vorteile von DH2go: unkomplizierter Austausch zu den Funktionen, dieselben Oberflächen für alle Teilnehmer*innen und Leiter*innen und gute Anpassbarkeit an die Erfordernisse bestimmter Kurse und ihrer Teilnehmer.

Weiter mussten wir den Wert eines guten Supportsystems erst durch Erfahrung kennenlernen. Die Nutzung von Infrastruktur - auch 'alle Studierenden nutzen ihre eigenen Laptops im Kurs' ist Infrastruktur - setzt ein funktionierendes Supportsystem voraus. Die Entwicklung eines strukturierten Systems mit auf die Zielgruppe abgestimmten Manuals und einem First- und Second-Level-Support kann daher als wichtiger stabilisierender Faktor für alle mit DH2go durchgeführten Kurse und Workshops gelten.

In der Lehre zeigten sich ebenfalls einige positive Effekte. Da alle Teilnehmer*innen nur eine einzige Software installieren müssen, sind die üblichen Installationsprobleme kein Zeitfaktor mehr. Darüber hinaus können Teilnehmer*innen die Vorgehensweisen der Kursleiter*innen dank der identischen Benutzeroberflächen direkt replizieren. Ein weniger direktes, unerwartetes Resultat ist die erhöhte Fragebereitschaft bei technischen Problemen: statt das Problem auf eigene Fehler zu beziehen, wird es als Bug empfunden, den man melden sollte. Ein didaktischer Vorteil, wenn auch das Supportvolumen steigt.

Zurzeit ist DH2go auf Anfrage als Service nutzbar oder über unsere Dokumentation und Images auf den eigenen Server-Architekturen replizierbar. Für unsere fortschreitende Arbeit an Konzept und Implementierung freuen wir uns auf spannende Gespräche.

DiaCollo für GEI-Digital

Ein experimentelles Projekt zur weiteren Erschließung digitalisierter historischer Schulbuchbestände

Niëländer, Maret

nielaender@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Germany

Jurish, Bryan

jurish@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Scheel, Christian

scheel@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Germany

Seit 2009 digitalisiert die Forschungsbibliothek des Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut (GEI) mit Förderung der DFG seine historischen Bestände. Die „Digitale Schulbuchbibliothek GEI-Digital“¹ umfasst mittlerweile etwa 6000 vor 1920 erschienene Werke, v.a. deutschsprachige Realienkundebücher, Fibeln und Lesebücher sowie Bücher für die Fächer Geographie, Geschichte, Politik und Religion. Die Digitalisate, Metadaten und OCR-generierten Volltexte können online genutzt werden und stehen in verschiedenen Formaten unter der Lizenz CC0 zum Download zur Verfügung (Hertling / Klaes 2018a, 2018b).

Das Projekt

Das experimentelle, GEI-intern geförderte Projekt „DiaCollo für GEI-Digital“² zielte darauf ab, den aktuellen digitalen Bestand mit etablierten computerlinguistischen Werkzeugen zu verbinden und dabei die Passfähigkeit von Daten, Werkzeugen und Bedarfen der Nutzer:innen zu testen und ggf. zu erhöhen. Es schließt damit an frühere Projekte an, bei denen mit verschiedenen Partnern unterschiedliche Ansätze für die weitere digitale Erschließung der Bestände erprobt und entwickelt wurden: so etwa zur Visualisierung der Metadaten im Projekt „GEI-Digital Visualized“³ und zur Nutzung von Volltexten und Metadaten für Filterung, Gruppenvergleiche, Suchen in Verbindung mit Topic Modells u. ä. im Projekt „Welt der Kinder“.⁴ Das Projekt ist somit Bestandteil von Bedarfserhebung und Benchmarking für Tool-Entwicklungen am GEI (De Luca et.al. 2019).

Vorgehen

Ende 2020 wurden die Daten aller Werke, die bis zu diesem Zeitpunkt mit automatisch generierten Volltexten zur Verfügung standen über die bestehenden APIs gesammelt, nach TEI konvertiert und zum *GEI-Digital-2020* Korpus zusammengefasst.

Für die maschinelle Vorverarbeitung und Indexierung wurden Werkzeuge und Workflows genutzt, die am Zentrum Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften speziell für historische deutschsprachige Texte entwickelt und genutzt werden. Diese Werkzeuge sind dafür optimiert, möglichst vorlagentreue digitale Volltexte um Zusatzinformationen anzureichern, um so z. B. Frequenz- und diachrone Kollokationsanalysen und komplexe Suchen unter Einbeziehung von Wortarten zu ermöglichen.

Hierfür wurde eine Instanz der ebenfalls am Zentrum Sprache genutzten und entwickelten D*- und DiaCollo-Software für das GEI aufgesetzt. Das open-source Werkzeug DiaCollo wurde von Bryan Jurish in Zusammenarbeit mit Historiker:innen entwickelt, um den Wortgebrauch über die Zeit sowohl im Distant Reading zu untersuchen und zu visualisieren, als auch die Ergebnisse jederzeit am konkreten Beleg in der Quelle überprüfen zu können (Jurish

2018; Jurish / Nieländer 2020). Üblicherweise wird DiaCollo mit historischen Referenzkorpora oder mehrere Jahrgänge umfassenden Zeitungs- und Zeitschriften-Korpora eingesetzt.



Abb. 1: Startseite von „DiaCollo für GEI-Digital“

Usability und Nachnutzbarkeit

- Korpus und Werkzeuge sind über die Webseite des Projektes nutzbar.
- Die Benutzeroberflächen von D* und DiaCollo wurden für Benutzer:innen mit einem gewissen Maß an Vorerfahrung mit korpuslinguistischen Methoden und Terminologie entwickelt. Um die Usability für Nutzer:innen aus anderen Disziplinen und für Laien zu erhöhen wurde ein umfangreiches Tutorial erstellt, das die Benutzeroberflächen, einige der anpassbaren Parameter sowie Beispielabfragen präsentiert. Auch Vorverarbeitung, Indexierung und einige Besonderheiten des Korpus werden im Tutorial vorgestellt (Nieländer / Jurish 2021).
- Um den Nutzer:innen intuitivere Einblicke in die Zusammensetzung des Korpus⁵ zu ermöglichen, wurden die Visualisierungen und Filterfunktionen des o.g. Projektes „GEI-Digital Visualized“ nachgenutzt, die 2017 in einer Kooperation mit der Fachhochschule Potsdam entwickelt worden waren.⁵ Zudem stehen die bibliographischen Metadaten aller Werke des GEI-Digital-2020 Korpus zum Download als Excel-Liste bereit.⁶
- Die verfügbaren Exportmöglichkeiten für Treffermengen wurden um ein KWIC/CSV Format ergänzt, um auch technisch wenig versierte Nutzer:innen in die Lage zu versetzen, diese z. B. in ein Tabellenkalkulationsprogramm zu exportieren um sie dort weiter zu bearbeiten oder archivieren zu können.
- Das *GEI-Digital-2020* Korpus wurde auch über das Zentrum Sprache zugänglich gemacht, wo es z.B. von der Community der Sprachwissenschaft und Germanistik nachgenutzt wird. Die historischen Schulbücher können dort vergleichend oder gemeinsam mit weiteren historischen Quellensammlungen der Jahre 1465–1969 untersucht werden.⁷

Befunde und Ausblick

Das Projekt verdeutlicht einmal mehr die Vorteile der Nachnutzung und der offenen, interoperablen Gestaltung von Datenbeständen und digitalen Werkzeugen. Als Anwendungsfall bereits

erprobter Abläufe war es verhältnismäßig ressourcenschonend realisierbar und half gleichzeitig, diese Abläufe weiter zu testen und optimieren. Die Möglichkeiten für digital gestützte Analysen historischer Schulbücher wurden erheblich erweitert auch wenn die Aussagekraft computerlinguistischer Analysen durch die Fehlerquote der automatischen Texterkennung einschränkt bleibt. Einige Charakteristika von Schulbüchern und der Korpuszusammensetzung haben sich als nicht optimal kompatibel mit den Logiken der Analysewerkzeuge erwiesen. Dies ist zum Teil durch die Nutzung von Filterfunktionen mit der DDC-Abfragesprache kompensierbar. Grundsätzlich zeigten sich im Projekt nutzer:innenseitige Bedarfe für die ausführlich dokumentierte und auf unterschiedliche Zielgruppen abgestimmte Gestaltung von Forschungsinfrastrukturen, die ggf. auch projektspezifische Korpuszusammenstellungen und modulare, individuell durchführbare Datenkuration erlauben.

Fußnoten

1. <http://gei-digital.gei.de/>
2. <https://diacollo.gei.de/>
3. <http://gei-digital.gei.de/visualized>
4. <http://wdk.gei.de/>
5. <https://diacollo.gei.de/gei-digital-2020/visualized/>
6. <https://diacollo.gei.de/wp-content/uploads/2021/04/gei-digital-2020.xlsx>
7. <https://www.dwds.de/d/korpora/dtaxl>

Bibliographie

Hertling, Anke / Klaes, Sebastian (2018): „Historische Schulbücher als digitales Korpus für die Forschung: Auswahl und Aufbau einer digitalen Schulbuchbibliothek“ in: Nieländer, Maret / De Luca, Ernesto William (eds.): *Digital Humanities in der internationalen Schulbuchforschung*. Göttingen: V&R unipress 21–44. DOI: 10.14220/9783737009539.21

Hertling, Anke / Klaes, Sebastian (2018): „»GEI-Digital« als Grundlage für Digital-Humanities-Projekte: Erschließung und Datenaufbereitung“ in: Nieländer, Maret / De Luca, Ernesto William (eds.): *Digital Humanities in der internationalen Schulbuchforschung*. Göttingen: V&R unipress 45–68. DOI: 10.14220/9783737009539.45

Jurish, Bryan (2018): „Diachronic Collocations, Genre, and DiaCollo“ in: Whitt, R. J. (ed.): *Diachronic Corpora, Genre, and Language Change*. Amsterdam: John Benjamins 42–64.

Jurish, Bryan / Nieländer, Maret (2020): „Using DiaCollo for historical research“ in: Simov, Kiril / Eskevich, Maria (eds.), *Selected Papers from the CLARIN Annual Conference 2019, Linköping Electronic Conference Proceedings* 172:5: 33–40. DOI: 10.3384/ecp2020172005

Nieländer, Maret / Jurish, Bryan (2021): *D* für Anfänger:innen: Ein Tutorial. Einfache und komplexe Suchanfragen, Frequenzanalysen und diachrone Kollokationsanalysen in der D*-Korpusmanagement-Umgebung*. urn:nbn:de:0220-2021-0088.

De Luca, Ernesto William / Fallucchi, Francesca / Ligi, Alessandro / Tarquini, Massimiliano (2019): „A Research Toolbox: A Complete Suite for Analysis in Digital Humanities“ in: Garoufallou E. / Fallucchi F. / William De Luca E. (eds): *Metadata and Semantic Research. MTSR 2019. Communications in Computer and Information Science*, vol 1057. Springer, Cham: 385–397. DOI: 10.1007/978-3-030-36599-8_35

Organigramme

Organigramm des Hofstaats von Königin Luise (1776–1810), Gemahlin von Friedrich Wilhelm III.

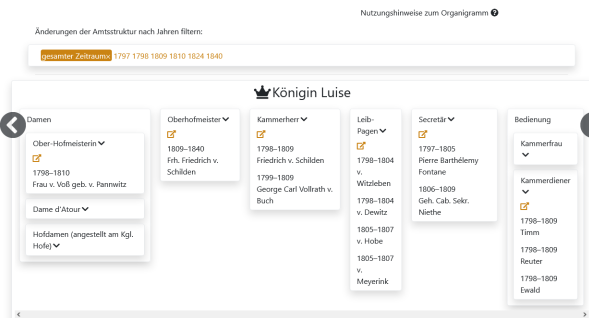


Abb. 2: Screenshot Organigramm des Hofstaats von Königin Luise (1776–1810)

Die Organigramme⁸ repräsentieren die Hofstaaten und ihre Strukturen (vgl. Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" 2021c). In Form eines ausklappbaren Strukturbaums wird die hierarchische Amtsstruktur der betrachteten Hofstaaten abgebildet. Neben den Bezeichnungen und Funktionen werden auch Informationen über die Besetzung der Ämter und Behörden, die Beziehungen zwischen diesen, sowie Veränderungen in der Amtsstruktur im Verlauf der Jahre abgebildet. Über einen Jahresfilter kann ein- und ausgeblendet werden, welche Ämter und Behörden wann hinzugefügt, entfernt, ausgesetzt oder restrukturiert wurden. Das Organigramm zum Hofstaat von Wilhelm I. wurde bereits veröffentlicht. Folgen werden die Hofstaaten vier weiterer Monarchen und ihrer jeweiligen Parallelhöfe (Ehefrauen, Kinder usw.). Im Ergebnis wird ein umfangreiches Bild der Hofstaaten, ihrer Strukturen und Verläufe gezeichnet. Die Hofstaaten werden in TEI-XML erfasst, die Verarbeitung und Darstellung erfolgt mittels XQuery und XSLT.

Adjutantenjournale

Friedrich Wilhelm IV. – Journal 1848

Abb. 3: Screenshot Adjutantenjournal Friedrich Wilhelm IV. Februar 1848

Die Journale der diensthabenden Flügeladjutanten des Monarchen, auch Adjutantenjournale⁹ genannt, bieten in Form einer eher klassischen Text-Bild-Visualisierung einen Einblick in den Tagesablauf der preußischen Monarchen für den Zeitraum von 1819–1913, da in ihnen die Termine, Orte und Treffen des Hofes festgehalten wurden (vgl. Akademienvorhaben "Anpassungs-

strategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" 2021b). Die Journale werden auf der Website sowohl als hochaufgelöste Scans der Originale sowie als edierte Texte angeboten. Sachanmerkungen und Verlinkungen zu den Registern der Projektwebsite ermöglichen ein schnelles Einlesen in das Thema und bieten durch ihre spezielle Auszeichnung eine eigene Datenbasis, die für weitere Forschungen genutzt werden kann. Die visuelle Darstellung des Textes mit der mitlaufenden Abbildung des Dokuments ermöglicht zudem einen schnellen Abgleich von Edition und Original. Die Journale werden in TEI-XML erfasst, die Verarbeitung und Darstellung erfolgt mittels XQuery, XSLT und JavaScript.

FAIR-Data als Voraussetzung für die Weiterverbreitung

Um die Weiternutzung visualisierter Daten zu ermöglichen ist es notwendig, diese nach den FAIR-Prinzipien zu publizieren. Alle in den Visualisierungen verarbeiteten Daten werden daher unter CC-BY-SA-Lizenz zum Download angeboten und über Schnittstellen verfügbar gemacht¹⁰, sodass auch die wissenschaftliche Gemeinschaft und Öffentlichkeit die Möglichkeit hat, mittels eigener Anwendungen an Kulturen der Erinnerung zur preußischen Monarchie mitzuschreiben und weiterzuforschen.

Fußnoten

1. <https://www.briefedition.alfred-escher.ch/briefe/>
2. <https://schleiermacher-digital.de/briefe/visual.xql>
3. <https://edition.onb.ac.at/sauer-seuffert/context:sauer-seuffert/methods/sdef:Context/get?mode=statistic>
4. <https://actaborussica.bbaw.de/>
5. <https://www.bbaw.de/en/bbaw-digital/telota>
6. <https://actaborussica.bbaw.de/wohntopographie/index.xql>
7. <https://leafletjs.com/>
8. <https://actaborussica.bbaw.de/organigramme/index.xql>
9. <https://actaborussica.bbaw.de/adjutantenjournale/index.xql>
10. <https://actaborussica.bbaw.de/vorhaben/index.xql?id=api>

Bibliographie

Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (2021a): "Das Akademienvorhaben Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" in: Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (eds.): *Praktiken der Monarchie*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften Version 4 vom 05.08.2021 <https://actaborussica.bbaw.de/v3/P0006476>.

Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (2021b): "Einleitung Adjutantenjournale" in: Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (ed.): *Praktiken der Monarchie*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften Version 4 vom 05.08.2021 <https://actaborussica.bbaw.de/v3/P0007683>.

Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (2021c): "Einleitung Organigramme" in: Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (ed.): *Praktiken der Monarchie*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften Version 4 vom 05.08.2021 <https://actaborussica.bbaw.de/v3/P0006306>.

Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (2021d): "Wohntopographie um 1800" in: Akademienvorhaben "Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786–1918)" (ed.): *Praktiken der Monarchie*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften Version 4 vom 05.08.2021 <https://actaborussica.bbaw.de/v3/P0006299>.

Rehbein, Malte (2017): "Informationsvisualisierung" in: Janidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): *Digital Humanities*. Eine Einführung. Stuttgart: J.B. Metzler 10.1007/978-3-476-05446-3_23 328–342.

Die Vermessung der (musikalischen) Welt

Rettinghaus, Klaus

klaus.rettinghaus@enote.com
Enote GmbH, Germany

Einleitung

Die Musik ist ein fundamentaler Bestandteil unseres kulturellen Gedächtnisses. Ihre Bewahrung und Überlieferung scheint jedoch – trotz einer schier unendlichen Zahl an verfügbaren audio-visuellen Trägermedien – nahezu ausschließlich der papiergebundenen Schriftlichkeit zuzufallen; das renommierte Riemann Musiklexikon erhob diese Schriftlichkeit sogar zum konstitutiv-konzeptionellen Bestandteil ihres Seins: „Es liegt im Wesen der abendländischen Musik, daß sie zur Schrift gebracht wird.“ (Gurlitt 1967: 641). Trotz der breit geführten Diskussion der vergangenen 50 Jahre (Möller 1997) behält die Aussage doch – gerade vor dem Idealbild der Idee einer absoluten Musik (Dahlhaus 1994) – einen wahren Kern.

Geschichte und aktuelle Lage

Die traditionelle Musiknotation hat sich über einen Zeitraum von 400 bis 500 Jahren kontinuierlich und behutsam weiterentwickelt, erst durchbrochen durch das Aufkommen der graphischen Notation im 20. Jahrhundert, die antrat, das althergebrachte Schrift-Bild durch neue Bildlichkeit zu ersetzen. (Finke 2019). Ungeachtet dessen wird bis heute die „klassische“ Musiknotation in der Schule gelehrt und ja, sie ist sogar noch immer zum Komponieren nützlich und hilfreich. Kurz gesagt: die traditionelle Notenschrift „aus Punkten und Strichen“ bildet ihren ganz eigenen Kosmos.

Umso erstaunlicher ist, dass trotz der Digitalisierungswelle der vergangenen Jahre, die insbesondere die Geisteswissenschaften erfasste, die Musiknotation im Wesentlichen in ihrer papiernen Körperlichkeit gefangen blieb. Eine *wirkliche* digitale Transformation der traditionellen Notenschrift bleibt ein Desiderat. Zwar bieten zahlreiche Sheet-Music-Apps, die über die einschlägigen Distributionsplattformen erhältlich sind, digitale Surrogate verschiedenster musikalischer Werke, im Gros handelt es sich dabei jedoch lediglich um Scans der (papiernen) Notenseiten und nicht um genuin digitalen Notensatz. Zugegebenermaßen ist eine wirkliche digitale Übersetzung des musikalischen Schriftbildes ein immenses Unterfangen, denn „eine ‚amtliche Rechtschreibung‘ existiert in der Musik nicht.“ (Weber 2015) Hinzu kommt, dass durch die langjährige Tradition der westlichen Musiknotation eine gewisse Erwartungshaltung an das Schriftbild auf Seiten der Leserschaft existiert. Und gerade Musikschaffende sind hier selten bereit Kompromisse einzugehen. Warum existiert von so vielen Werken der Musik eine Vielzahl an unterschiedlichen Ausgaben? In den seltensten Fällen, weil eine Ausgabe „richtiger“ (was auch immer das genau heißen mag) ist als eine andere. Interpretinnen und Interpreten schwören zumeist auf Ausgaben eines bestimmten Verlags, weil sie deren „Qualität“ schätzen; und das bezieht sich üblicherweise nicht auf editorische Grundsatzentscheidungen, sondern vielmehr auf den visuellen Gesamteindruck der Notation.

Seit der Industrialisierung des Notendrucks im 19. Jahrhundert herrscht die Überzeugung vor, es gäbe ein ideales Notenbild. Eine Vorstellung die nicht zuletzt durch notengrafische Großkonzerne wie C. G. Röder in Leipzig befeuert wurde, der einen Großteil der deutschen Musikverlagslandschaft belieferte (Beer 2005). Moderner computergestützter Notensatz daneben wirkt zumeist unbefriedigend. Was macht aber ein harmonisches Notenbild aus, wie sind die Größenverhältnisse im Druck? Ein Schlüssel zur Erkenntnis könnten die Notenschlüssel sein, denn diese finden sich häufig und werden im klassischen Notensatz in unterschiedlichen Größen gebraucht. So konstatiert Elaine Gould in ihrem Standardwerk zur Musiknotation: „A change of clef [...] is two-thirds of the size of the clef at the beginning of the staff.“ (Gould 2011: 7) Das klingt nach einem gesichertem Fakt. Ein flüchtiger Blick in klassische Editionen von der Wende des 19. zum 20. Jahrhundert liefert einen anderen Eindruck: dort sind Schlüsselwechsel bei einer Größe von 80–90% des Schlüssels vom Anfang des Systems deutlich größer. Auch andere Autorinnen und Autoren, sofern sie überhaupt ein bestimmtes Größenverhältnis nennen, geben üblicherweise eine Größe von 75% an (z. B. Gerou / Lusk 1996: 113). Helene Wanske schweigt diesbezüglich leider (Wanske 1988).

Ändern sich Verhältnisse bei unterschiedlichen Rastral-Größen? Sind die Relationen schlüssel-spezifisch? Wie gehen unterschiedliche Verlage vor? Hat sich der Notenstich in dieser Hinsicht über die Jahrzehnte gewandelt?

Idee und Ausblick

Um diese Fragen beantworten zu können, untersuchen wir tausende Seiten von Notendruck klassischer Musik. Moderne Editionen aus dem 20. Jahrhundert mit den Schwerpunkten Klavier- und Kammermusik haben wir selbst gescannt. Hinzu treten Ausgaben von IMSLP¹, darunter vor allem Drucke aus den „alten“ Gesamtausgaben, die bis zum 2. Weltkrieg erschienen sind und auch sinfonische Werke enthalten.

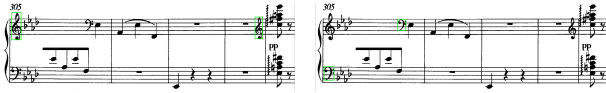


Abb. 1: Detektierte Schlüssel in unterschiedlichen Größen

Mit unserer selbstentwickelten OMR-Software sind wir in der Lage, sehr präzise Aussagen über Größen und Größenverhältnisse einzelner Symbole zu treffen. Erste Ergebnisse sollen auf einem Poster präsentiert werden. Dabei beschränken wir uns zunächst auf die Notenschlüssel, denn wie sich zeigt, bietet sich hier bereits ein weites Feld für Analysen. Diese sollen nur den Startschuss geben für weitere Untersuchungen an historischen Musikdrucken und uns helfen, bei der digitalen Transformation des klassischen Notensatzes ein ansprechenderes Layout zu erzielen. Die Ergebnisse sowie die Rohdaten sollen der Community unter einer freien Lizenz bereit gestellt werden.

Fußnoten

1. International Music Score Library Project. <https://imslp.org/>.

Bibliographie

Beer, Axel (2005): „Röder, Carl Gottlieb“, in: *MGG Online*, Laurenz Lütteken (ed.), Kassel, Stuttgart, New York 2016ff., zuerst veröffentlicht 2005, online veröffentlicht 2016, <https://www.mgg-online.com/mgg/stable/26863> [letzter Zugriff 14. Juli 2021].

Dahlhaus, Carl (1994): *Die Idee der absoluten Musik*. 3. Auflage. Kassel: Bärenreiter.

Finke, Gesa (2019): „Partituren zum Lesen und Schauen. Bildlichkeit als Merkmal graphischer Notation“, in: *Zeitschrift der Gesellschaft für Musiktheorie* 16/1, 21-39. <https://doi.org/10.31751/1001> [letzter Zugriff 14. Juli 2021].

Gerou, Tom / Lusk, Linda (1996): *Essential Dictionary of Music Notation*. Los Angeles: Alfred Publishing Co.

Gould, Elaine (2011): *Behind Bars*. London: Faber Music.

Gurlitt, Willibald / Eggebrecht, Hans Heinrich: (1967): *Riemann Musiklexikon*, Sachteil, 12. Aufl., Mainz: Schott 641.

Möller, Hartmut (1997): „Notation, Einleitung, Bewertungen von Notation“, in: *MGG Online*, Laurenz Lütteken (ed.), Kassel, Stuttgart, New York 2016ff., zuerst veröffentlicht 1997, online veröffentlicht 2016, <https://www.mgg-online.com/mgg/stable/13480> [letzter Zugriff 14. Juli 2021].

Töpel, Michael (1997): „Notation, 20. Jahrhundert, Entwicklungen seit 1950“, in: *MGG Online*, Laurenz Lütteken (ed.), Kassel, Stuttgart, New York 2016ff., zuerst veröffentlicht 1997, online veröffentlicht 2016, <https://www.mgg-online.com/mgg/stable/14135> [letzter Zugriff 14. Juli 2021].

Wanske, Helene (1988): *Musiknotation. Von der Syntax des Notenstichs zum EDV-gesteuerten Notensatz*. Mainz: Schott.

Weber, Fabian (2015): „Vom lesbaren zum schönen Partiturbild“, in: *nmz - neue musikzeitung* 64 (5) <https://www.nmz.de/artikel/vom-lesbaren-zum-schoenen-partiturbild> [letzter Zugriff 14. Juli 2021].

Digitale Texte vom Religionsfrieden bis hin zum Liebesbrief

Das Zentrum für digitale Editionen in Darmstadt stellt sich vor

Kalmer, Silke

silke.kalmer@tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Kampkaspar, Dario

dario.kampkaspar@tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Müller, Sophie

tonia-sophie.mueller@tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Seltmann, Melanie E.-H.

melanie.seltmann@tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Stegmeier, Jörn

joern.stegmeier@tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Wunsch, Kevin

kevin.wunsch@tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Einleitung

Digitale Arbeitsweisen sind heutzutage vermehrt Grundlage jeglicher editorischer Arbeit. Sie sind dabei nicht klar als abgegrenztes Werk zu sehen, sondern vielmehr als ein nie abgeschlossenes Werk (vgl. Sahle 2013: 8). Im Darmstädter Zentrum für digitale Editionen (ZEiD) an der ULB entwickeln wir verschiedene Ausgabe- und Arbeitsumgebungen, die den unterschiedlichen Rollen der Editionsutzer*innen gerecht werden: „reader, user and co-worker“ (vgl. Greve Rasmussen 2016: 126).

Das ZEiD deckt alle Aspekte der Aufbereitung von Texten für wissenschaftliche Editionen und alle Bereiche digitaler Editionen von der Planung bis zur Veröffentlichung ab. Es befasst sich mit weitreichenden Fragen der digitalen Editorik, etwa Organisation und veränderten wissenschaftlichen Praktiken von digitalen Editionen (vgl. Sahle 2013: 8). Das ZEiD bearbeitet nicht nur bibliothekseigene Bestände, sondern fungiert auch als Partner für externe Projekte wie „Europäische Religionsfrieden Digital“ und „Gruß & Kuss“ (vgl. Rapp et al. 2022). Unser Team besteht derzeit aus 8 Projektmitarbeiter*innen mit verschiedenen Zeitanteilen.

Workflow

Vorgesehen ist ein Workflow von der Texterfassung (OCR) der Digitalisate über das Erstellen und Bearbeiten der XML-Grundlage bis hin zur Realisierung der digitalen Edition als Online-Präsentation, welcher je nach Projektbedarf angepasst werden kann. Verschiedene eigens erstellte Transformationswerkzeuge dienen der Konvertierung von Texten aus verschiedenen Formaten wie etwa XML, JSON, WORD-DOCX oder PDF in ein TEI-basiertes (TEI Consortium 2021), an das DTABf angelehnte Basisformat, das spezielle Bedürfnisse des ZEiD berücksichtigt. Durch die Festlegung auf ein hauseigenes TEI-ULB-Basisformat ist die systematische Erfassung der Texte garantiert und die Einheitlichkeit der Texte aus verschiedenen Projekten in der Infrastruktur des Zentrums gegeben. Die Texte können in andere Formate konvertiert, annotiert und mit Metadaten angereichert werden. Weiterhin können Entitäten ausgezeichnet werden, die in einer zentralen Registerdatei verwaltet werden. Zudem ermöglicht das Basisformat die Ausgabe in verschiedenen Formaten, etwa JSON, DOCX, PDF und HTML.

Framework

Die XML-Dateien werden in exist-db (eXist Solutions 2021) abgelegt und mit Hilfe des Frameworks wdbplus (Kampkaspar 2018) in verschiedenen Präsentationsformen nutzerfreundlich präsentiert. Standardmäßig werden das entsprechende Digitalisat und der transkribierte Text nebeneinander dargestellt. Abweichende Darstellungsformen sind möglich, sodass auf die Besonderheiten der einzelnen Projekte eingegangen werden kann. Weitere Vorteile von wdbplus sind verschiedene APIs, mit deren Hilfe nicht nur einzelne Texte, sondern auch Metadaten einzelner Projekte abgerufen werden können. Auch Volltextsuchen können auf Projektebene sowie projektübergreifend realisiert werden.

Projekte

Das Projektportfolio des ZEiD umfasst mehrere Projekte, in denen eine Vielzahl an Textsorten mit druck- und handschriftlichen Originalen aufbereitet werden. Die Textsorten reichen dabei von frühneuzeitlichen (Hand-)Schriften über Verfassungstexte des 18. Jhdts. und Zeitungsdrucke aus drei Jahrhunderten bis hin zu handschriftlichen Liebesbriefen aus vier Jahrhunderten. Dabei adressieren wir unterschiedliche Herausforderungen. Herausragend sind hier die OCR und speziell die HTR und die Erstellung von eigens trainierten Modellen und deren Anwendung und die Qualitätskontrolle. Auch die Handhabung der heterogenen Ausgangsformate ist ein wesentlicher Bestandteil der Aufgaben.

Eines unserer bekanntesten Projekte ist „Europäische Religionsfrieden Digital“ (ULB Darmstadt 2021) in Kooperation mit der Akademie der Wissenschaften und der Literatur | Mainz und dem Leibniz-Institut für Europäische Geschichte. Dabei handelt es sich um eine rein digitale Edition von frühneuzeitlichen Religionsfriedensregelungen aus verschiedenen Regionen des heutigen Europas, bei der das ZEiD für das Konzipieren und Erstellen der digitalen Komponenten der Edition und das Entwerfen von Fragestellungen aus den Digital Humanities verantwortlich ist.

In Kooperation mit dem Institut für Sprach- und Literaturwissenschaft der TU Darmstadt das Projekt „Digitalisierung des Darmstädter Tagblatt“ durchgeführt. Das „Darmstädter Tagblatt“

erschien seit ca. 1739 über 3 Jahrhunderte hinweg in diversen Titelformen, Ausgaberrhythmen und Formaten, bis es 1986 nach 248 Jahren im „Darmstädter Echo“ aufging.

Im BMBF Citizen-Science-Projekt „Gruß & Kuss“ (Liebesbriefarchiv 2021) werden Liebesbriefe erschlossen, analysiert sowie erforscht. Durch das Verbundprojekt wird für alltagskulturelle und gefährdete Quellen, für die bisher kein staatlicher Sammlungsauftrag existiert, die dauerhafte Erforschung und Bewahrung in Gedächtnisinstitutionen erstmals sichergestellt. Hervorgegangen aus dem Liebesbriefarchiv der Universität Koblenz-Landau arbeiten im Projekt verschiedene Institute sowie die Bibliotheken der Universitäten Koblenz-Landau und Darmstadt sowie der Hochschule Darmstadt zusammen. Gemeinsam mit Bürger*innen wird hier u.a. überprüft, inwiefern der Workflow des ZEiDs auch auf heterogene handschriftliche Daten angewendet werden kann oder angepasst werden muss.

Das Projekt „Open Access Transformation by Cooperation“ (OATbyCO) wird vom BMBF gefördert. Darin entwickelt das ZEiD in Kooperation mit der Wissenschaftlichen Buchgesellschaft (wbG) anhand von 700 zum Teil nur noch gedruckt vorliegenden Titeln aus der Backlist der wbG einen XML-basierten Workflow zur digitalen Veröffentlichung, welcher als Modell für gleichartige kooperative Unternehmungen in diesem Bereich dienen soll. Mit dem Aufbau einer digitalen Infrastruktur zur Indexierung, Langzeitarchivierung und Dissemination werden die Titel anschließend im Open Access zur Verfügung gestellt werden.

Bibliographie

eXist Solutions (2021): eXist DB. <http://exist-db.org> [letzter Zugriff 29. November 2021].

Greve Rasmussen, Krista Stinne (2016): “Reading or Using a Digital Edition? Reader Roles in Scholarly Editions” in: Driscoll, Matthew James / Pierazzo, Elena (eds.): *Digital scholarly editing. Theories and practices*. UK: Open Book Publisher 119–136.

Kampkaspar, Dario (2018): “W. Digitale Bibliothek (wdbplus)”, in: *GitHub* <https://github.com/dariok/wdbplus> [letzter Zugriff 29. November 2021].

Liebesbriefarchiv (2021): “Projekt ‘Gruß & Kuss’”, in: *Liebesbriefarchiv* <https://liebesbriefarchiv.de> [letzter Zugriff 29. November 2021].

Rapp, Andrea / Büdenbender, Stefan / Dietz, Nadine / Dunkelmann, Lena / Gnau-Franké, Birte / Liesenfeld, Nina / Schmunk, Stefan / Selmann, Melanie E.-H. / Stäcker, Thomas / Werner, Stephanie / Wyss, Eva L. (2022): “Mein liebster Schatz! Das Citizen Science-Projekt Gruß & Kuss stellt sich vor”, in *DHd2022. Kulturen des digitalen Gedächtnisses*: Zenodo.

Sahle, Patrick (2013): *Digitale Editionsformen*. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels (= Schriften des Instituts für Dokumentologie und Editorik 7). Norderstedt: BoD.

TEI Consortium (2021): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [Version 4.3.0]. [31.08.2021]. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/> [letzter Zugriff 29. November 2021].

ULB Darmstadt (2021): “Europäische Religionsfrieden digital”, in *Universitäts- und Landesbibliothek Darmstadt* https://www.ulb.tu-darmstadt.de/forschen_publizieren/forschen/eured.de.jsp [letzter Zugriff 29. November 2021]

Digitalisierte Ego-Dokumente als Quellen für die historische Forschung

Adamczak, Katarzyna

katarzyna.adamczak@bsb-muenchen.de
Bayerische Staatsbibliothek, Germany

Štanzel, Arnošt

arnost.stanzel@bsb-muenchen.de
Bayerische Staatsbibliothek, Germany

Ego-Dokumente als Quellen für die historische Forschung bieten ein breit ausgefächertes und äußerst spannendes Untersuchungsfeld. Zum einen ermöglichen sie Einblicke auf die verschiedenen Ebenen der Lebens- und Gefühlswelt der Autorin/des Autors, zum anderen verläuft in ihnen eine von Forscherinnen und Forschern zu definierende Grenze zwischen dem, was vom Ego „absichtlich oder unabsichtlich enthüllt“¹ oder aber verborgen wurde.

Neben den (auto)biographisch gefärbten Zügen besitzen Ego-Dokumente weitere Merkmale, die sie für die Forschung sowie aus informationstechnologischer und bibliothekarischer Sicht interessant machen.

Zunächst umfassen sie Textsorten, die freiwillig oder unfreiwillig entstanden. Zur erstgenannten Kategorie zählen Texte wie Autobiographien, Memoiren, Reiseberichte, Tagebücher oder Briefe, die u.U. für eine Veröffentlichung bestimmt waren. Der zweiten Kategorie gehören Texte an, bei denen an eine explizite längerfristige Überlieferung oder Veröffentlichung nicht oder nur unter bestimmten Voraussetzungen gedacht war, z.B. Schriftstücke aus administrativen Kontexten: Testamente, Verhörprotokolle oder Zeugenbefragungen. Sodann geben Ego-Dokumente Auskünfte über sonst in historischen Schriftquellen wenig vertretene Gruppen, wie etwa Frauen, Bauern, Arbeiter, Handwerker oder Soldaten. Damit helfen sie, Erfahrungszusammenhänge und Lebenswelten der Unter- und Mittelschichten zu rekonstruieren, die in autobiographischen Texten von herausragenden historischen Persönlichkeiten kaum oder ungenügend beleuchtet wurden. Schließlich bedienen Ego-Dokumente verschiedene Medien: neben den klassischen Manuskripten und Akten sind sowohl Bilder – wie etwa Lithografien oder Fotografien –, als auch Ton- und Filmaufnahmen zu berücksichtigen.

Die Bayerische Staatsbibliothek ist im Besitz tausender gedruckter Tagebücher, Autobiographien und Memoiren, aber auch Fotografien und Filmen. Neuerdings publiziert sie im Rahmen eines Projektes des Fachinformationsdienstes Ost-, Ostmittel- und Südosteuropa bislang unveröffentlichte Selbstzeugnisse digital. Der Publikationsdienst steht wissenschaftlichen Institutionen und auch Privatpersonen offen und umfasst Materialien mit Bezug zum östlichen und südöstlichen Europa, die Quellencharakter haben und in Deutschland vorliegen, aber durch kommerzielle Verlage nicht veröffentlicht werden. In diesem Zusammenhang wurden bisher ausgewählte Ego-Dokumente aus dem Archiv der Forschungsstelle Osteuropa an der Universität Bremen sowie Ego-Dokumente aus dem Nachlass des Osteuropahistorikers Martin Winkler (1893-1982), die der Bayerischen Bibliothek vermacht wurden, digitalisiert und über das Forschungsportal *os-*

mikon im Open Access bereitgestellt.² Derzeit wird die Bereitstellung des aus Privatbesitz stammenden Nachlasses der deutsch-russischen Medizinerin Elsa Winokurov (1883-1983) sowie von einigen Selbstzeugnissen aus dem Bestand des Instituts für deutsche Kultur und Geschichte Südosteuropas an der LMU München vorbereitet. Die Resonanz auf die bisher veröffentlichten Materialien ist in der Fachcommunity groß. So gab es hoch interessierte Rückmeldungen zum Fotoalbum von Helmuth Schröder über dessen Kriegsgefangenschaft in Sibirien³, und der Nachlass Winokurov dient einem Projektkurs des Elitestudiengangs Osteuropastudien an der LMU im WS 2020/2021 als Arbeitsgrundlage.⁴

Die physikalische Vielfalt der digitalisierten Materialien (handschriftlich verfasste Manuskripte, born digital, Einzelblätter, Fotos, Fotonegative, Filme, Audioaufnahmen) ging einher mit der Erarbeitung von technischen Workflows, bei denen größtenteils auf bestehende Best Practices, die im Kontext anderer Projekte an der Bayerischen Staatsbibliothek entwickelt wurden, zurückgegriffen werden konnte. Als besonders herausfordernd erwiesen sich jedoch die Digitalisierung von Dias sowie das Verfahren zur Angabe von Wasserzeichen und Bildunterschriften – beides musste in mehreren Schritten per Trial and Error optimiert werden.

Das Poster fokussiert zwei Aspekte der digitalen Bereitstellung von Ego-Dokumenten: Erstens die digitale Transformation unveröffentlichter Selbstzeugnisse, beginnend mit deren Anwerbung und Auswahl über die Klärung von Rechtsfragen bis hin zu Digitalisierung, Katalogisierung, Langzeitarchivierung und Online-Bereitstellung; zweitens die Verwendung von digitalisierten Ego-Dokumenten in Forschung und Lehre und der daraus resultierenden Kooperationsmöglichkeiten zwischen Privatpersonen, Forschungsinstituten, Universitäten und Bibliotheken.

Es handelt sich somit um den Sachstand und um Ausblicke eines Projektes, das seit 2019 an der Bayerischen Staatsbibliothek betrieben wird und zugleich um wissenschaftliche Zugänge zu spannenden historischen Quellen im Kontext der Mikrohistorie, Migrationsgeschichte und teilweise auch der *herstory*.

Fußnoten

1. J. Presser „Uit het werk van dr J. Presser“, Amsterdam 1969, S. 286, zitiert nach W. Schulze „Ego-Dokumente: Annäherung an den Menschen in der Geschichte? Vorüberlegungen für die Tagung »EGO-DOKUMENTE«“, in: Ders. „Ego-Dokumente. Annäherung an den Menschen in der Geschichte“, Berlin u.a. 1996, S. 11-30; hier: S. 15.
2. Siehe <https://www.osmikon.de/publizieren/ego-dokumente-veroeffentlichen>.
3. Siehe <https://ego-dokumente.osmikon.de/BV045328295/>.
4. Dr. Kornelia Konczal/Dr. Arpine Maniero: Projektkurs „Elsa Winokurov (1883-1983): eine deutsch-russische Biographie“.

Bibliographie

Schulze, Winfried (1996): "Ego-Dokumente: Annäherung an den Menschen in der Geschichte? Vorüberlegungen für die Tagung 'EGO-DOKUMENTE'", in: Ders. (Hg.): *Ego-Dokumente. Annäherung an den Menschen in der Geschichte*, Berlin u.a.: Akademie Verlag 11-30.

Doing (Digital) History

Kollaborative Formen der Erforschung von Geschichte in sozialen Medien im Projekt #SocialMediaHistory

Berg, Mia

mia.berg@rub.de
Ruhr-Universität Bochum, Deutschland

Lorenz, Andrea

andrea.sarah.lorenz@uni-hamburg.de
Universität Hamburg, Deutschland

Kontext

Medien speichern, vermitteln und strukturieren Gedächtnis¹ und Geschichte. Sie können sowohl Ergebnis von Erinnerungsprozessen sein als auch einen Erinnerungsanlass bieten (Zierold 2006: 136f.). Erst durch und in unterschiedlichsten Gedächtnismedien² wird das kollektive Gedächtnis konstruiert und durch Akte kollektiver Erinnerung hervorgebracht (Erl 2004: 3). Das Internet hat dabei nicht nur zur digitalen Transformation bestehender Inhalte geführt, sondern neue digitale Gedächtnisinhalte und Erinnerungs- und Geschichtspraktiken generiert. Die Trennung des kollektiven Gedächtnisses in ein kulturelles „Langzeitgedächtnis“ und ein kommunikatives „Kurzzeitgedächtnis“ weicht zunehmend auf (Assmann 2002: 246). Das zeigen insbesondere die sozialen Medien, deren parallele „Gedächtniscommunities“ und „networked memories“ sich durch Pluralisierung und Fragmentierung individueller und kollektiver Erinnerungen auszeichnen (Bartoletti 2011: 100). Die niedrigen Produktions- und Zugangs-schranken potenzieren geschichtsbezogene Inhalte, Akteur*innen und Praktiken. Für Historiker*innen werden Geschichtsbilder und diskursive Aushandlungen sichtbar. Es entsteht eine diverse, partizipative Erinnerungslandschaft, die die Deutungs- und Diskurs-hoheit etablierter Akteur*innen wie Institutionen oder Forscher*innen in Frage stellt (König 2020: 76). An die Stelle gesellschaftlicher und institutioneller Relevanzstrukturen treten die Interessen der User*innen. Soziale Medien verweisen so auf die Kontingenz und Selektivität von Erinnerungsprozessen (Zierold 2006: 188). Das digitale Gedächtnis zeigt dabei im Brennglas, was aus historischer Perspektive schon immer zu hinterfragen war: das Verhältnis von Erinnern und Vergessen, Fragen der Überlieferung, Auswahl und Speicherung, des Originals, der Partizipation und Sichtbarkeit, der Zugänglichkeit und nicht zuletzt gesellschaftlicher Macht.

Forschungsstand

Hannes Burkhardt hat jüngst festgehalten, dass die „Relevanz des Internets für die Vergegenwärtigung von Vergangenheit heute kaum überschätzt werden“ könne (Burkhardt 2021: 13). Die Ge-

schichtswissenschaft arbeitet zwar seit Jahren digital und hat mit der Digital History eine eigene „digitale“ Disziplin herausgebildet. Vor allem das in audiovisuellen sozialen Medien geformte digitale Gedächtnis wurde bisher jedoch nicht umfangreicher untersucht.³ Die Kopplung an Datenstrukturen und globale Konzerne führt zu technischen, ethischen und rechtlichen Herausforderungen. Zentral ist vor allem die Frage, inwiefern Geschichte als Big Data überhaupt ausgewertet werden kann, wenn Plattformen wie Instagram nur eingeschränkte APIs zur Verfügung stellen oder automatisierte Datenerhebungen vollständig verbieten.⁴ Dabei sind gerade Instagram und TikTok auch in Bezug auf geschichtsbezogene Inhalte besonders Nutzer*innen- und Reichweitenstark: Allein #history wurde auf Instagram 41 Millionen Mal geteilt und auf TikTok 23 Milliarden Mal aufgerufen (Stand: November 2021). Um sich den neuen Formaten historischer Erzählung anzunähern, müssen sich geschichtswissenschaftliche Werkzeuge und Infrastrukturen ändern. Zentral stellen sich Fragen nach der Zugänglichkeit, Archivierbarkeit und (automatisierten) Auswertbarkeit der entstehenden Daten (König 2020, Kiechle 2018). Da im Fach häufig nicht einmal flächendeckend empirische Methoden curricular verankert sind, sind viele Historiker*innen im Umgang mit der „Computer Mediated History“ (Kiechle 2021) auf die Digital Humanities angewiesen. Bereits im Zuge der DHd 2020 wurde ein fachwissenschaftlicher Bedarf an niedrigschwelligen Services ausgemacht, der auch auf den Umgang und die Forschung mit Social Media-Daten übertragbar ist (Hermes / Klinke / Demmer 2020: 184f.).

Projekt

Das Projekt „SocialMediaHistory“ erforscht seit März 2021 zusammen mit Citizen Scientist, wie Geschichte auf Instagram und TikTok stattfindet, analysiert und produziert werden kann.⁵ Die ursprünglich rein geschichtswissenschaftliche Perspektive muss dabei um eine informatische ergänzt werden, um die Datenmengen des digitalen Gedächtnisses erschließen und handeln zu können. Darüber hinaus müssen die technischen und kommerziellen Bedingtheiten der Inhalte reflektiert und selbst Datenbankarbeit und Sammlungsaufbau geleistet werden. In den nächsten drei Jahren sollen Geschichtswissenschaft und Digital Humanities deshalb stärker verzahnt werden.

Konkret geschieht dies in einem ersten Schritt im Zuge eines Projektseminars am IDH der Universität zu Köln, das von Jürgen Hermes im WiSe 21/22 angeboten wird. Ausgehend von im Projekt formulierten Fragestellungen entwickeln Studierende in interdisziplinärer Zusammenarbeit Lösungsansätze für (1) die automatisierte Erhebung und Auswertung von Instagram- und TikTok-Kommentaren (Scraping, Sentiment Analysis), (2) die Identifizierung und Auswertung vergangenheitsbezogener Hate Speech (Topic Modeling) sowie (3) die Darstellung von Begleitdiskursen auf Twitter (Netzwerkanalyse). Gemeinsam sollen Möglichkeiten und Grenzen digitaler Methoden ausgelotet werden – auf technischer und rechtlicher Ebene und auf Ebene der historischen Erkenntnis.

Ausblick

Im weiteren Projektverlauf sollen die erarbeiteten Tools angewandt und evaluiert werden. Das Projekt verfolgt damit eine kollaborative Doing (Digital) History im doppelten Sinne: Durch die eigene Produktion von geschichtsbezogenem Content und als

methodischer Zugang zu und Reflexion über Beschaffenheit und Auswertbarkeit geschichtsbezogener Social Media-Inhalte.

Das Poster möchte zentrale Fragestellungen und gewonnene Erkenntnisse vorstellen. Auf diese Weise möchte das Projekt einen Beitrag dazu leisten, neue Perspektiven auf eine gemeinsame Wissen(schaft)skultur zu liefern und neue Zugänge in der Geschichtswissenschaft zu etablieren.

Fußnoten

1. Gedächtnis soll hier in Anlehnung an Martin Zierold (2006) als Fundus und Archiv verstanden werden.
2. Der Begriff verweist Aleida Assmann (2009) folgend auf die Rolle von Medien als externer Speicher und Träger des kulturellen Gedächtnisses.
3. Zum Forschungsstand siehe Burkhardt (2021).
4. Vgl. <https://www.instagram.com/about/legal/terms/before-january-19-2013/>
5. Weitere Informationen: www.socialmediahistory.de.

Bibliographie

Assmann, Jan (2002): "Das kulturelle Gedächtnis", in: *Erwägen, Wissen, Ethik* [EWE] 13: 239-247.

Assmann, Aleida (2009): *Erinnerungsräume. Formen und Wandlungen des kulturellen Gedächtnisses* (4. Aufl.). München: C.H.Beck.

Balbi, Gabriele / Ribeiro, Nelson / Schafer, Valérie / Schwarzenegger, Christian (eds.) (2021): *Digital Roots. Historicizing Media and Communication Concepts of the Digital Age* (Studies in Digital History and Hermeneutics Vol. 4). Berlin / Boston: De Gruyter.

Bartoletti, Roberta (2011): "Memory and Social Media: New Forms of Remembering and Forgetting", in: Pirani, Bianca Maria (eds.): *Learning from Memory: Body, Memory and Technology in a Globalizing World*. Cambridge: Scholars Publishing 82-111.

Blaney, Jonathan / Miligan, Sarah / Steer, Mary / Winters, Jane (2021): *Doing digital history. A beginner's guide to working with text as data*. Manchester: Manchester University Press.

Bunnenberg, Christian / Logge, Thorsten / Steffen, Nils (2021): "SocialMediaHistory. Geschichtemachen in sozialen Medien", in: *Historische Anthropologie* 29-2: 267-283.

Burkhardt, Hannes (2021): *Geschichte in den Social Media. Nationalsozialismus und Holocaust in Erinnerungskulturen auf Facebook, Twitter, Pinterest und Instagram* (Beihefte zur Zeitschrift für Geschichtsdidaktik, Band 23). Göttingen: V&R unipress.

Erll, Astrid (2004): "Medium des kollektiven Gedächtnisses – ein (erinnerungs-)kulturwissenschaftlicher Kompaktbegriff", in: Erll, Astrid / Nünning, Ansgar (eds.): *Medien des kollektiven Gedächtnisses. Konstruktivität – Historizität – Kulturspezifität*. Berlin / New York: De Gruyter 3-22.

Föhr, Pascal (2019): *Historische Quellenkritik im Digitalen Zeitalter*. Glückstadt: vvh Verlag Werner Hülsbusch.

Hermes, Jürgen / Klinke, Harald / Demmer, Dennis (2020): "Public Humanities Tools: Der Bedarf an niederschweligen Services", in: *Book of Abstracts Dhd 2020: Spielräume* 184-186. <https://zenodo.org/record/3666690#.YO8NJZgzaUk> [letzter Zugriff 24. November 2021].

Kiechle, Oliver (2018): "Archivierung von Social Media? Store local!", in: Blog. *Diskrete Werte. Digitalia – Quellenkritik –*

Public History. <https://digitalia.hypotheses.org/56> [letzter Zugriff 24. November 2021].

Kiechle, Oliver (2021): " 'One person's Data is another person's noise.' – Flame Wars, SPAM und Bots in Born Digital Sources", in: Blog. *Digitale Geschichtswissenschaft*. <https://digitalhist.hypotheses.org/2389> [letzter Zugriff 24. November 2021].

König, Mareike (2017): "Digitale Methoden in der Geschichtswissenschaft. Definitionen, Anwendungen, Herausforderungen", in: *BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen* 1-2: 7-21.

König, Mareike (2020): "Geschichte digital. Zehn Herausforderungen", in: Arendes, Cord et al. (eds.): *Geschichtswissenschaft im 21. Jahrhundert. Interventionen zu aktuellen Debatten*, Berlin / Boston: De Gruyter 67-76.

Möller, Katrin (2021): "Die Modellierung des zeitlichen Vergleichs als Kernkompetenz von Digital History", in: Blog. *Digitale Geschichtswissenschaft*. <https://digitalhist.hypotheses.org/2399> [letzter Zugriff 24. November 2021].

Russel, Matthew A. / Klassen, Mikhail (2019): *Mining the Social Web. Data Mining Facebook, Twitter, LinkedIn, Instagram, Github, and More* (3. Aufl.). Sebastopol: O'Reilly.

Vlassenroot, Eveline / Chambers, Sally / Lieber, Sven / Michel, Alejandra / Geeraert, Friedel / Pranger, Jessica / Birkholz, Julie / Mechant, Peter (2021): "Web-archiving and social media: an exploratory analysis", in: *International Journal of Digital Humanities* 1-3. <https://doi.org/10.1007/s42803-021-00036-1> [letzter Zugriff 24. November 2021].

Zierold, Martin (2006): *Gesellschaftliche Erinnerung. Eine medienkulturwissenschaftliche Perspektive* (Media and Cultural Memory / Medien und kulturelle Erinnerung, Band 5). Berlin / New York: De Gruyter.

Ein Thesaurus für die digitale Edition der Ästhetikvorlesungen von Friedrich Schleiermacher

Kelm, Holden

kelm@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Klappenbach, Lou

klappenbach@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Einleitung

Die Ästhetikvorlesungen des Philosophen und Theologen Friedrich Schleiermacher¹ (1768-1834) sind Bestandteil des ästhetischen Diskurses der klassischen deutschen Philosophie nach Kant (vgl. Jaeschke / Arndt 2012), sind aber als Bestandteil dieses Diskurses noch nicht hinreichend untersucht worden. Das sich die-

sem Desiderat widmende DFG-Projekt², das mit dem Akademienvorhaben „Schleiermacher in Berlin 1808-1834“³ assoziiert ist, stellt die Frage, wie die Ästhetikvorlesungen aufgrund von digitalen Methoden erschlossen, wie sie in ihrem Verlauf untersucht und in ihrem philosophiehistorischen Kontext kritisch dargestellt werden können. Das geplante Poster illustriert eine für dieses Projekt zentrale digitale Methode: das von der Digital Humanities-Arbeitsgruppe TELOTA⁴ entwickelte Tool ediarum.SKOS.

Erfassung und Analyse der Themen und Begriffe der Ästhetikvorlesungen mit dem Tool ediarum.SKOS

Das DFG-Projekt schließt an die Hybridedition von Schleiermachers Ästhetikvorlesungen im Rahmen der *Kritischen Gesamtausgabe* Schleiermachers (Schleiermacher 2021) an und ist durch drei zentrale Forschungsfragen gegliedert: 1.) Wie kann die Nutzung und Erschließbarkeit der digitalen Edition durch thematische und begriffliche Zugänge sinnvoll erweitert werden? 2.) Wie kann der Verlauf der drei Ästhetikvorlesungen, die Schleiermacher 1819, 1825 und 1832/33 an der Berliner Universität hielt, aufgrund der überlieferten Dokumente näher bestimmt und in Bezug auf mögliche Veränderungen der Konzeption untersucht werden? 3.) Wie können die Ästhetikvorlesungen in ihrer Eigenbedeutung und in ihrem diskursiven Umfeld mit digitalen Mitteln erörtert werden?

Zur Bearbeitung dieser Forschungsfragen wurde beschlossen, die Annotation, Definition und Referenzierung von Themen mit einem digitalen Thesaurus vorzunehmen. Die Modellierung der Themen- und Begriffsfelder basiert auf dem „Simple Knowledge Organisation System“⁵ (SKOS), das auf dem Resource Description Framework (RDF) und RDF-Schema aufbaut (Isaac/Summers 2009), und welches in XML serialisiert werden kann. Das Akademienvorhaben „Schleiermacher in Berlin 1808-1834“ erschließt verschiedene Quellen für digitale und für Druck-Editionen mit *ediarum*⁶, einem oXygen-Framework für den Author-Modus des oXygen XML-Editors.⁷ Da semantische Modellierungsformen wie Thesauri bisher nicht in *ediarum* integriert sind, wird im Rahmen des Ästhetik-Projekts ein eigenes Framework entwickelt: ediarum.SKOS. Es bietet eine nutzer:innenfreundliche Oberfläche, um SKOS-basierte Thesauri zu erstellen und zu bearbeiten.



Abb. 1: Symbolleiste des ediarum.SKOS Frameworks im Autormodus des oXygen XML Editors

Das Framework⁸ kann als oXygen Add-on⁹ installiert werden und umfasst derzeit verschiedene Grundfunktionen, wie die Erstellung von Konzepten (skos:Concept), von Sammlungen (skos:Collection), die Vergabe von Labels (skos:prefLabel, skos:altLabel) und die Einfügung von Definitionen (skos:definition). An Beziehungstypen zwischen Konzepten sind derzeit Hierarchien (skos:narrower und skos:broader) und Assoziationen (skos:related) implementiert. Ebenfalls Teil von ediarum.SKOS ist ein SKOS-Reasoner in XSLT, der für die Nutzer:innen „per Knopfdruck“ Inferenzen aus den Beziehungen zieht. Zur Verknüpfung von Konzepten des Thesaurus mit historischen Dokumenten innerhalb einer Edition, kann ediarum.SKOS mit weiteren *ediarum*-Modulen (wie ediarum.BASE) sowie mit projektspezifischen *ediarum*-Frameworks kombiniert werden.

```
<skos:Concept rdf:about="https://schleiermacher-digital.de/thesaurus/aesthetik/ed_sdw_siv_r4b">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:prefLabel xml:lang="de">Geschichte der Ästhetik als Wissenschaft</skos:prefLabel>
  <skos:inSchema rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/krk_mhq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/ekz_3xq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/pjt_3xq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/egs_lyq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/fqw_lyq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/dct_vyq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/pfv_vyq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/kwz_1zq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/hkz_mzq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/hcz_xzq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/rz2_zzq_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/r42_d1z_cqb"/>
  <skos:narrower rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik/imf_fir_cqb"/>
  <skos:topConceptOf rdf:resource="https://schleiermacher-digital.de/thesaurus/aesthetik"/>
</skos:Concept>
```

Abb. 2: Datenmodellierung am Beispiel des Konzeptes "Geschichte der Ästhetik als Wissenschaft"

Methodisch wurde so vorgegangen, dass in den Textdokumenten der Ästhetikvorlesungen aufgrund von Lektüre und Interpretation eine Gruppe von 14 Themen identifiziert und jedem Thema ein vorläufiges Begriffsfeld zugewiesen wurde. Mithilfe des Tools ediarum.SKOS konnten die identifizierten Themen in Form eines Thesaurus organisiert werden. Für jedes Thema wurde ein Eintrag angelegt und dieser mit weiteren Informationen wie Bezeichnungen, Definitionen und Beziehungen zu anderen Einträgen ausgestattet. Die Themen wurden dann in den TEI-XML-Dateien annotiert und auf den entsprechenden Eintrag im Thesaurus referenziert. Für die Auswahl und Überprüfung der themenspezifischen Begriffe wurde „voyant tools“¹⁰ hinzugezogen (terms, trends, keywords in context); die sich daraus ergebenden signifikanten Begriffe wurden dann in den Thesaurus integriert und ihre Beziehung zu den Themen bestimmt. Aufgrund dieser Modellierung konnten semantische Beziehungen zwischen Themen und Begriffen in den Ästhetikvorlesungen explizit formalisiert und dadurch in computerlesbarer Weise erfasst werden (Rehbein 2017: 163). Dies ermöglicht, die verschiedenen Themen der Ästhetik Schleiermachers abzufragen, zu durchsuchen, zu analysieren und über das World Wide Web mit kontrollierten Vokabularen anderer Projekte und mit anderen Thesauri zu verknüpfen (Harpring 2010; Zaytseva / Đurčo 2020).

Ausblick

Die Erfassung der Themen und Begriffe mit ediarum.SKOS bildet die Grundlage für die Einbettung und Visualisierung des Thesaurus im Rahmen der digitalen Editionsplattform *schleiermacher digital*¹¹. Damit soll die Edition der Ästhetikvorlesungen um eine Erschließungsoption erweitert werden, die den Zugang zum Korpus durch einschlägige Themen und/oder durch mit diesen Themen verknüpfte signifikante Sachbegriffe ermöglicht. Da-

durch wird es künftig auch möglich, thematische und begriffliche Verknüpfungen mit anderen Ästhetikvorlesungen der Epoche herzustellen, um den ästhetischen Diskurs in autorenübergreifender Perspektive rekonstruieren zu können.

Dafür soll ediarum.SKOS über die derzeit implementierten Grundfunktionen hinaus weiterentwickelt und über das Ästhetik-Projekt hinaus nutzbar gemacht werden. Künftig sollen weitere Funktionen zur Verknüpfung von Konzepten mit externen Ressourcen im Sinne von linked open data hinzukommen. Dem Open Science Prinzip folgend, werden Code und Dokumentation im Projektverlauf zur Nachnutzung und Weiterentwicklung unter GNU-Lizenz auf Github publiziert und über Zenodo versioniert.

Fußnoten

1. https://de.wikipedia.org/wiki/Friedrich_Schleiermacher [letzter Zugriff 15. Juli 2021].
2. Zum DFG-Projekt „Schleiermachers Ästhetikvorlesungen im Kontext. Zur Reflexion und Anwendung digitaler Methoden in der Konstellationsforschung“ von Holden Kelm vgl. URL: <https://gepris.dfg.de/gepris/projekt/448730446?context=projekt&task=showDetail&id=448730446> [letzter Zugriff 15. Juli 2021].
3. <https://www.bbaw.de/forschung/schleiermacher-in-berlin-1808-1834-briefwechsel-tageskalender-vorlesungen> [letzter Zugriff 15. Juli 2021].
4. <https://www.bbaw.de/bbaw-digital/telota> [letzter Zugriff 15. Juli 2021].
5. <https://www.w3.org/TR/skos-primer/> [letzter Zugriff 15. Juli 2021].
6. <https://www.ediarum.org> [letzter Zugriff 15. Juli 2021].
7. <http://www.oxygenxml.com> [letzter Zugriff 15. Juli 2021].
8. Das Framework kann über die URL <http://telota.bbaw.de/ediarum/skos/edit/update.xml> installiert werden.
9. Informationen zum Installieren von oXygen Add-ons finden sich in der Dokumentation unter <https://www.oxygenxml.com/doc/versions/23.1/ug-editor/topics/howto-install-plugins.html> [letzter Zugriff 15. Juli 2021].
10. <https://voyant-tools.org/> [letzter Zugriff 15. Juli 2021].
11. <https://schleiermacher-digital.de/>

Bibliographie

Akademienvorhaben Schleiermacher in Berlin 1808–1834 (o.J.): *schleiermacher digital, Briefwechsel, Tageskalender, Vorlesungen von Friedrich Schleiermacher 1808–1834. Eine Publikation des Akademienvorhabens Schleiermacher in Berlin 1808–1834 der Berlin-Brandenburgischen Akademie der Wissenschaften*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften <https://schleiermacher-digital.de/> [letzter Zugriff 15. Juli 2021].

Rehbein, Malte (2017): "Ontologien" in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): *Digital Humanities. Eine Einführung*, Stuttgart: J.B. Metzler 328-342 10.1007/978-3-476-05446-3_23.

Schleiermacher, Friedrich Daniel Ernst (2021): "Vorlesungen über die Ästhetik" in: Holden Kelm (ed.): *Kritische Gesamtausgabe Schleiermachers, Abt. II, Bd. 14*, unter Verwendung vorbereitender Materialien von Wolfgang Virmond, Berlin: De Gruyter 10.1515/9783110537758.

Dumont, Stefan / Fechner, Martin (2014/2015): "Bridging the Gap: Greater Usability for TEI encoding", in: *Journal of the Text Encoding Initiative* [Online] 8 <http://journals.openedition.org/jtei/1242> [letzter Zugriff 15. Juli 2021].

Harpring, Patricia / Baca, Murtha (2013): *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. Los Angeles, California: Getty Research Institute https://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/index.html [letzter Zugriff 15. Juli 2021].

Isaac, Antoine / Summers, Ed (2009): *W3C Working Group Note 18 August 2009. SKOS Simple Knowledge Organization System Primer* <https://www.w3.org/TR/skos-primer/> [letzter Zugriff 15. Juli 2021].

Jaeschke, Walter / Arndt, Andreas (2012): *Die Klassische Deutsche Philosophie nach Kant. Systeme der reinen Vernunft und ihre Kritik 1785-1845*. München: C.H. Beck.

Zaytseva, Ksenia / Ďurčo, Matej (2020): "Controlled Vocabularies and SKOS" Version 1.1.0. Edited by Matej Ďurčo and Tanja Wissik. DARIAH-Campus [Training module] <https://campus.dariah.eu/resource/controlled-vocabularies-and-skos> [letzter Zugriff 15. Juli 2021].

Entdeckung der Korrespondenz Alexander von Humboldts durch Such- und Visualisierungsfunktionen

Lecroq, Axelle

axelle.lecroq@bbaw.de
BBAW, Germany

Einleitung

Dieses experimentelle Projekt¹ zielt darauf ab, die Korrespondenz von Alexander von Humboldt zu entdecken, zu erforschen und zu visualisieren. Ausgangspunkt der Idee ist das Findbuch Alexander von Humboldts Bibliothek, der sich in der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) befindet. Bislang ist die Sammlung nur über ein in den 1950er Jahren angelegtes Karteikartensystem zugänglich. Lediglich das Findbuch wurde von Anne McKinney im Rahmen ihres Praktikums für das Projekt "Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung" digital reproduziert.

Die ursprüngliche Idee war, zumindest einen Teil der Sammlung digital zugänglich zu machen und sie mit Hilfe neuer Recherche-tools zu entdecken. Mit diesem Ziel vor Augen sollte das Findbuch mit modernen Handschriftendatenbanken korreliert werden.

Geschichte

Das Projekt der Rekonstruktion der Korrespondenz Alexander von Humboldts begann mit der Gründung einer Alexander-von-Humboldt-Kommission bei der Deutschen Akademie der Wissenschaften zu Berlin im Jahr 1956. 1960 richteten die Akademien der Wissenschaften in Ost- und Westdeutschland sowie die Österreichische Akademie der Wissenschaften ein gemeinsames Ersuchen um internationale Unterstützung einer Edition von Humboldts Briefwechsel an Akademien, Archive, Bibliotheken und Sammler. Bereits zwei Jahre später waren Kopien von rund 7.600 Briefen aus aller Welt in Berlin eingetroffen (Schuchardt 2010, S.50-56).

Auf der Grundlage dieser stetig wachsenden Sammlung nahm 1970 die Alexander-von-Humboldt-Forschungsstelle ihre Arbeit auf. Von 1973 bis 2014 veröffentlichte ein Team von Forschern 42 Bände mit Korrespondenzen, Monografien und Sammelbänden. Seit 2015 setzt ein neues Projekt "Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung" diese langjährige Arbeit im Sinne der Digital Humanities fort und bereitet eine digitale und gedruckte Edition von Alexander von Humboldts Reisetagebüchern und wissenschaftlichen Manuskripten aus seinem Nachlass vor.

Die Daten

Der Kalliope-Verbundkatalog (<https://kalliope-verbund.info/>) ist sicherlich der größte Katalog von Archiven teilweise deutschsprachiger Institutionen. Die Daten der von Alexander von Humboldt (AvH) gesendeten und empfangenen Briefe wurden von der API (SRU Server) des Kalliope-Verbundes im Dublin Core-Format abgerufen.

Damit dieses digitale Projekt repräsentativ für die jahrelange Arbeit der BBAW ist, war es wichtig, auch Daten aus dem Ausland einzubeziehen. Man einigte sich zunächst auf die Daten der Bibliothèque nationale de France (BnF, <https://data.bnf.fr/>), die eine benutzerfreundliche API anbietet. Daraufhin wurden die in der Bibliothèque nationale de France aufbewahrten und über deren Online-Katalog zugänglichen Daten der Korrespondenz Alexander von Humboldts im csv-Format abgerufen.

Die in der Suchhilfe der BBAW aufgelisteten Institutionen sind bei weitem nicht nur europäisch, sondern es wurden auch die bei der American Philosophical Society (<https://diglib.amphilsoc.org/>) aufbewahrten Dokumente über Alexander von Humboldt im EAD-Format abgerufen.

Quantität

Auf der Grundlage von Berechnungen aus dem Jahr 1962 wurde geschätzt, dass Humboldt bis zu 3.000 Briefe pro Jahr schrieb. Hochgerechnet ergab dies eine geschätzte Gesamtzahl von 35.000 bis 50.000 Briefen aus Humboldts Hand; diese Schätzung ist auch heute noch gültig (Biermann und Lange 1962, S. 226). Bei den von Humboldt erhaltenen Briefen, die heute größtenteils als verschollen gelten, gingen Biermann und Lange von etwa 100.000 Briefen aus.

Year	Analogue collection			Digital collection	
	Letters by avH	Letters to AvH	Other	Letters by AvH	Letters to AvH
1962	7000	600			
1965	8.800	1.400			
1974	10.500	2.700			
2001	12.000	3.000			
2021	8.690	2.215	2.175	3465*	1467*

Abb.1: Anzahl der Dokumente in der Humboldt-Sammlung (* Anzahl der Einträge. Das bedeutet, dass ein Eintrag für mehr als einen Buchstaben stehen kann. Dies ist häufig bei Daten aus der BnF der Fall, wo ein Eintrag für mehrere Dutzend Dokumente stehen kann.)

Bearbeitung der Daten

Jeder Datensatz wurde bereinigt und homogenisiert. Die vier unterschiedlich formatierten Datensätze – CSV der BnF, XML-EAD der APS, Dublin Core des Kalliope-Verbundkatalogs und die Excel-Tabelle des Findbuchs – wurden in einem JSON-Format zusammengeführt. Diese JSON-Datei bildet dann unsere durchsuchbare Datenbank.

Um die Briefe auf einer Karte zu visualisieren, wurden außerdem neue Daten für jeden Brief hinzugefügt:

- Geopoint, Geoname ID und humboldt digital edition identifier (edh) ID für den Ort der Institution
- Geopoint, Geoname ID und edh ID für den Ort der Briefe

Verwendete Tools

Da es nicht möglich war, eine vollständige Website zu entwickeln, die eine umfassende Benutzererfahrung ermöglicht, wurde beschlossen, zunächst Jupyter-Notebooks zu verwenden. Diese interaktiven und leistungsstarken Notebooks haben den Vorteil, dass sie zahlreiche Widget-Möglichkeiten und Datenvisualisierungen bieten. Für die Datenvisualisierung wurden mehrere Bibliotheken verwendet, unter denen die wichtigsten sind:

- ipywidgets (<https://ipywidgets.readthedocs.io/en/latest/>) bietet zahlreiche interaktive HTML-Widgets innerhalb der Zell-Outputs eines Jupyter Notebooks.
- ipyleaflet (<https://ipyleaflet.readthedocs.io/en/latest/>) ermöglicht die Erstellung von interaktiven Karten innerhalb von Jupyter Notebooks.
- pandas (<https://pandas.pydata.org/>) bietet Werkzeuge zur Analyse und Manipulation von Daten.

Sie eignet sich besonders für Datenstrukturen.

Die Visualisierungen und Suchfunktionen

Die Suchfunktionen durchsuchen verschiedene Elemente: Absender, Empfänger, Empfangs- oder Versandort und Archivzentrum. Ein Dropdown-Menü ermöglicht es dem Benutzer, den gewünschten Wert auszuwählen.

Optimale Zoomstufe

Der Algorithmus berechnet die optimale Zoomstufe, d.h. die niedrigste Zoomstufe bei der alle Punkte auf der angezeigt werden können.

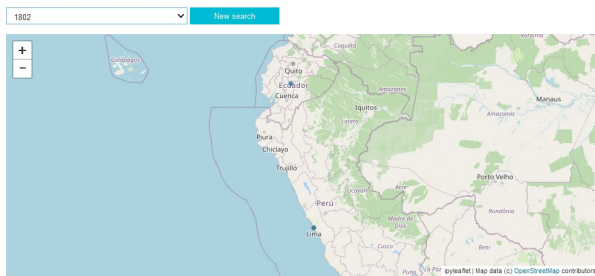


Abb. 2: Kartenvisualisierung für die Datumssuche '1802 mit optimale Zoomstufe'.

Zugang zur Informationen

Alle Karten sind interaktiv. Um zusätzliche Informationen zu erhalten, kann der Nutzer auf einen Punkt klicken, wodurch ein PopUp-Menü erscheint.

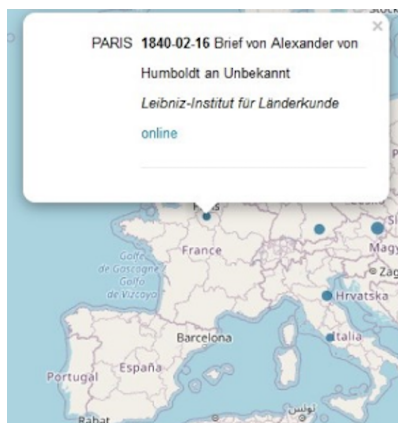


Abb. 3: Beispiel von einem PopUp Menu in einer Visualisierung

Dynamische Suchfunktion

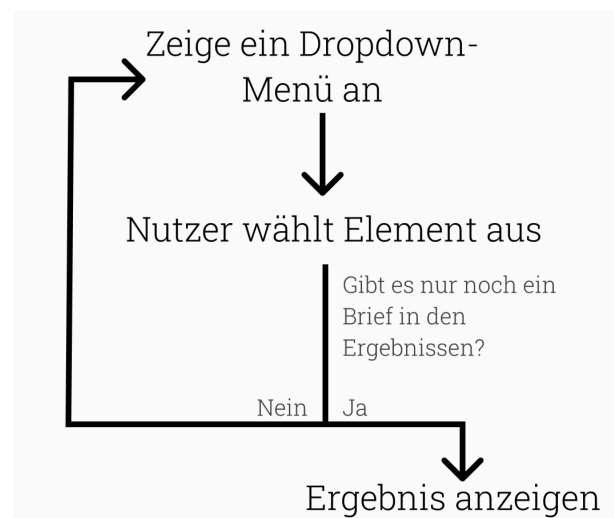


Abb. 4: Logik der dynamischen Suchfunktion

Abb. 5: Darstellung der dynamischen Funktion in Jupyter Notebook: Eine Folge von Dropdown-Menüs, bis es keine weiteren suchbaren Elemente mehr gibt oder nur noch ein Ergebnis (einen Brief) verfügbar ist

Darstellung aller Daten auf einer Karte

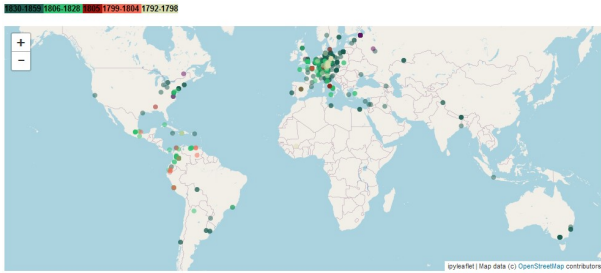


Abb. 6: Erste Variante der Darstellung aller Daten: Jede Farbe stellt einen Lebensabschnitt Humboldts dar. Die Legende ist noch nicht interaktiv. Es wäre eventuell interessant, wenn der Nutzer auf die verschiedenen Lebensabschnitte von Alexander von Humboldt klicken könnte, um zu sehen, wie sich seine Korrespondenz im Laufe der Zeit und in den verschiedenen Lebensabschnitten verändert hat.

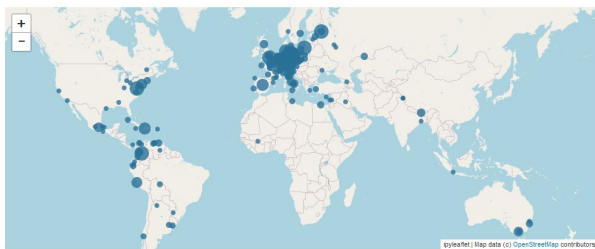


Abb. 7: Zweite Variante der Darstellung aller Daten: Der Radius der Punkte ist proportional zur Anzahl der empfangenen und gesendeten Briefe des Ortes. Diese Karte macht deutlich, dass Alexander von Humboldts Korrespondenz alle Kontinente berührt

Fußnoten

1. Mehr über das Projekt auf Github: <https://github.com/edition-humboldt-collection/corresp-humboldt-dataviz>

Bibliographie

Biermann, Kurt-R. und Fritz G. Lange (1962): „Die Alexander-von-Humboldt-Briefausgabe“. *Forschungen und Fortschritte* 36, 8: S. 225–230.

Biermann, Kurt-R. (1965): „Der Zugang an Briefen Alexander von Humboldts hält an“. *Spektrum. Mitteilungsblatt für die Mitarbeiter der Deutschen Akademie der Wissenschaften zu Berlin* 11, 2: S. 55–58.

Biermann, Kurt-R. (1974): „Die Alexander-von-Humboldt-Forschung an der Akademie der Wissenschaften der D.D.R. – Ergebnisse und Ziele“. *Boston Studies in the Philosophy of Science* 15: S. 295–305.

ERNiE's bibliographical details (Encyclopedia of Romantic Nationalism in Europe, ed. Joep Leerssen; Amsterdam: Study Platform on Interlocking Nationalisms, 2015) : <https://ernie.uva.nl/> (<https://ernie.uva.nl/>).

Hotson, H., & Wallnig, T. (Eds.). (2019): *Reassembling the Republic of Letters in the Digital Age*. doi:10.17875/gup2019-1146

Humboldt, Alexander von (1973). *Die Jugendbriefe Alexander von Humboldt 1787–1799*. Hg. und erläutert von Ilse Jahn und Fritz G. Lange. Berlin: Akademie-Verlag.

Lecroq, Axelle (2021), Entre enrichissement et développement de projets : l'utilisation de données externes pour la correspon-

dance d'Alexander von Humboldt, sous la direction d'Ariane Pinche, École nationale des chartes (PSL).

Mapping the Republic of Letters, Stanford University, 2013. <http://republicofletters.stanford.edu/>,

Schuchardt, Gregor (2010): *Fakt, Ideologie, System. Die Geschichte der ostdeutschen Alexander von Humboldt-Forschung*. Stuttgart: Franz Steiner Verlag.

Schwarz, Ingo (2001): „Zur Geschichte der Alexander-von-Humboldt-Forschung an der Berlin-Brandenburgischen Akademie der Wissenschaften“. In: *Die Berliner und Brandenburger Lateinamerikaforschung in Geschichte und Gegenwart. Personen und Institutionen*. Hg. von Gregor Wolff, S. 107–127. Berlin: Wissenschaftlicher Verlag Berlin.

FDM-Awareness in Zeiten von Corona Sammelkarten zum Forschungsdatenmanagement „Daten & Datteln digital“

Mollenhauer, Elisabeth

e.mollenhauer@uni-koeln.de

Universität zu Köln, Germany

Rau, Felix

f.rau@uni-koeln.de

Universität zu Köln, Germany

FDM-Vermittlung vor Corona

Als Forschungsdatenkompetenzzentrum der Philosophischen Fakultät der Universität zu Köln ist das Data Center for the Humanities (DCH) für die aktive Unterstützung von Geisteswissenschaftler:innen an der Fakultät und darüber hinaus bei Fragen zum Forschungsdatenmanagement (FDM) zuständig. Dies umfasst alle Aspekte des Forschungsdatenlebenszyklus, von der initialen Forschungsfrage über die Datenerhebung und -analyse bis hin zur Archivierung, Publikation und Nachnutzbarmachung von Forschungsdaten. Dies schließt auch Aspekte der Öffentlichkeitsarbeit mit ein, um Wissenschaftler:innen über Fragen des Umgangs mit Forschungsdaten zu informieren und die Awareness für die Notwendigkeit von umfassendem Forschungsdatenmanagement zu steigern.¹

In einer Umfrage zu Forschungsdaten an der Philosophischen Fakultät 2018 schätzte die überwiegende Mehrheit der befragten Wissenschaftler:innen ihre FDM-Kenntnisse als durchschnittlich (45,2 %) bis gering (29,6 %) ein. Als Konsequenz wurde die Beratungstätigkeit des DCH sowie die aktive Unterstützung von Wissenschaftler:innen beim Datenmanagement weiter verstärkt und Schwellen reduziert. Mit einer 2019 eingerichteten offenen, wöchentlich stattfindenden Sprechstunde konnten die durch das DCH durchgeführten Beratungen noch einmal signifikant erhöht werden. Mit dem FDM-Workshop für Promovierende an der a.r.t.e.s. Graduate School for the Humanities Cologne und

der FDM-Übung für Masterstudierende der Informationsverarbeitung, Medieninformatik und Linguistik haben FDM-Aspekte außerdem Eingang in die universitäre Lehre gefunden und werden so frühzeitig an Nachwuchswissenschaftler:innen herangetragen (vgl. Blumtritt et al. 2020a, 2020b).²

Für das Sommersemester 2020 wurde in Zusammenarbeit mit dem Dekanat der Philosophischen Fakultät eine mehrteilige Vortragsreihe zu FDM-Themen geplant, vorbereitet und beworben (s. Abb. 1). Die als praktisch orientierte „How-to“-Reihe konzipierten Vorträge sollten unter dem Titel „Daten & Datteln“ Informationsvermittlung mit Kaffee, Keksen und Trockenobst in einer ungezwungenen Atmosphäre verbinden.



Abb. 1: Ankündigung der Vorträge im Sommersemester 2020 (Layout und Illustration: Julia Sorouri)

Digitale FDM-Vermittlung während der Corona-Pandemie

Als Mitte März 2020 deutlich wurde, dass neben Lehre und Beratung auch eine Vortragsreihe in Präsenz nicht möglich sein würde, wurde entschieden, die Vorträge durch digitale Sammelkarten zu ersetzen und über den digitalen Newsletter der Fakultät monatlich zu versenden.³ Die Sammelkarten wurden gleichzeitig auch über den Twitter-Account des DCH kommuniziert und fanden so auch über die Grenzen der Fakultät hinaus Anklang.⁴ Ergänzt wurden die Sammelkarten im Wintersemester 2020/21 durch zwei digitale Kurzvorträge für Mitglieder der Fakultät.⁵

Die monatliche Kommunikation der Sammelkarten konnte so der durch Kontaktbeschränkungen reduzierten Sichtbar- und Niederschwelligkeit des DCH und seines Angebots entgegenwirken.⁶ Gleichzeitig wurde regelmäßig über den Fortbestand des Beratungs- und Betreuungsangebot des DCH im digitalen Raum informiert. Aufgrund der positiven Resonanz auf Twitter – und der andauernden Pandemie – wurde die Reihe in den darauffolgenden Semestern durch weitere Sammelkarten erweitert und ist seitdem fortlaufend.

Entwicklung und Umsetzung des Sammelkartensets

Für das Sommersemester 2020 (und auch für die folgenden) wurden vorab vier Sammelkarten mit je einem FDM-Thema konzipiert und als kompakter Text in einem entworfenen Template umgesetzt. Zentrale Aspekte der Sammelkarten wurden mit Icons und Diagrammen veranschaulicht. Als Download wurden die Karten öffentlich auf der Webseite des DCH mit einem vertiefendem Text zur Verfügung gestellt und sind dort mit einer individuellen URL aufrufbar.⁷

Die ersten acht Sammelkarten bieten einen Einstieg in grundlegende FDM-Aspekte und somit auch in die Grundlagen guter wissenschaftlicher Praxis (vgl. DFG 2019) und Themen der Data Literacy (vgl. Wuttke & Helling 2020). Sie betreffen sämtliche Etappen des Forschungsdatenlebenszyklus und nehmen konkret Bezug auf das Service-Portfolio des DCH einerseits sowie durch das Datenzentrum adressierte Forschungsbereiche andererseits:⁸

- Forschungsdatenmanagement und Drittmittelförderung
- Datenmanagementpläne
- Archivierung
- Digitales Publizieren
- Langlebigkeit und Pflege von Daten
- Backup und Datensicherheit
- Nachhaltige Softwarekuratierung
- Persistente Identifier: DOI, ORCID und Co.

Dabei hat sich gezeigt, dass bereits auf Vermittlung ausgerichtete Konzepte wie die FAIR-Prinzipien (vgl. Wilkinson et al. 2016)⁹ oder die 3-2-1-Backup-Regel (vgl. Briney et al. 2020: 8–9) aufgrund ihrer Prägnanz sich besonders für das Format der Sammelkarte eignen (s. Abb. 2). Fortgeführt wurde die Reihe mit der Vorstellung eigener Ressourcen und Services, wodurch insbesondere auch fachspezifische FDM-Aspekte und die Kompetenzbereiche des DCH – audiovisuelle (AV) Daten und lexikalische Ressourcen – abgedeckt wurden.

HOW TO Reihe zum Forschungsdatenmanagement in den Geisteswissenschaften

6

Datenverlust kann katastrophale Konsequenzen für die wissenschaftliche Arbeit haben. Ein durchdachter Backup-Workflow kann das verhindern und ist daher elementarer Bestandteil guten Forschungsdatenmanagements.

 x 3   x 2  x 1

Grundlage ist die 3-2-1-Regel: Daten sollten immer

- dreimal vorliegen (Original und 2 Kopien),
- auf zwei Medientypen gespeichert und
- einmal als Offsite-Kopie gesichert werden.

Die Sicherung sollte dabei regelmäßig, automatisiert und möglichst über institutionelle Infrastruktur geschehen.

Kontaktieren Sie uns, wenn Sie Beratung oder Unterstützung zu Backup und Sicherheit Ihrer Forschungsdaten suchen, gerne via E-Mail.

info-dch@uni-koeln.de
@dch_cologne
<https://dch.phil-fak.uni-koeln.de>

DCH
Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

Universität zu Köln
In Zusammenarbeit mit dem Institut
der Philosophischen Fakultät

Daten & Datteln

 **HOW TO KEEP IT SAFE**
Backup und Datensicherheit

Abb. 2: Sammelkarte 6 – Backup und Datensicherheit (Layout: Elisabeth Mollenhauer, basierend auf Arbeiten von Julia Sorouri)

Ausblick zur Nachnutzung

Die zunächst nur temporär geplanten Sammelkarten haben sich mittlerweile als wesentlicher Bestandteil der Öffentlichkeitsarbeit des DCH auf Twitter etabliert. In postpandemischer Zukunft sollen diese als gedrucktes Set in kleiner Auflage in Beratung, Schulung und Lehre als Gesprächseinstieg und praktische Merkhilfen zum Mitnehmen eingesetzt werden. Die über Twitter kommunizierten Sammelkarten zu generischen FDM-Themen sollen weiterhin auch universitätsextern einen Mehrwert für die geisteswissenschaftliche Community erzeugen.

Fußnoten

1. https://dch.phil-fak.uni-koeln.de/sites/dch/user_upload/Satzung_DCH__11.07.2018_.pdf [letzter Zugriff 1. Dezember 2021].
2. <https://dch.phil-fak.uni-koeln.de/sichtbarkeit-und-lehre/umfrage-forschungsdaten>; <https://dch.phil-fak.uni-koeln.de/sichtbarkeit-und-lehre> [letzter Zugriff 1. Dezember 2021].
3. Die Idee, FDM-Inhalte auf ein Sammel- oder Spielkartenformat zu konzentrieren und zu vermitteln, wurde bereits im „DANS Data Game“ (<https://web.archive.org/web/20211201115745/https://dans.knaw.nl/en/news/dans-data-game/>; <https://web.archive.org/web/20211201120148/https://dans.knaw.nl/en/news/dans-data-game-also-online-avail->

able/) des Data Archiving and Networked Services (DANS) in Den Haag und in den „Research Data ScaryTales“ des Thüringer Kompetenznetzwerks Forschungsdatenmanagement (TKDM) umgesetzt (vgl. Gerlach et al. 2020). Auch das Dramenquartett (<https://dramenquartett.github.io/>) verfolgt ein didaktisches Ziel (vgl. Fischer et al. 2018a, 2018b) [letzter Zugriff 1. Dezember 2021].

4. https://twitter.com/dch_cologne [letzter Zugriff 1. Dezember 2021].

5. <https://dch.phil-fak.uni-koeln.de/daten-und-datteln/vortraege> [letzter Zugriff 1. Dezember 2021].

6. Siehe hierzu auch die „Research Data ScaryTales“ des TKDM: <https://web.archive.org/web/20211024045932/https://forschungsdaten-thueringen.de/rdm-scarytales/articles/overview.html> [letzter Zugriff 1. Dezember 2021].

7. <https://dch.phil-fak.uni-koeln.de/daten-und-datteln/sammelkarten> [letzter Zugriff 1. Dezember 2021].

8. <https://dch.phil-fak.uni-koeln.de/fdm-services>; <https://dch.phil-fak.uni-koeln.de/forschung> [letzter Zugriff 1. Dezember 2021].

9. <https://www.go-fair.org/fair-principles/> [letzter Zugriff 1. Dezember 2021].

Bibliographie

Blumtritt, Jonathan / Helling, Patrick / Mathiak, Brigitte / Neufeind, Claes / Rau, Felix / Schildkamp, Philip / Wieners, Jan Gerrit (2020a): "Geisteswissenschaftliches Forschungsdatenmanagement in der Lehre – Konzepte, Methoden, Erfahrungen", in: *DHd 2020 "Spielräume. Digital Humanities zwischen Modellierung und Interpretation". 7. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. Konferenzabstracts* 318–321 <https://doi.org/10.5281/zenodo.4608524>.

Blumtritt, Jonathan / Helling, Patrick / Mathiak, Brigitte / Neufeind, Claes / Rau, Felix / Schildkamp, Philip / Wieners, Jan Gerrit (2020b): "Geisteswissenschaftliches Forschungsdatenmanagement in der Lehre – Konzepte, Methoden, Erfahrungen", in: *DHd 2020 "Spielräume. Digital Humanities zwischen Modellierung und Interpretation". 7. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. [Poster]* <https://doi.org/10.5281/zenodo.3697519>.

Briney, Kristin A. / Coates, Heather / Goben, Abigail (2020): "Foundational Practices of Research Data Management", in: *Research Ideas and Outcomes* 6: e56508 <https://doi.org/10.3897/ri-o.6.e56508>.

Deutsche Forschungsgemeinschaft (DFG) (2019): *Guidelines for Safeguarding Good Research Practice. Code of Conduct* <https://doi.org/10.5281/zenodo.3923601>.

Fischer, Frank / Kittel, Christopher / Milling, Carsten / Trilcke, Peer / Wolf, Jana (2018a): "Dramenquartett – Eine didaktische Intervention", in: *DHd 2018 "Kritik der digitalen Vernunft". 5. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. Konferenzabstracts* 397–398 <https://doi.org/10.5281/zenodo.4622596>.

Fischer, Frank / Kittel, Christopher / Milling, Carsten / Schultz, Anika / Trilcke, Peer / Wolf, Jana (2018b): "Dramenquartett – Eine didaktische Intervention", in: *DHd 2018 "Kritik der digitalen Vernunft". 5. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. [Poster]* <https://doi.org/10.6084/m9.figshare.5926363.v1>.

Gerlach, Roman / Rex, Jessica / Lang, Kevin / Neute, Nadine / Schröter, Annett / Schwartz, Volker

(2020): *Research Data ScaryTales* [Data set] <https://doi.org/10.5281/zenodo.4066679>.

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, IJsbrand Jan, et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship", in: *Sci Data* 3: 160018 <https://doi.org/10.1038/sdata.2016.18>.

Wuttke, Ulrike / Helling, Patrick (2020): "Barcamp Data Literacy: Datenkompetenzen in den digitalen Geisteswissenschaften vermitteln: Erkenntnisse aus dem Workshop der AG Datenzentren während der DHd 2020 an der Universität Paderborn", in: *Bausteine Forschungsdatenmanagement* 2 (November): 49–55 <https://doi.org/10.17192/bfdm.2020.2.8276>.

Fidus Writer als Alternative zum DH ConValidator? Ein Prototyp

Gebhard, Henning

s2hegebh@uni-trier.de
Universität Trier, Germany

Hintergrund: DH ConValidator

Kollaborative Autorschaft, digitales Publizieren und Langzeitarchivierung sind Themen, die in den Digital Humanities viel diskutiert werden (Ernst 2015; DHd-AG Digitales Publizieren 2016; Neuroth et al. 2010). XML-TEI hat sich als für diese Zwecke angemessenes Dateiformat etabliert und wird unter Anderem für den Einreichungsprozess der Abstracts für die Jahrestagungen des DHd-Verbands und die internationale Digital Humanities-Konferenz, aber u.a. auch beim *Journal of the Text Encoding Initiative* verwendet (vgl. Branden und Holmes 2014; Blanke et al. 2014). XML-TEI ist zwar ein offenes, sehr expressives und damit für "semantisches Publizieren" (Shotton 2009, Schöch 2020) und Langzeitarchivierung geeignetes Format, das zudem in den Digital Humanities weit verbreitet ist. Es ist jedoch auch vergleichsweise unhandlich und für viele daher nicht das Format der Wahl beim Verfassen ihrer Texte. Der aktuelle Lösungsansatz – die Nutzung eines DH ConValidator genannten Online-Tools zur Umwandlung von Office-Dokumenten in XML-TEI – ist fehleranfällig und unbequem sowohl für die Einreichenden als auch für diejenigen, die das Book of Abstracts erstellen.¹

Alternativer Ansatz: Fidus Writer mit Erweiterungen

Vor diesem Hintergrund wurde im Rahmen einer Masterarbeit der Prototyp einer möglichen Alternative zum Einsatz des DH ConValidators entwickelt (Gebhard 2021). Diese Alternative besteht in der Verwendung eines webbasierten WYSIWYG Editors für das nahtlose Verfassen und die Einreichung des Abstracts. Im Rahmen der Masterarbeit wurde ein solcher Prototyp auf Grundlage von "Fidus Writer" entwickelt, einem Open Source Editor mit starkem Fokus auf wissenschaftliches Schreiben.² Die-

ser ist in Python/Django implementiert und kann als Web-basierter Editor zur Verfügung gestellt werden. Er erlaubt nicht nur das kollaborative Verfassen von Texten inklusive Abbildungen und Formelsatz, sondern auch die Eingabe relevanter Metadaten und die Verwaltung der Literaturangaben. User:innen können unterschiedliche Rollen mit angepassten Berechtigungen zugeteilt werden, um den Review-Prozess zu unterstützen. Die gewünschte Dokumentstruktur kann über Templates auf die Bedürfnisse von Konferenzen zugeschnitten werden. Wie eine ausführliche Anforderungsanalyse (u.a. mit leitfadengestützten Interviews mit verschiedenen Stakeholdern) gezeigt hat, fehlt für die Eignung zur Benutzung im Rahmen von DH-Konferenzen allerdings u.a. die Anbindung an TEI-basierte Workflows. Daher wurde ein Exporter neu entwickelt, mit dem Fidus Writer Dokumente im TEI Format gespeichert werden können, entweder für Archivzwecke oder für die weitere Verwendung in üblichen Publikationspipelines.

Der Schreib- und Einreichungsprozess mit Fidus Writer

Statt des bisherigen Einreichungsverfahrens mit den Word-Processor-Templates und dem DH ConValidator könnte bei Verwendung des erweiterten Fidus Writer der Schreib- und Einreichungsprozess folgendermaßen ablaufen: Die Konferenz-Organisator:innen setzen eine FidusWriter-Instanz auf und hinterlegen ein geeignetes Dokument-Template. Die Autor:innen legen dort einen Account und ein neues Dokument an. Sie schreiben online gemeinsam ihren Text und binden die Literatur über Zotero und/oder die Literaturverwaltung von Fidus Writer ein. Wenn der Text fertig ist, exportieren die Autor:innen ihren Beitrag (als ZIP-Datei mit XML-TEI, der BibTex-Datei und den Abbildungen). Diese Datei reichen Sie in ConFTool ein. Die Konferenz-Organisator:innen erhalten XML-TEI, das nach einem recht strikten Schema valide ist (siehe unten) und können daraus HTML und PDF generieren. Die XML-TEI-Datei kann archiviert werden.

Ein Beitrag zur Standardisierung: Fidus Writer und jTEI-Schema

Ein angenehmer Nebeneffekt dieser Arbeit war die Notwendigkeit, ein striktes, weitgehend auf jTEI beruhendes Schema für das Exportformat zu definieren. Dieses erlaubt nicht nur die Überprüfung des Exporters selbst. Es könnte auch in Schritt in die Richtung sein, dass der gesamte Einreichungsprozess unabhängiger von einzelnen Tools und im Idealfall völlig plattformagnostisch wird. Überdies stellt dieses Schema ein verlässliches Development Target für weitere Programme dar, die für die Publikation oder sonstige Nachnutzung entwickelt werden und unterstützt damit deren Wiederverwendbarkeit (vgl. Schreibman 2009).

Fazit und Ausblick

Mit dem Pilotprojekt zum Fidus Writer soll nicht nur eine Alternative zum Einreichungsprozess mit dem DH ConValidator entwickelt werden. Der Ansatz hat auch zum Ziel, einen nutzungsfreundlichen und zeitgemäßen Schreib- und Einreichungsprozess mit einem hohen Anspruch an die Datenqualität und die Nutzung offener Formate und Standards für die Digital Humanities-Com-

munity zu verbinden. Nächste Schritte in diesem Prozess sollen nun der testweise Einsatz bei einer kleineren Konferenz sein, bevor der Einsatz auch bei der Jahrestagung des DHd-Verbands angedacht wird.

Posterpräsentation

Für das Poster ist geplant, die für eine Einreichung bei der DHd-Konferenz wesentlichen Schritte und ihre Unterstützung durch den im Rahmen der Masterarbeit erweiterten Fidus Writer zu illustrieren. Sofern vor Ort möglich, könnte der Fidus Writer auch an einem Laptop im Einsatz demonstriert und durch die Konferenzteilnehmer:innen getestet werden.

Fußnoten

1. Der DH ConValidator wurde ursprünglich im Kontext der DH2012 in Hamburg von Marco Petris entwickelt und wird seitdem bei der internationalen DH-Konferenz und der DHd-Jahrestagung eingesetzt. Er interagiert mit ConfTool und verwendet OxGarage für die Formatttransformation. Siehe: <https://github.com/ADHO/dhconvalidator>.
2. Fidus Writer wird maßgeblich von Johannes Wilm entwickelt. Siehe: <https://www.fiduswriter.org/>.

Bibliographie

- Blanke, Tobias / Pierazzo, Elena / Stokes, Peter A.** (2014): „Digital Publishing Seen from the Digital Humanities“. *Logos* 25 (2): 16–27. <https://doi.org/10.1163/1878-4712-11112041>.
- Branden, Ron van den / Holmes, Martin** (2014): „Journal of the Text-Encoding Initiative Article Schema. Schema and Guidelines for Encoding an Article for the Journal.“ *Journal of the Text-Encoding Initiative*. https://tei-c.org/release/doc/tei-p5-exemplars/pdf/tei_jtei.doc.pdf.
- DHd-Arbeitsgruppe Digitales Publizieren** (2016): "Working Paper Digitales Publizieren". *DHd Working Papers*, 1. Wolfenbüttel: HAB / DHd-Verband. <http://dhd-wp.hab.de/?q=ag-text>.
- Ernst, Thomas** (2015): "Vom Urheber zur Crowd, vom Werk zur Version, vom Schutz zur Öffnung? Kollaboratives Schreiben und Bewerten in den Digital Humanities". In: *Grenzen und Möglichkeiten der Digital Humanities*. Hg. von Constanze Baum / Thomas Stäcker. 2015 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). DOI: 10.17175/sb001_021.
- Gebhard, Henning** (2021): *Building a TEI-driven online tool for authoring, submitting and publishing of conference abstracts based on Fidus Writer*. Trier: Universität Trier.
- Kaden, Ben / Kleineberg, Michael** (2017): "Zur Situation des digitalen geisteswissenschaftlichen Publizierens – Erfahrungen aus dem DFG-Projekt „Future Publications in den Humanities.“" *Bibliothek Forschung und Praxis* 41, no. 1. <https://doi.org/10.1515/bfp-2017-0009>.
- Neuroth, Heike / Oßwald, A. / Scheffel, R. / Strathmann, S. / Huth, Karsten** Hrsg. (2010): *nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Göttingen: nestor.
- Schöch, Christof** (2021): "Open Access für die Maschinen." In *Die Zukunft des kunsthistorischen Publizierens*, ed. Maria Effinger, Hubertus Kohle. Heidelberg: ART-Books. DOI: <https://doi.org/10.11588/arthistoricum.663.c9210>.

Shotton, David (2009): „Semantic Publishing: The Coming Revolution in Scientific Journal Publishing“. *Learned Publishing* 22 (2): 85–94. <https://doi.org/10.1087/20092020>.

Schreibman, Susan (2009): "The Text Encoding Initiative: An Interchange Format Once Again." In *Jahrbuch Für Computerphilologie* 10, 12–24. Mentis Verlag, 2009.

GitMA-Poster CATMA-Daten via Git abrufen und mittels Python-Bibliothek weiterverarbeiten

Meister, Malte

meister@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Vauth, Michael

vauth@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Gerstorfer, Dominik

gerstorfer@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Etwas zu erinnern heißt nicht, es abzuspeichern, sondern auch, es abzurufen und weiter zu prozessieren. Denn nur im produktiven Anschluss erhält die Erinnerung eine Bedeutung. Diese Beobachtung trifft *a fortiori* auf technische Speichersysteme zu. Der Nutzen einer Software wird, gerade in den Digital Humanities, über die Möglichkeiten bestimmt, die erzeugten Daten zu exportieren, zu konvertieren, zu archivieren und in anderen Systemen weiterzuverarbeiten. In diesem Poster werden wir neue Möglichkeiten vorstellen, Daten aus CATMA (Gius et al. 2021) abzurufen und nachzunutzen.

CATMA (Computer Assisted Text Markup and Analysis) ist eine kollaborative Textannotations- und Analyse-Plattform, die in den Digital Humanities gut etabliert ist und von vielen Projekten aktiv genutzt wird. Annotationsexporte waren, besonders im XML-TEI Format, schon seit Version 3 ein wichtiger Bestandteil der CATMA-Software (Petris & Meister 2016; Petris 2017). Der Datenzugriff war aber bis einschließlich CATMA 5 nur über die graphische Benutzeroberfläche (GUI) möglich. Seit der Version 6.0 werden die von den Nutzer:innen erzeugten Daten in einem auf Git basierenden Backend gespeichert und versioniert.

Der genaue Aufbau der Datenstrukturen wird auf der CATMA Webseite dokumentiert (Petris 2020): Jedes Dokument, jede Annotation Collection einschließlich der Annotationen, sowie jedes Tagset einschließlich der zugehörigen Tags werden im Backend einzeln repräsentiert. Besonders wichtig für die Weiterverarbeitung der Annotationsdaten sind die Informationen, mit denen die einzelnen Annotationen repräsentiert werden:

- eine Referenz auf das entsprechende Dokument
- die genaue Platzierung der annotierten Textspanne (als sogenannte Start und End-Offsets, welche sich auf die Zeichen-Positionen im Dokument beziehen)

- eine Referenz auf das verwendete Tag (die Annotationskategorie) und das Tagset (eine benannte Kollektion von Tags) aus dem es stammt
- eventuell Properties (vordefinierte erweiternde Eigenschaften) und deren Werte
- Autor:in der Annotation
- Zeitpunkt der Annotation

Die Nutzer:innen können sowohl auf eigene als auch auf mit ihnen geteilte Daten in Form von Git Repositorien zugreifen. Diese stellen damit eine Art Programmierschnittstelle (API) zum Abruf von CATMA-Annotationen dar, welche auf den lokalen Rechner heruntergeladen oder in anderen Tools weiterverarbeitet werden können.

Im Fachbereich für Digital Philology an der TU Darmstadt ist außerdem eine Python-Bibliothek entstanden, die einen einfachen Zugriff auf die Git Repositorien zulässt. Sie ermöglicht die Weiterverarbeitung der Annotationen mit gängigen Python Datascience-Tools, zum Beispiel als Pandas DataFrame. Mit der Python-Bibliothek lassen sich unter anderem Berechnungen des Inter Annotator Agreement oder Visualisierungen zum Annotationsfortschritt und zur Annotationsexploration erstellen. Damit ermöglichen wir nicht nur die Annotationsauswertung, sondern auch die schnelle Identifizierung von Annotationsfehlern, die unmittelbar korrigiert werden können.

Insgesamt ist das zentrale Anliegen des Git Access, CATMA-Daten direkt verfügbar zu machen, damit Nutzer:innen nicht unbedingt an die schon in CATMA vorhandenen Funktionalitäten gebunden sind. Dadurch kann der Workflow zwischen Annotation, Annotationsauswertung und Annotationsüberarbeitung deutlich schneller werden. Das ist besonders für Nutzer:innen relevant, die sich – unter anderem im Rahmen von Forschungsprojekten – um die Organisation und Evaluierung von Annotationen kümmern.

Mit unserem Poster werden wir diesen Workflow detailliert darstellen. Das Poster soll also auch als eine Art Bedienungsanleitung für die Nutzung des CATMA Git Access fungieren und Best Practices zeigen. Dabei werden wir folgende Schritte abdecken:

1. Voraussetzungen für den Zugriff auf die CATMA GitLab API
2. Installation der CATMA Python Pakete (bzw. eines Docker Image, welches alle Erfordernisse abdeckt)
3. Clonen der Repositories
4. Zugriff auf die Daten mit Python
5. Beispiele für die Annotationsexploration und -auswertung

CATMA erscheint zum Beispiel im TAPoR Toolverzeichnis, sowie in „Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften“ (Frey-Endres & Simon 2021).

Bibliographie

Frey-Endres, Marcel / Simon, Tobias (2021): „Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften“. In: *Digital Philology | Working Papers in Digital Philology* 02/2021. Darmstadt: TUPrints. URL: https://tuprints.ulb.tu-darmstadt.de/17850/1/Digital_Philology_Working_Papers_in_Digital_Philology_vol002.pdf [letzter Zugriff 24. November 2021]

Gius, Evelyn / Meister, Jan Christoph / Meister, Malte / Petris, Marco / Bruck, Christian / Jacke, Janina / Schumacher,

Mareike / Gerstorfer, Dominik / Flüh, Marie / Horstmann, Jan (2021): CATMA 6 (Version 6.3). Zenodo. DOI: 10.5281/zenodo.1470118. URL: <https://catma.de/> [letzter Zugriff 24. November 2021]

Petris, Marco (2017): „TEI Export Format“. In: CATMA. URL: <https://catma.de/documentation/tei-export-format/> [letzter Zugriff 6. Juli 2021].

Petris, Marco (2020): „Git Access“. In: CATMA. URL: <https://catma.de/documentation/git-access/> [letzter Zugriff 6. Juli 2021].

Petris, Marco / Meister, Malte (2016): „Technology and Versions“. In: CATMA. URL: <https://catma.de/documentation/technology-and-versions/> [letzter Zugriff 6. Juli 2021].

Grenzüberschreitendes Textmining von Historischen Zeitungen

Das impresso-Projekt zwischen Text- und Bildverarbeitung, Design und Geschichtswissenschaft

Ehrmann, Maud

maud.ehrmann@epfl.ch

École polytechnique fédérale de Lausanne - EPFL

Bunout, Estelle

estellebunout@gmail.com

Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Clematide, Simon

siclemat@ifi.uzh.ch

University of Zurich

Düring, Marten

marten.during@uni.lu

Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Fickers, Andreas

andreas.fickers@uni.lu

Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Guido, Daniele

daniele.guido@uni.lu

Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Kalyakin, Roman

roman@kalyakin.com

Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Kaplan, Frederic

frederic.kaplan@epfl.ch
École polytechnique fédérale de Lausanne - EPFL

Romanello, Matteo

matteo.romanello@unil.ch
École polytechnique fédérale de Lausanne - EPFL

Schroeder, Paul

hello@youtag.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Ströbel, Philip

pstroebel@cl.uzh.ch
University of Zurich

van Beek, Thijs

thijs@midasweb.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Volk, Martin

martin.volk@uzh.ch
University of Zurich

Wieneke, Lars

lars.wieneke@uni.lu
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

impresso. Media Monitoring of the Past ist ein interdisziplinäres Forschungsprojekt, in dem Wissenschaftler und Wissenschaftlerinnen aus der Computerlinguistik, Design und den Geschichtswissenschaften an der Anreicherung eines Korpus aus schweizerischen und luxemburgischen Zeitungen arbeiten. Ziel ist es, die Qualität von Textmining-Werkzeugen für historische Zeitungstexte zu verbessern, letztere mit zusätzlichen Informationen anzureichern und schließlich mit Hilfe einer neu entwickelten Benutzeroberfläche in den historischen Forschungsprozess zu integrieren.

impresso adressiert die Herausforderungen, die große Sammlungen von digitalisierten und mit Daten angereicherten Zeitungen mit sich bringen. Diese lassen sich in fünf Kategorien zusammenfassen:

1. Dokumenten-Silos: Portale für digitalisierte Zeitungen bieten aufgrund rechtlicher Restriktionen und digitalisierungspolitischer Zwänge zwangsläufig unvollständige, nicht repräsentative Sammlungen, die einer automatisierten Verarbeitung in unterschiedlicher Qualität unterzogen wurden.

2. Große Mengen, großes Wirrwarr: Zeitungsdigitalisate sind durch Unvollständigkeit, Inkonsistenzen und Duplikate gekennzeichnet.

3. Textauschen: Unvollkommene OCR, fehlerhafte Artikelsegmentierung und das Fehlen geeigneter linguistischer Ressourcen beeinträchtigen die Robustheit von Bild- und Text-Mining-Algorithmen erheblich.

4. Generosity (Whitelaw 2015): Suche und Auffinden relevanter Inhalte in solch großen und heterogenen Korpora.

5. Transparenz: Kritische Beurteilung der inhärenten Verzerrungen in digitalisierten Quellen und den daraus extrahierten Daten.

Parallel zu *impresso*, haben in jüngster Zeit eine Reihe von ähnlich gelagerten Forschungsprojekten computergestützte Methoden für die Analyse digitalisierter historischer Zeitungen angewandt (Ridge et al. 2019). In diesem Beitrag präsentieren wir unseren Ansatz, den obigen Herausforderungen mit Hilfe eines interdisziplinären Teams und eines Co-Design-Ansatzes gerecht zu werden (Allen and Sieczkiewicz 2010; Atanassova 2014; Hechl et al. 2021; Ehrmann 2019). Im Folgenden geben wir einen Überblick über die wesentlichen Schritte der Dokumenten- und Textverarbeitung und deren Repräsentation innerhalb des Interfaces. Für jeden Schritt werden wir dessen Mehrwert, überwundene Schwierigkeiten und zukünftige Zielsetzungen darlegen. Ein Beispiel soll aber zunächst den Mehrwert des multilingualen, schweizerisch-luxemburgischen Korpus, der semantischen Anreicherungen und der Benutzeroberfläche illustrieren: Eine einfache Stichwortsuche nach Artikeln, die über die Schlacht um Arnheim seit dem Jahr 1944 berichten stößt in diesem Korpus schnell an ihre Grenzen: Die Stichwörter "arnheim" (deutsche Schreibweise) oder "arnhem" (niederländisch, englisch, französisch) liefern eine Vielzahl von irrelevanten Treffern verursacht durch den dortigen Fußball-Club Vitesse Arnhem, Werbeanzeigen und OCR-Fehler. Im Vergleich zu anderen Benutzeroberflächen für historische Zeitungen kann eine solche Suche in der *impresso*-App erheblich präziser formuliert und erweitert werden. Im Folgenden werden die wesentlichen Arbeitsschritte (Siehe auch Abb. 1) und Komponenten vorgestellt und, wo sinnvoll, durch das obige Beispiel illustriert.

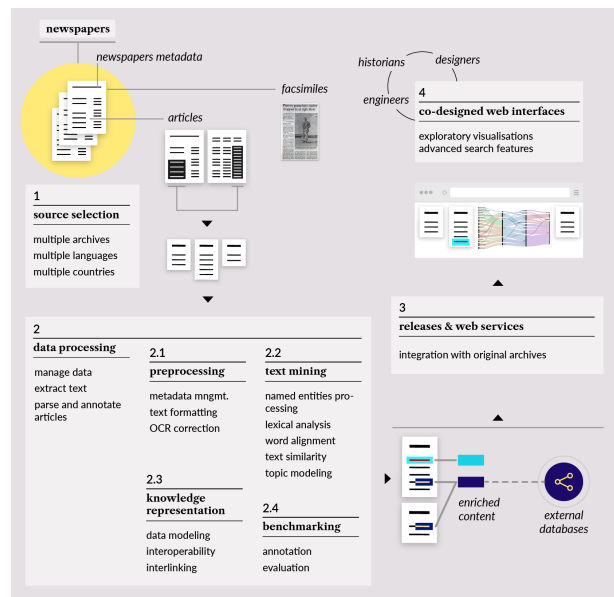


Abb. 1: Überblick über wesentlichen Arbeitsschritte des *impresso*-Projekts

Datenerfassung und Vorverarbeitung

Digitalisierte Zeitungen sind in “Silos” über viele Institutionen mit unterschiedlichen Zugriffsrichtlinien und Einrichtungen verstreut. Aus rechtlicher und administrativer Perspektive werden wir auf unsere Strategien zur Akquise und Inventarisierung digitaler Zeitungssammlungen eingehen und über unsere Vereinbarungen mit Datenanbietern berichten. Aus technischer Perspektive werden wir die Workflows skizzieren, die für den Umgang mit heterogenen Bild- und OCR-Formaten (Optical Character Recognition) entwickelt wurden, ebenso wie unsere Bemühungen um deren Standardisierung.

Digitale Dokumentenverarbeitung

Unser Ausgangspunkt sind – im Idealfall – Texte und Textblöcke, wie sie von OCR und OLR (Optical Layout Recognition) ausgegeben werden. Aufgrund der bereits erwähnten Heterogenität sind Texte und Layoutelemente jedoch meist “verrauscht” und müssen sorgfältig bewertet, korrigiert und zuweilen mittels OCR neu erfasst werden. Während des Projekts haben wir deshalb eine mehrsprachige OCR-Qualitätsbeurteilung, HTR-basierte Systeme für die Erkennung von Frakturschrift (Ströbel 2019) und die semantische Segmentierung von Zeitungen unter Verwendung textueller und visueller Merkmale (Barman 2019) getestet.

Lexikalische Verarbeitung

Nach der Texterkennung, bestand der nächste Schritt im Aufbau des impresso-Korpus in der Anwendung einer Reihe von linguistischen Vorarbeiten, einschließlich Sprachidentifikation, Tokenisierung, Normalisierung der historischen Rechtschreibung und Lemmatisierung. Historischer Sprachwandel und die Mehrsprachigkeit unseres Zeitungskorpus (französisch, deutsch, luxemburgisch) haben diese Aufgaben verkompliziert. Als ergänzende Ressourcen sind diese Vorarbeiten aber für nachfolgende Textverarbeitungsaufgaben nützlich. Desweiteren haben wir verteilte Repräsentationen von Wörtern (word embeddings) für jede Sprache im Korpus berechnet. N-Gram-Visualisierungen spiegeln die Veränderungen im Wortgebrauch in unserem Korpus im Laufe der Zeit wider; korpuspezifische word embeddings ermöglichen es dem Benutzer, verwandte Wörter zu finden und auch über verschiedene Sprachen hinweg zu vergleichen. Innerhalb des Interfaces dienen word embeddings zur Anzeige von Vorschlägen für die Schlagwortsuche und für N-gram-Analysen, indem sie synonyme oder verwandte Begriffe anzeigen (“arnheim”, “arnhem”), aber auch auf historische Schreibvarianten und häufige OCR-Fehler (“arnehm”) hinweisen. Die Benutzeroberfläche ermöglicht ebenfalls den Vergleich mehrerer N-grams in Kombination mit Suchanfragen oder innerhalb einzelner Artikel-Sammlungen, beispielsweise um die Präsenz der Schlacht von Arnheim mit jener von anderen Schlachten zu vergleichen.

Benannte Entitäten

Benannte Einheiten wie Namen von Personen, Orten und Organisationen liegen der Semantik von Texten zugrunde und sind von

entscheidender Bedeutung bei deren Interpretation. Ihre automatische Erkennung und Disambiguierung unterstützt das information retrieval und die Exploration großer Textsammlungen erheblich. Sie zeigen die wechselnden Kontexte auf, in denen Personen, Institutionen und Orte über Sprachen, Zeit und Zeitungen hinweg erscheinen. Durch die Verknüpfung werden beispielsweise die Entitäten “Michail Gorbatschow” und “Mikhail Gorbachev” derselben Person zugeordnet und durch kontextualisierende Informationen wie Lebensdaten angereichert. Innerhalb des Interfaces kann die Verteilung der Entitäten innerhalb des Korpus, auch über Sprachgrenzen hinweg verfolgt werden.

Topic Modelling

Wir ermitteln sprachspezifisch, welche “Themen” in unserem Zeitungskorpus vorkommen, um daraus eine Thematik für den Benutzer abzuleiten. Zu diesem Zweck werden mehrere topic models (über das gesamte Korpus, pro Zeitung, pro Zeitraum) berechnet und die Themen den Zeitungsartikeln zugeordnet. Topics erlauben es, überwältigende Ergebnismengen zu reduzieren, bestimmte Themenaspekte wie “Sport”, “Militär” oder “Kunst” ein- bzw. auszuschließen. Im obigen Beispiel würden also sportbasierte Themen aus- und militär-basierte Themen eingeschlossen. Innerhalb des Interfaces visualisiert eine Graph-Visualisierung Überschneidungen zwischen Topics. Knoten repräsentieren Topics, eine Kante zwischen zwei Knoten ist gewichtet und basiert auf der Zahl der Artikel, in denen beide Topics erkannt wurden.

Text reuse

Text reuse erkennt und vergleicht ähnliche Textpassagen und liefert Cluster von wiederverwendeten Text-Passagen, die sich in großen Sammlungen von Dokumenten auffinden lassen. Innerhalb des impresso-Projekts haben wir passim (Romanello 2018; Smith et al. 2014) verwendet. Das Interface zeigt Text reuse Passagen innerhalb von Artikeln an und eine eigene Komponente erlaubt die Stichwortsuche in allen Text reuse Clustern mit diversen Filter-Möglichkeiten um beispielsweise die Zirkulation von Agenturberichten über die Schlacht von Arnheim in der Presse-landschaft zu rekonstruieren.

The screenshot displays the 'Text Reuse Explorer' interface. At the top, there's a search bar with 'arnheim' entered. Below it, a cluster of results is shown for 'Cluster #c128849402104', which contains 2 articles from 1844 to 1944 with a 61.8% lexical overlap. The interface is divided into sections: 'OVERVIEW', '2 PASSAGES', and '8 CONNECTED CLUSTERS'. The 'PASSAGES' section shows two text snippets from historical newspapers, one dated Friday, October 13, 1944, and another from Friday, October 13, 1944. The 'CONNECTED CLUSTERS' section shows related clusters like '#c128849402104' and '#c128849402104'. The interface also includes a 'REFINE' section with filters for 'ORDER BY' and 'DATE RANGE'.

Abb. 2: Text Reuse Explorer

Bildähnlichkeits-Suche

Eine auf (Seguin 2018) basierende visuelle Suchmaschine ergänzt die Textsuche und erlaubt es, Kopien und einander ähnliche visuelle Elemente, wie z.B. Anzeigen, Fotos, Zeichnungen und Karten zu suchen. Die Bildähnlichkeitssuche erlaubt drei Formen der Interaktion: ausgehend von einem Bild können ähnliche Bilder gefunden werden, Bilder können über eine Stichwortsuche im umgebenden Text gefunden werden, externe Bilder können hochgeladen werden um zu prüfen, ob ähnliche Bilder im Korpus vorhanden sind.

Systemarchitektur

Text- und Bildverarbeitungskomponenten müssen in eine modulare Systemarchitektur integriert werden, die auch eine API, ein middle layer und ein front end umfasst. Wir haben eine technische Dokumentation veröffentlicht, die Informationen zu allen Schritten in der Aufbereitung und Anreicherung unseres Korpus und einer API enthält.

Die Benutzeroberfläche wurde mit dem Ziel entwickelt, Quellen- und datenkritisches Textmining für die breite Masse von historisch arbeitenden WissenschaftlerInnen anzubieten und neue Arbeitsabläufe für die Exploration historischer Zeitungen zu ermöglichen. Um eine hohe Anschlussfähigkeit an gängige Forschungspraxis zu gewährleisten, haben wir den Schwerpunkt auf das Auffinden von relevanten Inhalten gelegt. Inspiriert wurde die Gestaltung der Benutzeroberfläche durch die Prinzipien der Generosity, einem sinnbildlichen sich-öffnen des Korpus und der Transparenz, die eine informierte Nutzung und kritische Einordnung der gemachten Beobachtungen ermöglichen soll. Weiter angeregt durch die Gestaltung durch das in mehreren Workshops gesammelte Benutzerfeedback und durch das übergreifende Ziel, nahtlos zwischen close und distant reading Perspektiven zu wechseln. Um eine möglichst breite Anwendbarkeit und unterschiedlichste Workflows zu gewährleisten, wurde die Benutzeroberfläche so gestaltet, dass sie NutzerInnen im Rahmen des technisch möglichen die freie Kombination aller obenstehenden semantischen Anreicherungen und den auf ihnen basierenden Komponenten erlaubt. Dies wird ermöglicht durch eine Reihe von Komponenten, die innerhalb der gesamten Benutzeroberfläche verfügbar sind.

Search

Suchoperationen gehören zu den am häufigsten genutzten Interaktionen mit digitalisierten Quellen. Innerhalb der Benutzeroberfläche wurden deshalb Merkmale der klassischen "erweiterten" Suche mit semantischen Anreicherungen verknüpft um im Sinne der Generosity unterschiedliche Zugänge in das Material zu ermöglichen. Neben klassischen UND/ODER/NICHT Interaktionen, schlägt die Suche verwandte oder synonyme Suchbegriffe auf Basis des Korpus vor ebenso wie relevante Entitäten und Topics. In Kombination mit Filtern, die sowohl auf Bestandsmetadaten und semantischen Anreicherungen basieren, ermöglicht es die Suche in einem iterativen Prozess hochkomplexe Suchanfragen zu formulieren.

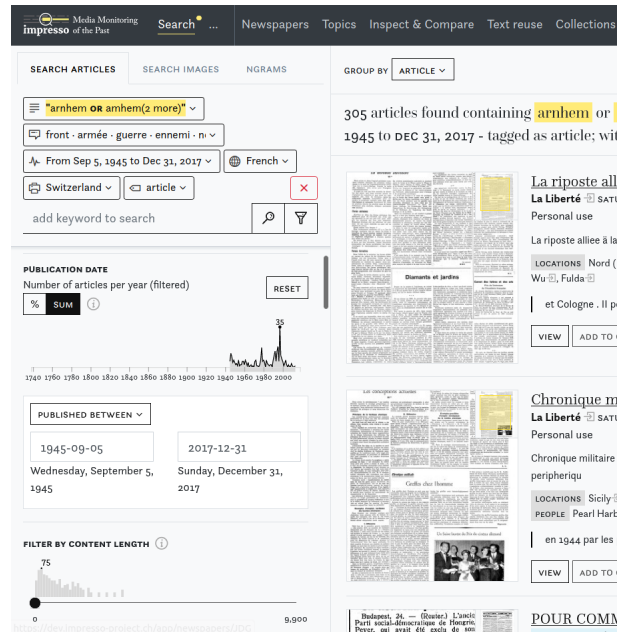


Abb. 3: Suche mit Hilfe von word embeddings, topics, Spracherkennung und Bestandsmetadaten

Collections

Ermöglichen den manuellen oder Suchabfragen-basierten Aufbau themenspezifischer Sammlungen mit bis zu 10.000 Artikeln, die als eigene Objekte ebenfalls durchsucht und gefiltert werden können.

Inspect&Compare

Inspect&Compare erlaubt den Vergleich von Suchanfragen und Sammlungen hinsichtlich ihrer Überlappungen und Differenzen mit Hilfe von Zeitleisten und Balkendiagrammen. Beispielsweise ließen sich eine Sammlung von Artikeln zu Arnheim mit einer Sammlung über die Schlacht von El-Alamein hinsichtlich ihrer Präsenz in der Medienberichterstattung vergleichen. Ebenso eignet sich die Komponente für den iterative Aufbau von Sammlungen durch experimentelle Variationen von Suchanfragen und dem Vergleich ihrer Ergebnisse. Figure X zeigt beispielsweise an, dass die angelegte Sammlung in (A) 182 Artikel mit dem Begriff „vi-tesse arnhem“ enthält.

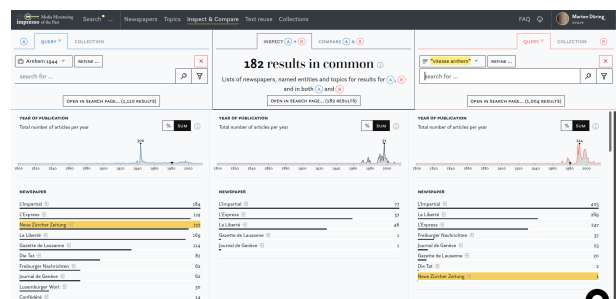


Abb. 4: Inspect&Compare, Vergleich einer Sammlung mit einer Suchabfrage

Überblicks-Fenster

Zu jedem Objekt innerhalb der Benutzeroberfläche, wie z.B. genannten Entitäten, Topics oder Zeitungen erlaubt ein korrespondierendes Überblicks-Fenster deren Präsenz innerhalb der laufenden Suche bzw. des Korpus zu prüfen und weitere Informationen, wie z.B. Metadaten abzurufen.

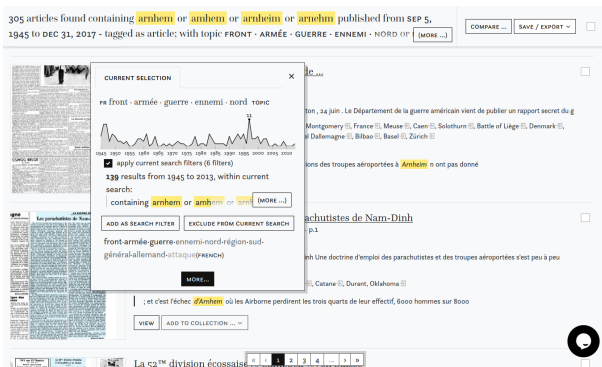


Abb. 5: Überblicks-Fenster eines topics

Export

Artikelsammlungen können mit allen semantischen Anreicherungen für die weitere Analyse außerhalb des Interfaces als .csv Dateien exportiert werden, je nach Rechtssituation auch inklusive des Artikel-Volltextes.

Artikelempfehlungen

Such- und Filteroperationen dienen dazu, relevante Inhalte innerhalb des Korpus zu identifizieren und neue Facetten auf das Quellenmaterial zu erhalten. Ein automatisiertes Empfehlungssystem hilft dabei, relevante Inhalte zu finden, die außerhalb des Suchbereichs der NutzerInnen liegen. Artikelempfehlungen basieren auf topics, zeitlicher Distanz, benannten Entitäten und text reuse, die voneinander unabhängig gewichtet werden können. Dies könnten Artikel sein, die zwar die Schlacht von Arnheim thematisieren und ähnliche benannte Entitäten erwähnen, aber nicht einem militär-basierten Topic zugeordnet wurden und deshalb aus den vorigen Ergebnislisten ausgeschlossen wurden.

Korpus-Übersicht

Visualisierung des Gesamtbestandes inklusive aller Bibliotheks-Metadaten, die auch Lücken, Verzerrungen im Korpus einschließt.

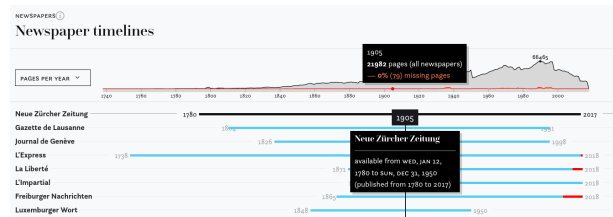


Abb. 6: Korpus-Übersicht

Lehr-/Lernmaterialien

Die Gestaltung der Benutzeroberfläche hat sich an den durchschnittlichen technischen Kompetenzen der historisch arbeitenden GeisteswissenschaftlerInnen orientiert, mit dem Ziel, NutzerInnen zu fordern ohne sie zu überfordern. Gleichzeitig, dem Anspruch der Transparenz folgend, sollten möglichst alle relevanten Methoden und Entscheidungen im Einreichungsprozess erklärt und dokumentiert werden. Diesen Zwecke erfüllen eine Reihe von FAQs, Blogartikeln, Tutorials und Videos, die das Projekt und die Funktionalitäten der Benutzeroberfläche dokumentieren.

Die Kombination aus semantischer Anreicherung, Design und historischen Forschungsinteressen hat - wie wir hier illustriert haben - zu einer Vielzahl von neuen Interaktionsmöglichkeiten mit historischen Zeitungen geführt. In Kombination tragen diese dazu bei, relevante Inhalte besser zu finden, close und distant reading zu integrieren und Vergleichsperspektiven einzunehmen.

Bibliographie

- Allen, Robert, and Robert Siczekiewicz.** 2010. "How Historians Use Historical Newspapers." *Proceedings of the American Society for Information Science and Technology* 47. <https://doi.org/10.1002/meet.14504701131>.
- Atanassova, Rossitza.** 2014. "Improving the Discovery of European Historic Newspapers." In *IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge*. Lyon, France. <http://library.ifla.org/1038/>.
- Barman, Raphaël.** 2019. "Historical Newspaper Semantic Segmentation Using Visual and Textual Features."
- Ehrmann, Maud.** 2019. "Historical Newspaper User Interfaces: A Review." In *IFLA 85 th . Athens, Greece: IFLA*. <http://library.ifla.org/2578/>.
- Hechl, Stefan, Pierre-Carl Langlais, Jani Marjanen, Sarah Oberbichler, and Eva Pfanzelter.** 2021. "Digital Interfaces of Historical Newspapers: Opportunities, Restrictions and Recommendations." *Journal of Data Mining & Digital Humanities* Histoinformatics (January). <https://doi.org/10.46298/jdmdh.6121>.
- Ridge, Mia, Giovanni Colavizza, Laurel Brake, Maud Ehrmann, Jean-Phillipe Moreux, and Andrew Prescott, eds.** 2019. "The Past, Present and Future of Digital Scholarship with Newspaper Collections." In *DH 2019 Book of Abstracts*.
- Romanello, Matteo.** 2018. "Detecting Text Reuse in Newspapers Data with Passim." Presented at the Hacking the News Workshop in conjunction with DHN 2018, Helsinki. <http://dig-hum-nord.eu/>.
- Seguin, Benoît Laurent Auguste.** 2018. "Making Large Art Historical Photo Archives Searchable." Lausanne: EPFL.

Smith, David A., Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. "Detecting and Modeling Local Text Reuse." In *IEEE/ACM Joint Conference on Digital Libraries*, 183–92. <https://doi.org/10.1109/JCDL.2014.6970166>.

Ströbel, Phillip. 2019. "Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images." Presented at the *Digital Humanities 2019: Complexities, Utrecht, December 7*. https://www.conftool.pro/dh2019/index.php?page=brows%20eSessions&path=adminSessions&print=export&ismobile=%20false&form_session=481&presentations=show.

Whitelaw, Mitchell. 2015. "Generous Interfaces for Digital Cultural Collections" 9 (1). <http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html>.

Informationstechnologische Gedächtnisarbeit in der Rezensionenzeitschrift RIDE

Henny-Krahmer, Ulrike

ulrike.henny-krahmer@uni-rostock.de
Universität Rostock, Germany

Neuber, Frederike

neuber@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Scholger, Martina

martina.scholger@uni-graz.at
Universität Graz

Seit 2014 gibt das Institut für Dokumentologie und Editorik (IDE) die digitale Rezensionenzeitschrift RIDE (A Review Journal for Digital Editions and Resources) heraus.¹ Mit derzeit (Stand Juli 2021) 70 Rezensionenartikeln zu digitalen wissenschaftlichen Editionen, Textsammlungen sowie Tools und Forschungsumgebungen ist es RIDE in den letzten sieben Jahren gelungen, Rezensionen verstärkt in den Digital Humanities zu verankern. Sowohl RIDE als auch die vom IDE herausgegebenen Kriterienkataloge für digitale Ressourcen wurden u.a. im Rahmen der DHd-Konferenzen kritisch diskutiert (Henny 2018, Neuber et al. 2018, Resch und Rastinger 2020), als Suchkriterium zum Auffinden digitaler Editionen angeführt (Franzini 2012) oder als Unterstützung bei der Korpuserstellung (Ruiz Fabio et al. 2017) und Verbesserung bestehender Applikationen verwendet (Virtueller Forschungsverbund Edirom 2021)². Bei der DHd 2022 möchten wir mit einem Poster herausstellen, was es bedeutet, die Arbeit an einem digitalen Rezensionenjournal als "informationstechnologische Datenbankarbeit, als Aushandlung von Metadatenstandards, als digitalen Sammlungs- und Korpuserstellung, als Korpuserstellung, als FAIR Research Data Management zu begreifen" (Call for Papers).

Das IDE zeichnet für die digitale Infrastruktur und den Publikationsworkflow von RIDE selbst verantwortlich. Die Publikationsinfrastruktur besteht aus einem Zusammenspiel von eXist-db und Wordpress. Von Beginn an wurden die Rezensionen in XML/TEI codiert, als HTML und PDF publiziert und auch die zugrunde-

liegenden TEI-Versionen auf GitHub zum Download angeboten.³ Begleitend zu den Rezensionstexten werden in RIDE außerdem *Factsheets* veröffentlicht, die auf der Grundlage eines ausgefüllten Fragebogens für jede rezensierte Ressource erstellt werden und wesentliche Informationen in strukturierter Form enthalten. So wird es möglich, vergleichbare Auswertungen über alle Rezensionen hinweg vorzunehmen, die auf der RIDE-Webseite in Form von Diagrammen veröffentlicht werden.⁴

Seit etwa 2019 wurde der Workflow von RIDE mit Blick auf technologisch nachhaltige und standardisierte Lösungen sowie die Umsetzung der FAIR-Prinzipien überarbeitet und neu konsolidiert. Die Grundlage dafür bildete die Erstellung eines neuen TEI-Schemas in ODD ("One Document Does it all")⁵. Die Erfahrungen der ersten fünf Jahre haben durch die Rezensionspraxis und Bandbreite der Beiträge gezeigt, in welchen Bereichen das Datenschema enger gefasst und wo es noch erweitert werden musste. Das neue RIDE-Schema orientiert sich so weit wie möglich am Schema des *Journals of the Text Encoding Initiative* (jTEI).⁶ Auch die Abbildung der analytisch komplexen Fragebögen, einem Alleinstellungsmerkmal RIDEs gegenüber anderen Zeitschriften, wurden in das Schema integriert. Neue Elemente des RIDE-Datenmodells sind u.a. die detaillierte Sichtbarmachung der an der Publikation beteiligten Personen und ihrer Rollen, unter Einbeziehung von Normdaten.⁷ Die bisher erschienenen Bände wurden bereits dem neuen Schema entsprechend konvertiert.

RIDE ist eine von vielen digitalen Open Access-Zeitschriften in den DH, darunter beispielsweise die *Zeitschrift für digitale Geisteswissenschaften* (ZfDG), *Digital Humanities Quarterly* (DHQ) oder *Journal of the TEI* (jTEI), die Datensätze ihrer Beiträge veröffentlichen. Im Gegensatz zu den genannten Zeitschriften stellt RIDE aber nicht nur singuläre Daten (d.h. XML/TEI-Files), sondern auch den gesamten Datenbestand in zitierfähiger, nachhaltiger und (nach-)nutzerfreundlicher Form unter CC BY Lizenz auf GitHub und Zenodo (mit DOI, in verschiedenen Versionen) zur Verfügung (Institut für Dokumentologie und Editorik 2021). Die Metadaten der Beiträge werden über eine OAI-Schnittstelle ausgeliefert, die bereits von der Deutschen Nationalbibliothek genutzt wird.⁸ Darüber hinaus ist ein Großteil der Skripte, die zur Generierung der digitalen Rezensionen verwendet werden, ebenfalls öffentlich verfügbar.⁹ Damit ist RIDE nicht nur Open Access, sondern auch FAIR.

Das Poster wird die technischen Grundlagen RIDEs illustrieren und damit einen Blick "hinter die Kulissen" gewähren. Neben einer Beschreibung des derzeitigen Workflows soll der Beitrag demonstrieren, dass die strukturelle und semantische Codierung der Inhalte sowie die Begleitung der Rezensionen durch Kriterien und Fragebögen dazu führen, dass RIDE neben dem Angebot von Texten auch vermehrt datengestützte Zugänge bietet. Dies verstärkt den Blick auf die Gesamtheit der Rezensionen und die rezensierten Gegenstände und fördert rezensitionsübergreifende Analysen.¹⁰ Damit verändert sich die Kultur des digitalen Rezensierens und die herausgeberische Arbeit wird vermehrt zur informationstechnologischen Gedächtnisarbeit.

Fußnoten

1. <https://ride.i-d-e.de>

2. Siehe das Label "ride.a.12.4" in den Issues zur WeGA-WebApp: <https://web.archive.org/web/20210706124100/https://github.com/Edirom/WeGA-WebApp/labels/ride.a.12.4>

3. <https://github.com/i-d-e/ride>

4. Siehe <https://ride.i-d-e.de/data/charts-scholarly-editions/> für die Diagramme zu wissenschaftlichen Editionen und <https://ride.i-d-e.de/data/charts-text-collections/> zu Textsammlungen.
5. Siehe <https://github.com/i-d-e/ride/tree/master/schema> ; zu ODD vgl. <https://wiki.tei-c.org/index.php/ODD> .
6. <https://tei-c.org/guidelines/customization/jtei/>
7. Während das ODD-Schema bisher als technisches Schema im Einsatz ist, soll es als menschenlesbare Dokumentation des Datenmodells zukünftig ermöglichen, dass AutorInnen Einreichungen direkt in XML/TEI umsetzen.
8. <https://ride.i-d-e.de/data/>
9. Die Skripte sind unter einer GNU General Public License auf GitHub veröffentlicht und können nachgenutzt werden: <https://github.com/i-d-e/ride-scripts> .
10. Wie sie z.B. von RIDE selbst in Form der Diagramme angeboten werden oder von Resch et al. (2020) als externe Nutzerinnen durchgeführt wurden.

Bibliographie

- ACH / ADHO (ed.) (2007-2021): Digital Humanities Quarterly (DHQ). <http://www.digitalhumanities.org/dhq/>
- Franzini, Greta (2021): "Catalogue of Digital Editions." Version 4.0.0. <https://dig-ed-cat.acdh.oeaw.ac.at>
- Henny, Ulrike (2018): "Reviewing von digitalen Editionen im Kontext der Evaluation digitaler Forschungsergebnisse", in: Kamzelak, Roland S. / Steyer, Timo (eds.): *Digitale Metamorphose: Digital Humanities und Editionswissenschaft*. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 2). DOI: 10.17175/sb002_006
- Institut für Dokumentologie und Editorik (ed.) (2021): "Datasets of RIDE (A Review journal for digital editions and resources)." Version 2.1.0. GitHub.com. <https://github.com/i-d-e/ride>. DOI: 10.5281/zenodo.5081646
- MWW / DHd (ed.) (2016-2021): Zeitschrift für digitale Geisteswissenschaften (ZfdG). <https://zfdg.de/>
- Neuber, Frederike / Henny-Krahmer, Ulrike / Sahle, Patrick / Fischer, Franz (2018): "Alles ist im Fluss – Ressourcen und Rezensionen in den Digital Humanities", Panelorganisation bei der DHd2018, 28. Februar 2018, Köln. Invited Speakers: Rüdiger Hohls, Anne Baillot, Christof Schöch, Jürgen Hermes. Zenodo. DOI: 10.5281/zenodo.4622507
- Resch, Claudia / Rastinger, Nina (2020): "Digitale Editionen im Spannungsfeld zwischen Formalisierung und Interpretation: Rezensionen der Online-Zeitschrift RIDE als Gradmesser für die Zukunft", in: Schöch, Christoph / Helling Patrick (eds.): *DHd 2020. Konferenzabstracts*. Zenodo. DOI: 10.5281/zenodo.4621960
- Ruiz Fabo, Pablo / Bermúdez Sabel, Helena / Martínez Cantón, Clara / Calvo Tello, José (2017): "Diachronic Spanish Sonnet Corpus (DISCO)." Madrid: UNED. <https://github.com/pruizf/disco> . DOI: 10.5281/zenodo.3841583
- Text Encoding Initiative Consortium (ed.) (2011-2021): Journal of the Text Encoding Initiative (JTEI). <https://journals.openedition.org/jtei/>
- Virtueller Forschungsverbund Edirom (2021): "WeGA-WebApp." Version 4.3.0. GitHub.com. <https://github.com/Edirom/WeGA-WebApp> . DOI: 10.5281/zenodo.4487114

„Ja, jetzt ist das langweilig. Aber in zwanzig Jahren!“ Bereitstellung, Zugang und Analyse literarischer Blogs am Beispiel des Techniktagebuchs

Blessing, André

andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für maschinelle Sprachverarbeitung

Hess, Jan

jan.hess@dla-marbach.de
Deutsches Literaturarchiv Marbach, Germany

Jung, Kerstin

kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Institut für maschinelle Sprachverarbeitung

„Ja, jetzt ist das langweilig. Aber in zwanzig Jahren!“ – Wie der Untertitel des *Techniktagebuchs*¹ bereits suggeriert, versteht sich das 2014 von der Schriftstellerin Kathrin Passig ins Leben gerufene literarische Blog als Teil des kulturellen Gedächtnisses. Mehr als sechzig regelmäßige Autor:innen bloggen hier oftmals humorvoll-pointiert über ihre Erlebnisse, Erfahrungen, Erfolge und Misserfolge im Zusammenhang mit technischen Themen und Gegenständen. Bereits seit 2008 werden literarische Blogs wie das *Techniktagebuch* am Deutschen Literaturarchiv Marbach gesammelt und archiviert. Im Rahmen des Projekts SDC4Lit – Science Data Center for Literature² werden nicht nur der Workflow zur Archivierung von Literatur im Netz aktualisiert, sondern auch Methoden und Werkzeuge zur (explorativen) Analyse von elektronischer Literatur getestet und (weiter-)entwickelt, um sie in Form von Analyse-Pipelines über die SDC4Lit-Plattform zur Verfügung zu stellen (Schlesinger 2021).

Am Beispiel des *Techniktagebuchs* soll der hier vorgeschlagene Poster-Beitrag einerseits die Herausforderungen bei der wissenschaftlichen Arbeit mit diesen bislang wenig erforschten „born-digital“ Literaturformen³ aufzeigen sowie andererseits vor allem auch (explorative) Zugangs- und Analysemöglichkeiten literarischer Blogs präsentieren. Als Standard für die Archivierung von Webinhalten hat sich das WARC-Format (IPCC) etabliert. Resultierend aus der Blog-Software enthält der im Falle des *Techniktagebuchs* ca. neun Gigabytes große Blog-Crawl⁴ neben den 7677 Blogbeiträgen eine Vielzahl an Metaseiten wie dem aus Tags bestehenden Stichwortverzeichnis oder den Blog-Archivseiten, die für die reine Inhaltsanalyse störend sind und entsprechend vorab identifiziert werden müssen. Eine Besonderheit des *Techniktagebuchs* und zugleich Grund für die Auswahl von Tumblr als Blogging-Plattform⁵ ist zudem die Möglichkeit zur Rückdatierung einzelner Blogposts. Neben den aus aktuellen (technischen) Erlebnissen der Autor:innen resultierenden und entsprechend auf den tatsächlichen Schreibzeitpunkt datierten Beiträgen enthält das *Techniktagebuch* zahlreiche retrospektivisch verfasste und entsprechend zurückdatierte, sodass das eigentliche Publikationsdatum zum Teil über die in der jeweiligen URL ent-

haltene Tumblr-ID ermittelt werden muss. Bereits aus der Discrepanz zwischen dem eigentlichen Publikationsdatum und dem von den verschiedenen Autor:innen deklarierten Datum ergeben sich verschiedene Fragestellungen: Welche Autor:innen schreiben eher über Aktuelles, welche eher über Vergangenes? Lassen sich dabei jeweils thematische Schwerpunkte ermitteln? Wie verändern sich diese im zeitlichen Verlauf? Gibt es thematische "Trendsetter" unter den Autor:innen? Inwiefern beeinflussen sie sich gegenseitig?

Um sich Fragestellungen wie diesen nähern zu können, wurden die in den WARC-Records enthaltenen Daten zunächst mithilfe von WARC-Bibliotheken wie WARCIO⁶ und JWARC⁷ aufbereitet und die zu analysierenden Texte mithilfe von trafilatura (Barbaredi 2019) extrahiert. Mittels regelbasierter Informationsextraktion (Blessing 2014) wurden hieraus u. a. die in den Blogposts im Normalfall nur in Klammern am Textende genannten Autor:innen ermittelt. Basierend auf den so gewonnenen Informationen wurden verschiedene Netzwerke zur Visualisierung verschiedener Zusammenhänge erstellt. Analysiert wurden so beispielsweise die direkten Referenzierungen der Autor:innen innerhalb der Blogposts unter Berücksichtigung der deklarierten Publikationsdaten. Dabei wurde u. a. sichtbar, dass die Links zwar vorwiegend genutzt werden, um auf auch hinsichtlich des deklarierten Datums früher entstandene Blogposts zurück zu verweisen oder gar Fortsetzungsgeschichten in Form mehrerer Blogbeiträge aufzubauen. Die hier dennoch zahlreich zutage tretenden, vermeintlich anachronistischen Verweise auf künftige Beiträge verdeutlichen aber zugleich, dass die Autor:innen die Rückdatierungsoption durchaus kreativ zu nutzen wissen.

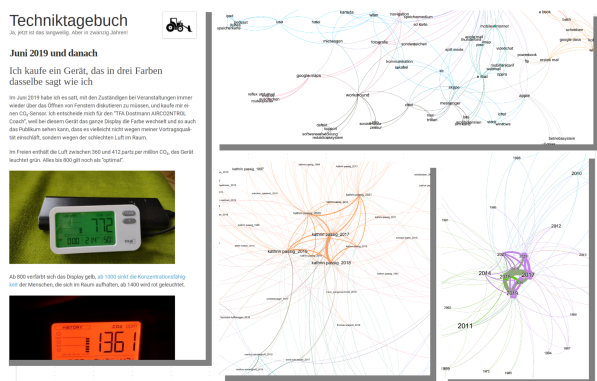


Abb. 1: Collage von Gephi-Visualisierungen zur Analyse des Technikagebuchs.

Netzwerkvisualisierungen zu den von den Autor:innen vergebenen Tags in Kombination mit dem Einsatz von Topic-Model-Analysen (Blei 2003) zeigen demgegenüber, dass diese manuell definierten Schlagwörter der Autor:innen nicht konsistent, mitunter eher humoristisch verwendet werden. Um das *Technikagebuch* korpuslinguistisch und stilometrisch auszuwerten, kamen bislang CQPWeb (Hardie 2012) sowie Stylo (Eder 2016) zum Einsatz, sodass auch hinsichtlich des Vokabulars und des individuellen Blog-Schreibstils Unterschiede aufgezeigt werden können. Interessant und überraschend ist beispielsweise bereits die später auf dem Poster zu findende Antwort auf die Frage nach den fünf Wörtern, die Kathrin Passig wesentlich häufiger nutzt als alle anderen Autor:innen.

Mit Blick auf die bereits erwähnte SDC4Lit-Plattform werden die genannten Methoden und Werkzeuge nicht nur hinsichtlich ihrer konkreten Nutzens bei der beispielhaften, explorativen Ana-

lyse des *Technikagebuchs* getestet, (weiter-)entwickelt und in Form von Analysepipelines kombiniert, sondern auch hinsichtlich ihrer Verwendbarkeit für die über das SDC4Lit-Repositorium zur Verfügung gestellten Materialien von Literatur im Netz. Letztlich sollen die erstellten Analyse-Pipelines demnach nicht nur auf anderen (literarischen) Blogs, sondern auch auf Werken der Netzliteratur Anwendung finden können, um möglichst niedrigschwellige Zugangsmöglichkeiten zu diesen trotz ihrer genuinen Digitalität eher vernachlässigten Literaturformen zu schaffen, die keinesfalls erst „in zwanzig Jahren“ in den (wissenschaftlichen) Fokus rücken sollten.

Fußnoten

1. Technikagebuch, <https://technikagebuch.tumblr.com/>, Ab-ruf: 15.7.2021.
2. SDC4Lit, <https://sdc4lit.de/>, letzter Zugriff: 15.7.2021.
3. In Untersuchungen literarischer Blogs finden computerge-stützte Analysemöglichkeiten bislang kaum Verwendung, vgl. Fassio (2021), Knapp (2012, 2014) und Ainetter (2006).
4. Der Blogcrawl wurde erstellt mithilfe von Brozzler (<https://github.com/internetarchive/brozzler>, letzter Zugriff: 30.11.2021) und Heritrix (<https://github.com/internetarchive/heritrix3>, letzter Zugriff: 30.11.2021).
5. Technikagebuch, <https://technikagebuch.tumblr.com/post/76963766003/20140218>, letzter Zugriff : 15.07.2021.
6. WARCIO: WARC (and ARC) Streaming Library, <https://github.com/webrecorder/warcio>, letzter Zugriff: 30.11.2021.
7. JWARC, Java library for reading and writing WARC files with a typed API, <https://github.com/iipc/jwarc>, letzter Zugriff: 30.11.2021.

Bibliographie

- Ainetter, Sylvia** (2006): Blogs - literarische Aspekte eines neuen Mediums. Eine Analyse am Beispiel des Weblogs Miagolare (= Innsbrucker Studien zur Alltagsrezeption 5). Wien: Lit Verlag.
- Barbaredi, Adrien** (2019): "Generic Web Content Extraction with Open-Source Software", in: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), 267–268.
- Blei, David M. / Ng, Andrew Y. / Jordan, Michael I.** (2003): "Latent dirichlet allocation", in: Journal of Machine Learning Research 3, 993–1022.
- Blessing, André / Kuhn, Jonas** (2014): "Textual Emigration Analysis (TEA)", in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) European Language Resources Association (ELRA), Reykjavik, Iceland, 2089–2093.
- Eder, Maciej / Rybicki, Jan / Kestemont, Mike** (2016): "Stylo-metry with R: a package for computational text analysis", in: R Journal, 8(1), 107–121.
- Fassio, Marcella** (2021): Das literarische Weblog. Praktiken, Poetiken, Autorschaften (= Praktiken der Subjektivierung 21). Bielefeld: Transcript.
- Hardie, Andrew** (2012): "CQPweb – combining power, flexibility and usability in a corpus analysis tool", in: International Journal of Corpus Linguistics, 17(3), 380–409.

IIPC: The WARC Format. <https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/>, letzter Zugriff: 15.07.2021.

Knapp, Lore (2012): "Christoph Schlingensief's Blog. Multimediale Autofiktion im Künstlerblog", in: Ansgar Nünning und Jan Rupp (Hrsg.), *Narrative Genres im Internet: Theoretische Bezugsrahmen, Mediengattungstypologie und Funktionen*. Trier: Wissenschaftlicher Verlag, 117-132.

Knapp, Lore (2014): *Künstlerblogs. Zum Einfluss der Digitalisierung auf literarische Schreibprozesse* (Goetz, Schlingensief, Herrndorf), Berlin: Ripperger & Kremers.

Schlesinger, Claus-Michael / Ulrich, Mona / Hein, Pascal / Blessing, André (2021): *Networks of Net Literature - Modeling, Extracting and Visualizing Link-Based Networks in the DLA corpus of net literature*. Bergen: ELMCIP 2021, <https://elm-cip.net/node/16380>, letzter Zugriff: 30.11.2021.

Kontrastive Textanalyse mit pydistinto

Ein Python-Paket zur Nutzung unterschiedlicher Distinktivitätsmaße

Du, Keli

duk@uni-trier.de
Universität Trier, Germany

Dudar, Julia

dudar@uni-trier.de
Universität Trier, Germany

Rok, Cora

rok@uni-trier.de
Universität Trier, Germany

Schöch, Christof

schoech@uni-trier.de
Universität Trier, Germany

Viele Wissenschaftsbereiche, die sich mit der quantitativen Textanalyse beschäftigen, wie die Korpuslinguistik oder die Computational Literary Studies (CLS) setzen verschiedene statistische Distinktivitätsmaße ein, um Elemente (z.B. Wortformen oder Wortarten) zu bestimmen, die charakteristisch für eine Textgruppe im Vergleich mit einer anderen Textgruppe sind. Tools wie z.B. *WordSmith* (Scott 2020) oder *AntConc* (Anthony 2005), die solche Analysen ermöglichen, sind weit verbreitet, haben jedoch einige Nachteile: Die meisten bieten nur häufigkeitsbasierte Maße (z.B. Log-Likelihood-Ratio Test oder Chi-Squared Test) an, die in vielen Fällen Ergebnisse produzieren, die für die kontrastive (explorative) Textanalyse nicht hilfreich sind (siehe u.a. Baker 2004 und Johnson and Ensslin 2006). Dispersionsmaße wie z.B. DP (Gries 2008) oder dispersionsbasierte Distinktivitätsmaße wie z.B. Zeta (Burrows 2007), die besser interpretierbaren Ergebnisse liefern (siehe Gries 2021; Schöch 2018), werden dagegen

nicht implementiert. Eine Ausnahme bildet *stylo*, das Zeta implementiert (Eder et al. 2016). Ein weiterer Nachteil ist, dass bei den meisten Tools nur ein oder zwei Maße für die Analyse ausgewählt werden können, was einen Vergleich der unterschiedlichen Maße erschwert. Gerade wenn Nutzende ihre Analysen anpassen und eigene Parametereinstellungen vornehmen oder verschiedene Datenformate nutzen wollen, erweisen sich die meisten Tools als ungeeignet.

Um den Einsatz relevanter Maße für die kontrastive Textanalyse zu erleichtern und das Bewusstsein für die Vielfalt der Maße zu schärfen, entwickeln wir im Rahmen des Projekts „Zeta and Company“ ein Python-Paket mit dem Namen *pydistinto*.¹ Ziel unseres Projekts ist es, zu einem tieferen Verständnis der verschiedenen Distinktivitätsmaße zu gelangen und Verbesserungen für deren Implementierung und Anwendung vorzuschlagen. Mithilfe von *pydistinto* können zwei Textkorpora mit unterschiedlichen Maßen verglichen werden.

Hierfür haben wir zunächst ein konzeptionelles Framework erstellt, auf dessen Basis die Maße in *pydistinto* implementiert werden (Du et al. 2021a). Das Framework definiert die Bereiche Preprocessing, Berechnung von Häufigkeiten, Korpusaufteilung sowie der eigentlichen Berechnung der Distinktivitätswerte, Visualisierung sowie quantitative und qualitative Evaluation der Ergebnisse.

In der Implementierung umfasst das Preprocessing die Tokenisierung, Lemmatisierung und das POS-Tagging der Texte. Danach werden die Texte je nach Parameter entweder segmentiert (dies wird bei der Berechnung von dispersionsbasierten Maßen empfohlen) oder als ganze Dokumente belassen. Die (absoluten, binären, relativen usw.) Worthäufigkeiten in den Segmenten bzw. Dokumenten werden in einer Matrix zusammengefasst. Als Nächstes werden die Segmente bzw. Dokumente in zwei Gruppen, ein Ziel- und ein Vergleichskorpus, aufgeteilt. Anschließend werden die Distinktivitätswerte auf Basis der Worthäufigkeits-Matrizen berechnet und die distinktiven Wörter für das Zielkorpus visualisiert. Die Implementierung des Moduls zur quantitativen Evaluation steht noch aus. Geplant ist hier, die statistischen Eigenschaften der Wortlisten zu analysieren und die Korrelation verschiedener Maße zu untersuchen (siehe, für Zwischenergebnisse, Du et al. 2021c). Bei der qualitativen Evaluation werden die ausgegebenen Wörter manuell interpretiert und ihre Relevanz für das Zielkorpus wird beurteilt.

Das Python-Paket wird auf Github veröffentlicht und steht somit zur freien Nutzung, eigenen Anpassung und weiteren Entwicklung zur Verfügung (Du et al. 2021b). Im *pydistinto* sind derzeit folgende Distinktivitätsmaße implementiert: Zeta, Ratio of Relative Frequencies, Gris' Deviation of Proportions based measure (Eta, siehe Du et al. 2021c), Welch's T-test, Wilcoxon Rank-sum Test, Kullback-Leibler Divergence, Chi-Squared Test, Log-Likelihood-Ratio Test, TF-IDF. Ein besonderer Vorteil des Pakets ist, dass es in einem Beginner-Modus und einem Profi-Modus genutzt werden kann. Im Beginner-Modus können auch weniger erfahrene Nutzende mit geringen Programmier- und Statistikkenntnissen Textkorpora vergleichen. Ziel- und Vergleichskorpus müssen hierfür lediglich als 'plain text' vorbereitet und einige Parameter wie z. B. Segmentlänge, Feature-Typen oder Anzahl der Top-Features eingestellt werden. Die Analyse wird dann automatisch durchgeführt und eine Visualisierung angeboten. Wer sich für die statistischen Eigenschaften der unterschiedlichen Maße interessiert und diese vergleichen möchte, kann den Profi-Modus verwenden. Die Nutzenden können dann selbst darüber bestimmen, welche Maße und statistischen Eigenschaften der Features (z.B. absolute Häufigkeit, relative Häufigkeit, Dispersion) für die Berechnung der Distinktivität kombiniert werden sol-

len. Es gibt in diesem Modus außerdem zusätzliche Möglichkeiten, die Daten zu visualisieren: so kann die Abhängigkeitsstruktur zweier statistischer Merkmale (z.B. Zeta-Wert und absolute Häufigkeiten der Features) auch durch ein Streudiagramm dargestellt werden.

Durch die Entwicklung des Pakets möchten wir auf der einen Seite eine reflektierte Nutzung statistischer Distinktivitätsmaße für die kontrastive Textanalyse erleichtern. Auf der anderen Seite soll das Paket ermöglichen, die Eigenschaften und Leistungsfähigkeit der Maße empirisch zu ermitteln und systematisch zu vergleichen.

Fußnoten

1. Das Projekt gehört zum DFG-geförderten Schwerpunktprogramm "Computational Literary Studies" (SPP 2207) und läuft von 2020-2023. Weitere Informationen unter <https://zeta-project.eu/de/>.

Bibliographie

Baker, Paul (2004): "Querying keywords: questions in difference, frequency, and sense in keyword analysis", in: *Journal of English Linguistics* 32 (4), pp. 346–59

Du, Keli / Dudar, Julia / Rok, Cora / Schöch, Christof (2021a): Implementation framework of measures of distinctiveness. Zenodo. <http://doi.org/10.5281/zenodo.5092328>

Du, Keli / Dudar, Julia / Schöch, Christof (2021b): pydistinto. Version 0.1.0. Verfügbar unter: <https://github.com/Zeta-and-Company/pydistinto>. DOI: <https://doi.org/10.5281/zenodo.5094346>.

Du, Keli / Dudar, Julia / Rok, Cora / Schöch, Christof (2021c): "Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness", in: *CHR 2021: Computational Humanities Research Conference*, November 17–19, 2021, Amsterdam, The Netherlands, <https://2021.computational-humanities-research.org/conference/>

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): "Stylometry with R: a package for computational text analysis", in: *R Journal*, 8(1): 107–21. <https://journal.r-project.org/archive/2016/RJ-2016-007/index.htm>

Gries, Stephan (2008): "Dispersions and adjusted frequencies in corpora", in: *International Journal of Corpus Linguistics*, Volume 13(4): 403–437. DOI: <https://doi.org/10.1075/ijcl.13.4.02gri>

Gries, Stephan (2021): "A New Approach to (Key) Keywords Analysis: Using Frequency, and Now Also Dispersion", in: *Research in Corpus Linguistics*, 9, 1–33. DOI: <https://doi.org/10.32714/ricl.09.02.02>

Johnson, Sally / Ensslin, Astrid (2006): "Language in the news: some reflections on keyword analysis using WordSmith Tools and the BNC", in: *Leeds Working Papers in Linguistics and Phonetics* 11, pp. 96–109. https://www.latl.leeds.ac.uk/wp-content/uploads/sites/49/2019/05/Johnson-Ensslin_2006.pdf

Laurence, Anthony (2005): "AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit", in: *Proceedings of IWLLeL 2004: An Interactive Workshop on Language e-Learning*, 7–13.

Schöch, Christof (2018): "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie", in: Bernhart, T., et al. (eds.), *Quantitative Ansätze in der Literatur- und*

Geisteswissenschaften. Berlin: de Gruyter, 77–94. <https://www.degruyter.com/viewbooktoc/product/479792>.

Scott, Mike (2020): WordSmith Tools, Version 8, Stroud: Lexical Analysis Software.

Leben, Werke und Datensilos Zur Verknüpfung und Visualisierung von im/materiellen Komponenten des kulturellen Erbes

Mayr, Eva

eva.mayr@donau-uni.ac.at
Universität für Weiterbildung Krems, Austria

Liem, Johannes

johannes.liem@donau-uni.ac.at
Universität für Weiterbildung Krems, Austria

High-Steskal, Nicole

nicole.high-steskal@donau-uni.ac.at
Universität für Weiterbildung Krems, Austria

Grebe, Anja

anja.grebe@donau-uni.ac.at
Universität für Weiterbildung Krems, Austria

Windhager, Florian

florian.windhager@donau-uni.ac.at
Universität für Weiterbildung Krems, Austria

Hintergrund

In den letzten Jahren wurde die Digitalisierung der Objektsammlungen von zahlreichen Kulturerbe-Institutionen vorangetrieben. Materielle Kulturgüter aus europäischen Museen, Archiven und Bibliotheken sind als Digitalisate in großem Umfang auf transnationalen Plattformen wie Europeana.eu einer breiten Öffentlichkeit und der Wissenschaft zugänglich gemacht worden. Gleichzeitig, aber davon unabhängig, wurde immaterielles Kulturerbe – wie biografisches Wissen über bedeutende nationale Persönlichkeiten – digital erfasst und als strukturierte und verknüpfte Aggregate in Biographie-Datenbanken verfügbar gemacht. Diese Entwicklungen bieten eine gute Basis für eine digital vermittelte Rezeption, Analyse und Kommunikation von historischen Beständen zum kulturellen Erbe. Jedoch verhindern fehlende Verknüpfungen (zwischen den biographischen Datenbanken und Europeana) und Standardisierungen (zwischen den biographischen Datenbanken der verschiedenen Nationen) sowie mangelnde (Maschinen-)Lesbarkeit und Sichtbarkeit lokaler Datensammlungen oft eine optimale Nutzung – für die wissenschaftliche Analyse durch Expert*innen ebenso wie für ein besseres Verständnis von

kulturhistorischen Themen für die interessierten Öffentlichkeit und eine leichtere Exploration von materiellem gemeinsam mit immateriellem Kulturerbe.

Das H2020-Projekt InTaVia (*In/Tangible Cultural Heritage: Visual Analysis, Communication and Curation*, <https://intavia.eu>) will solche Barrieren reduzieren, materielles und immaterielles Kulturerbe digital zusammenführen und eine synoptische Betrachtung von Leben und Werken der europäischen Kulturgeschichte ermöglichen. Das Konsortium harmonisiert zu diesem Zweck nationale Kulturdatenbestände und entwickelt ein prototypisches Informationsportal für die visuelle Analyse, Kuratierung und Kommunikation von hybriden (i.e. im/materiellen) Kulturdaten auf multiplen Ebenen der Aggregation.

Konzept

Visualisierungen erlauben die Gewinnung von Überblicken und Einsichten in voluminöse und komplexe Datenbestände, indem sie z.B. enthaltene zeitliche, geographische, relationale oder kategoriale Muster, Verteilungen und Zusammenhänge sichtbar machen. Dadurch ergänzen Visualisierungen zur Option der direkten Kontemplation und Interpretation einzelner Objekte distante Blicktechniken auf Sammlungen des kulturellen Erbes und ermöglichen deren offene Exploration – für wissenschaftliche Expert*innen genauso wie für die interessierte Öffentlichkeit (Windhager, Federico et al. 2018).

Der Fokus von Visualisierungen des kulturellen Erbes lag dabei bisher meist auf **materiellen Objekten** in den Feldern von Bildern und Skulpturen (z.B. Glinka et al. 2016), Texten (Jänicke et al. 2017), oder von Dokumenten der performativen Künste (z.B. Khulusi et al. 2020). Durch die Öffnung von musealen Objektdatenbanken und Verknüpfung mit komplementären kulturellen Datenbeständen kann jedoch ein reichhaltigeres Verständnis von Objekten in diversen interpretativen und kontextuellen Horizonten ermöglicht werden (Mayr & Windhager 2020). So kann etwa biographisches Wissen über das Leben von Künstler*innen dabei helfen Muster in einer Visualisierung ihrer Objektsammlungen besser zu verstehen (Mayr et al. 2019).

Auf der **immateriellen** Seite des Kulturerbes fokussiert InTaVia auf **biographische Narrative**, die als lexikalische Texte erfasst und in den letzten Jahren mithilfe von NLP Verfahren in strukturierte Ereignis-Daten überführt und als prosopographische Datenbanken aggregiert wurden (Fokkens et al. 2014; Schlögl & Lejtovicz 2017). Diese Transformation von biographischen Texten in maschinenlesbare Formate eröffnet auch neue Möglichkeiten für deren visuelle Repräsentation (Windhager, Schlögl et al. 2018).

Über Personenreferenzen (z.B. die gemeinsamen Normdatei GND) lassen sich materielle Objektdaten (z.B. aus Europeana, aber auch aus lokalen Datenbanken einzelner Kultureinrichtungen) mit dem Wissen über “immaterielle” biographische Ereignisketten und die darüber verknüpften Akteure (z.B. Ersteller, Auftraggeber, Besitzer, oder assoziierte Organisationen) verknüpfen. Die Entwicklung von synoptischen Visualisierungsmethoden, mit denen Datenkonstellationen von “Leben und Werk” neu analysiert und kommuniziert werden können, bildet ein Hauptziel des Projekts InTaVia. Dass hiermit die klassische Grundidee der “biographischen Kritik” und Interpretation (vgl. Vasari 1550/2008; oder Johnson 1781/1868) relativ spät ihrer ersten digitalen und visuellen Remedialisierung zugeführt wird, schmälert leider keine der assoziierten informationstechnischen (Figure 1) und gestalterischen Herausforderungen (Windhager 2020), insbesondere unter

den Vorzeichen einer benutzerzentrierten Designansatzes (Mayr et al. 2018).

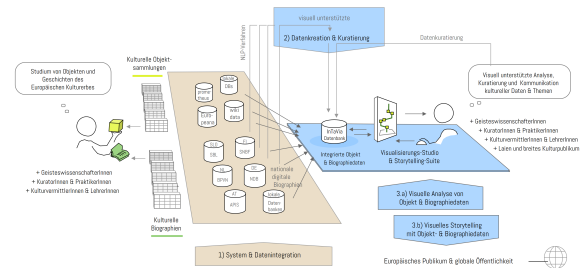


Abb. 1: Architektur der InTaVia-Plattform

Partizipative und prototypische Entwicklungen

In zwei partizipativen Design-Workshops mit 41 internationalen Expert*innen aus verschiedenen Feldern des Kulturerbes – mit wissenschaftlichem, genauso wie mit praktischem Hintergrund – wurde die konzeptuelle Architektur der geplanten InTaVia Informationsplattform diskutiert und auf ihre Passung mit den Bedürfnissen und Prozessen der potenziellen Nutzer*innen hin evaluiert. Aus den reichhaltigen Ergebnissen dieses Prozesses, die die Grundlage für die weiteren technischen Entwicklungen in InTaVia bilden, waren drei von besonderer Relevanz.

Trotz der bereits initial geplanten reichhaltigen Datenbasis – welche die über 50 Mio. Objekte der Europeana mit Beständen von vier nationalen Biographiedatenbanken aus Österreich, Slowenien, den Niederlanden und Finnland verknüpft – formulierten viele Workshopteilnehmer*innen den Bedarf nach Optionen der Einbringung und Einbindung von eigenen Objekt- oder Biographie-Datenbeständen. Ebenso wurden Befürchtungen dokumentiert, dass Fehler in den probabilistisch extrahierten Daten ebenso wie widersprüchliche Einträge in verschiedenen Datenbanken (z.B. weil eine Person in mehreren Biographieportalen mit unterschiedlichen Daten zu finden ist) zu Verzerrungen in der Datenbasis und deren Visualisierung führen könnten.

Daraus wurde die Notwendigkeit einer benutzerseitigen **Datenkuratierung** abgeleitet, die den künftigen Benutzer*innen die Möglichkeit gibt, eigene Daten einzuspeisen und mittels NLP zu strukturieren, die Ergebnisse von NLP-Prozessen zu korrigieren, fehlende Verknüpfungen zwischen Daten herzustellen oder fehlerhafte Verknüpfungen zu korrigieren und bestehende Quellenkonflikte aufzulösen oder sie als solche auch im Rahmen der Visualisierungen explizit sichtbar zu machen.

In den präsentierten Möglichkeiten der visuellen Analyse und Kommunikation sehen die Teilnehmer*innen ein hohes Potenzial für ihre berufliche Arbeit mit kulturellen Objekten und Akteuren. Betont wurde auch der Bedarf nach einem flexiblen Tool, das sich an aktuelle Daten, Themen und Fragestellungen anpassen lässt, und in dem Nutzer*innen verschiedene Ebenen und Informationen nach Bedarf ein- und ausblenden können, um etwa die thematische und visuelle Komplexität an diverse Zielgruppen der Kulturvermittlung anpassen zu können.

Visualisierungen werden im Projekt an mehreren Stellen eine Rolle spielen: (1) Zunächst sollen Benutzer*innen ohne technisches Vorwissen beim Verständnis der NLP Prozesse und bei der Datenkuratierung durch Visualisierungen unterstützt werden (z.B. indem unterschiedliche extrahierte Entitäten mit ihren Wahrscheinlichkeiten visualisiert werden, anstatt ausschließlich der wahrscheinlichsten Ergebnisse). (2) Interaktive Visualisierung erlauben die Exploration und Analyse der Daten in verschiedenen Kombinationen (Objekte / Objekte und Akteure / Akteure) und Aggregationszuständen (von einzelnen Individuen bis hin zu verknüpften Datenkonstellationen wie Institutionen oder Regionen) entlang verschiedener Datendimensionen (wie Zeit, Raum, Relationen oder Mengen, vgl. Figure. 2).

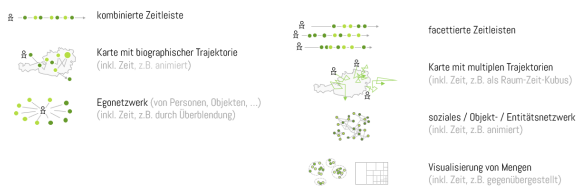


Abb. 2: Verschiedene Optionen, um Objekte (hellgrün) synoptisch mit Biographiedaten (dunkelgrün) zu visualisieren: von Einzelakteuren (links) hin zu Personengruppen (rechts)

Für jede dieser Datendimensionen sind in der Folge diverse Aggregationen denkbar. So können Informationen von einzelnen Individuen und Objekten ausgehend zu signifikant größeren Datenkonstellationen aggregiert werden - wie zum Beispiel Gruppen, kulturellen Institutionen oder raum-zeitlichen Regionen (Figure 3, unten). Der essentiellen Bedeutung der zeitlichen Orientierung dieser Daten wird zudem durch multiple Enkodierungen Rechnung getragen, um Nutzer*innen verschiedene analytische Perspektiven mit komplementären Stärken und Schwächen zu offerieren.

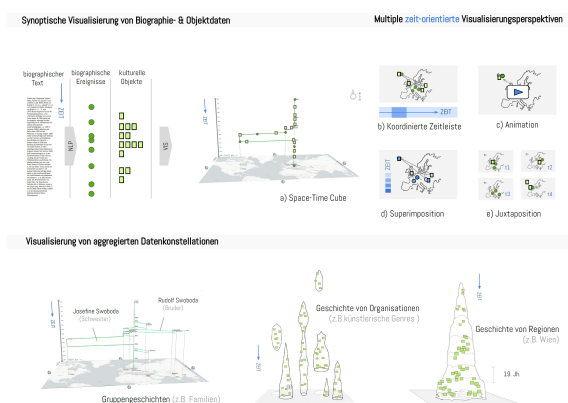


Abb. 3: Verschiedene Optionen der Visualisierung von Zeit (oben) und der Aggregation von Objekt- und Biographiedaten in der geographischen Raumzeit (unten).

(3) Schließlich können auch *narrative Visualisierungen* die Kommunikation der Ergebnisse von solchen visuellen Analysen unterstützen: Dazu werden in einem Editor narrative Techniken mit Visualisierungen kombiniert und mit audiovisuellem Mate-

rial und Texten angereichert. Durch die so ermöglichte Kombination von narrativ-sequentieller Präsentation der Informationen und freier Exploration der Visualisierung, können diese narrativen Visualisierungen auch der interessierten Öffentlichkeit einen Zugang zu der geplanten Datenbasis eröffnen.

Von den Berufsprofilen und Bedürfnissen der Teilnehmer*innen wurden in einem iterativen Prozess 10 **Personas** als prototypische Benutzer*innen von DH Expert*innen bis hin zu interessierten Laien abgeleitet, die Richtlinien für die weitere technischen Entwicklung in InTaVia enthalten: Vom fachlichen und technischen Vorwissen über zu erwartende Nutzungs- und Aktivitätsprofile sowie wichtige Einschränkungen (z.B. rechtliche Bedenken, aber auch bestimmte technische Bedürfnisse) sollen diese Personas sicherstellen, dass die zukünftigen Benutzer auch zwischen den geplanten Evaluationszyklen in die Entwicklung der DH-Technologien einfließen. In die Ausarbeitung der entsprechenden Profile flossen neben der Teilnehmer*innen-Analyse auch User Stories der NFDI4Culture-Initiative (2019) ein, ebenso wie Lernerfahrungen aus gescheiterten DH Projekten (Dombrowski 2014, 2019).

Ausblick

Das Projekt InTaVia befindet sich noch in einem frühen Stadium, dennoch erlauben die hier präsentierten Workshopergebnisse bereits die Ableitung von Nutzer*innenbedürfnissen und Anforderungen an DH Technologien, die auch für andere Projekte mit heterogener Nutzer*innen-Basis in Forschungs- und Arbeitsfeldern zu verlinkten Daten von Nutzen sein können. Die entwickelten Personas wurden auf der Projekt-Website zugänglich gemacht und stehen auch anderen Wissenschaftler*innen zur Verfügung.

Die Zusammenführung von Biographie- und Objektdaten in einem synoptischen Wissensgraphen und in einem korrespondierenden Visualisierungssystem, ermöglicht es, auch über alternative Formen der Objektmodellierung und -visualisierung nachzudenken: Bisher wurden digitale kulturelle Objekte oft nur mit statischen Metadaten verknüpft, jedoch lassen sich diese in InTaVia auch als geschichtliche Objekte mit veränderlichen Eigenschaften (z.B. dynamischen Klassifikationen, Standorten, oder Personenrelationen) modellieren und entsprechende Objektbiographien visualisieren. Eine solche Modellierung erhöht zwar die Anforderungen an die vorhandenen Digitalisierungsprozesse in Kulturinstitutionen, andererseits eröffnet sie auch neue Möglichkeiten für die DH-gestützte Forschung beispielsweise zur Provenienz von Objekten.

Ein weiterer Punkt von zentraler Relevanz in einem DH-Projekt ist die Frage der Datenqualität und von Unsicherheiten in den Daten als Herausforderung für die Modellierung und Visualisierung von kulturellen Daten (Windhager et al. 2019). Der Austausch im Rahmen von verschiedenen Workshops innerhalb des Konsortiums und mit den potenziellen Nutzer*innen zeigte, dass die Qualität der Daten, die Vielfalt der Einflüsse auf selbige und die Transparenz von Unsicherheiten in den Daten wichtige Faktoren für das Vertrauen der Benutzer*innen in die entwickelten Technologien sind. So ist es unerlässlich, dass im Projekt, zumindest jene Einflüsse auf die Datenqualität, die den gewählten Prozessen und Methoden geschuldet sind (z.B. Verknüpfung mehrerer Datenbanken, Anwendung von NLP Prozessen) auch transparent kommuniziert und visualisiert werden.

Danksagung. Das Projekt InTaVia wird von der Europäischen Kommission im Rahmen des H2020 Research and Innovation Pro-

gramme, Grant Agreement No. 101004825 gefördert. Wir möchten uns für die Unterstützung durch das Projektkonsortium (Vrije Universiteit Amsterdam, NL, Research Centre of the Slovenian Academy of Sciences & Arts, SI, Aalto University, FI, University of Southern Denmark, DK, Austrian Academy of Sciences, AT, University of Stuttgart, DE, Fluxguide, AT, University of Helsinki, FI) und die Teilnehmer*innen der Co-Design-Workshops herzlich bedanken.

Bibliographie

Dombrowski, Q. (2014): "What Ever Happened to Project Bamboo?" in: *Literary & Linguistic Computing* 29: 326–339. <https://doi.org/10.1093/lilc/fqu026>

Dombrowski, Q. (2019): *Towards a Taxonomy of Failure*. <http://quinndombrowski.com/?q=blog/2019/01/30/towards-taxonomy-failure>

Fokkens, A. / Ter Braake, S. / Ockeloen, N. / Vossen, P. / Legêne, S. / Schreiber, G. (2014): „BiographyNet: Methodological Issues when NLP supports historical research.“ in: *LREC*, 3728–3735.

Glinka, K. / Pietsch, C. / Dilba, C. / Dörk, M. (2016): "Linking structure, texture and context in a visualization of historical drawings by Frederick William IV (1795–1861)." in: *International Journal for Digital Art History*, (2). <https://doi.org/10.11588/dah.2016.2.33530>

Jänicke, S. / Franzini, G. / Cheema, M. F. / Scheuermann, G. (2017): „Visual text analysis in digital humanities.“ in: *Computer Graphics Forum* 36/6: 226–250.

Johnson, S. (1868): *Lives of the Most Eminent English Poets: With Critical Observations on Their Works*. (Neuaufgabe, Original veröffentlicht um 1781). AT Crocker.

Khulusi, R. / Kusnick, J. / Meinecke, C. / Gillmann, C. / Focht, J. / Jänicke, S. (2020). A survey on visualizations for musical data. In: *Computer Graphics Forum* 39/6: 82–110.

Mayr, E. / Salisu, S. / Filipov, V. / Schreder, G. / Leite, R. / Miksch, S. / Windhager, F. (2019): „Visualizing biographical trajectories by historical artifacts: A case study based on the photography collection of Charles W. Cushman.“ in: *Workshop on Biographical Data in a Digital World 2019*, Varna, Bulgaria.

Mayr, E. / Schreder, G. / Windhager, F. (2018): „Digital Humanities - Eine benutzerzentrierte Perspektive.“ in: *Digital Humanities im deutschsprachigen Raum* (DHD) 2018, Köln.

Mayr, E. / Windhager, F. (2020): „Vor welchem Hintergrund und mit Bezug auf was? Zur polykontextualen Visualisierung kultureller Sammlungen.“ In: U. Andraschke / S. Wagner, (Hrsg.): *Objekte im Netz. Wissenschaftliche Sammlungen im digitalen Wandel*, Bielefeld: transcript, 235–245.

NFDI4Culture (2019): *NFDI4Culture User Stories*. https://nfdi4culture.de/fileadmin/files/NFDI4Culture_UserStoryAll.pdf

Schlögl, M. / Lejtovicz, K. (2017): "APIS-Eine Linked Open Data basierte Datamining-Webapplikation für das Auswerten biographischer Daten.“ In: *DHD 2017, Book of Abstracts*, Bern. doi: 10.5281/zenodo.3684825

Vasari, G. (2008): *The Lives of the Artists* (Neuaufgabe, Original veröffentlicht um 1550). Oxford, UK: Oxford University Press.

Windhager, F. (2020): *A Synoptic Visualization Framework for Artwork Collection Data and Artist Biographies*. Doctoral Thesis, Universität Wien. <http://othes.univie.ac.at/65279/1/67771.pdf>

Windhager, F. / Federico, P. / Schreder, G. / Glinka, K. / Dörk, M. / Miksch, S. / Mayr, E. (2018): „Visualization of cultural heritage collection data: State of the art and future challenges.“ In: *IEEE Transactions on Visualization & Computer Graphics* 25: 2311–2330.

Windhager, F. / Salisu, S. / Mayr, E. (2019): "Exhibiting uncertainty: Visualizing data quality indicators for cultural collections. in: *Informatics*. Special Issue on Uncertainty in Digital Humanities, 6: <https://doi.org/10.3390/informatics6030029>

Windhager, F. / Schlögl, M. / Kaiser, M. / Bernad, Á. / Salisu, S. / Mayr, E. (2018). „Beyond one-dimensional portraits: A synoptic approach to the visual analysis of biography data.“ in: A. Fokkens / S. ter Braake / R. Sluijter / P. Arthur / E. Wandl-Vogt (Eds.): *Proceedings of BD-2017 - Biographical Data in a Digital World*. Linz: CEUR. <http://ceur-ws.org/Vol-2119/paper11.pdf>

Linked Open Data für die Literaturgeschichtsschreibung Das Projekt "Mining and Modeling Text"

Hinzmann, Maria

hinzmannm@uni-trier.de
Universität Trier, Germany

Schöch, Christof

schoech@uni-trier.de
Universität Trier, Germany

Dietz, Katharina

dietz@uni-trier.de
Universität Trier, Germany

Klee, Anne

klee@uni-trier.de
Universität Trier, Germany

Erler-Fridgen, Katharina

erler@uni-trier.de
Universität Trier, Germany

Röttgermann, Julia

roettger@uni-trier.de
Universität Trier, Germany

Steffes, Moritz

steffesm@uni-trier.de
Universität Trier, Germany

Zielsetzung und Projektstruktur

Im Umgang mit dem stetig wachsenden „digitalen Kulturerbe“ bietet die Weiterentwicklung der systematischen Datenerhebung und Wissensrepräsentation bisher nicht ausgeschöpfte Po-

tentiale für die Literaturgeschichtsschreibung. Vor diesem Hintergrund werden im Projekt *Mining and Modeling Text (MiMoText)* quantitative Methoden der Informationsextraktion („Mining“) und Datenmodellierung („Modeling“) ineinander verschränkt, um ein literaturgeschichtliches Wissensnetzwerk aufzubauen (vgl. Abb. 1).¹ Der Fokus liegt zunächst auf französischen Romanen (1751–1800). Die Übertragbarkeit in andere Domänen wird berücksichtigt. Zentrales Anliegen ist es, den Bereich der quantitativen Methoden zur Extraktion, Modellierung und Analyse geisteswissenschaftlich relevanter Informationen aus umfangreichen Textsammlungen weiterzuentwickeln und aus interdisziplinärer (geistes-, informatik- und rechtswissenschaftlicher) Perspektive zu erforschen.

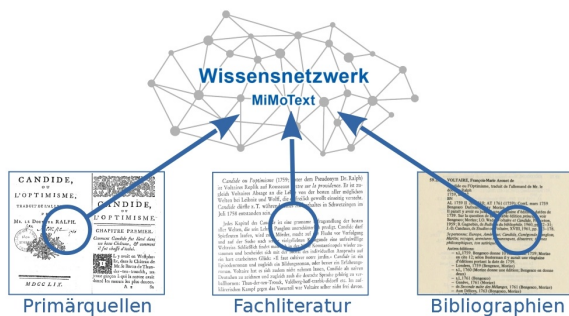


Abb. 1: Informationsquellen des Wissensnetzwerks.

Bezogen auf den Gegenstandsbereich ist ein Ausgangspunkt, dass die über rund zwei Jahrhunderte akkumulierten literaturhistorischen Forschungserkenntnisse größtenteils nicht unmittelbar nutzbar sind, da diese sehr umfangreich sind, nicht digital vorliegen oder auf unterschiedliche Orte und Quellen verteilt und in unterschiedlichen Sprachen publiziert sind. *MiMoText* begegnet diesem Desiderat, indem es Informationen aus drei unterschiedlichen Quellentypen im Aufbau des fachspezifischen Wissensnetzwerks miteinander verknüpft: Metadaten aus Nachweissystemen, Texteigenschaften aus Primärtexten, Sachinformationen aus Fachliteratur. In vier Teilbereichen (Research Areas/RAs) werden Lösungen für die Informationsextraktion, ihre Modellierung, die rechtlichen Rahmenbedingungen sowie die Infrastruktur erarbeitet (vgl. Abb. 2).

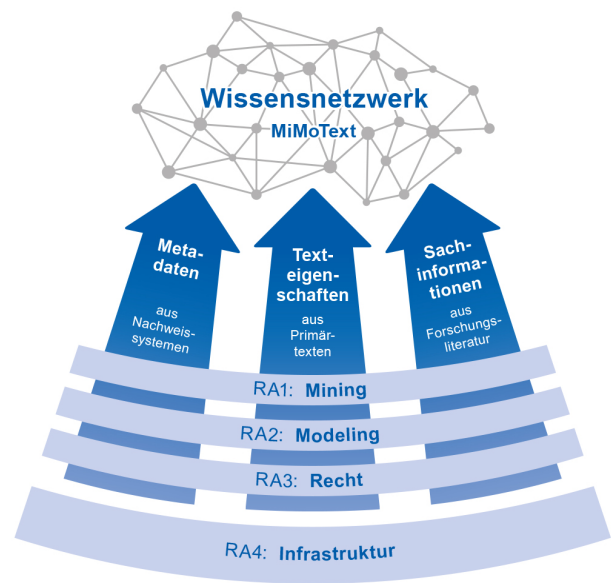


Abb. 2: Projektübersichtsgrafik.

Open Science-Prinzipien

Für die Bereitstellung der Daten sowie die Arbeit und Infrastruktur im Projekt sind Open Science-Prinzipien tragend. Dies betrifft u.a. die Veröffentlichung FAIRer Daten (Röttgermann/Schöch 2020) im Open Access (Schöch 2021), die Nutzung von Open Source-Tools wie *INCEPTION* (Klie et al. 2018) und *OCR4all* (Reul et al. 2019) sowie Wikibase als Infrastruktur, die dem Linked Open Data (LOD) folgt.² Im rechtswissenschaftlichen Teilbereich werden die rechtlichen Rahmenbedingungen von Text und Data-Mining in den Geisteswissenschaften auch im Hinblick auf eine offenere und nachhaltigere Nutzung von Forschungsdaten erforscht (vgl. Erler-Fridgen 2021a; Erler-Fridgen 2021b; Erler-Fridgen 2021c; Raue/Schöch 2020; Schöch et al. 2020).

Mining – Extrahieren von Informationen aus drei Quellentypen

Bibliographische Daten

Für unsere Domäne ist die 1977 von Mylne, Martin und Frautschi veröffentlichte *Bibliographie du genre romanesque français 1751-1800* (BGRF) zentral, da sie die Grundgesamtheit der Romane definiert. Die BGRF wurde von Andreas Lüschoff aufwendig erschlossen und nach aktuellen bibliographischen Standards modelliert (Lüschoff 2020).

Primärliteratur

Im Teilprojekt zur Primärliteratur wird schrittweise ein Korpus von etwa 200 Romanen aufgebaut, wovon bereits reichlich 100 Texte in XML-TEI verfügbar sind (vgl. Röttgermann 2021; Klee/

Röttgermann 2020).³ Eine Reihe von Analyseverfahren wurde bereits auf diesen Textbestand angewandt, darunter Topic Modeling (vgl. Klee/Röttgermann 2020), NER (Orte und Figuren) sowie explorativ Sentiment Analyse.⁴

Sekundärliteratur

Das mittelfristige Ziel besteht darin, durch überwachtetes Lernen die automatische Extraktion von Aussagen (RDF-Tripel) zu ermöglichen. Um Trainingsdaten zu generieren und Tripel in das Wissensnetzwerk einspeisen zu können, werden aktuell literaturgeschichtliche Texte in INCEption annotiert. Die Daten sollen als Statements über eine noch zu entwickelnde toolübergreifende Pipeline in unsere projektspezifische Wikibase importiert werden.⁵

Modeling – Repräsentation und Vernetzung von Wissen

Im Aufbau des LOD-Wissensnetzwerks werden literaturgeschichtliche Aussagetypen in einer systematischen Ontologie modelliert und die extrahierten Informationen als RDF-Tripel repräsentiert (vgl. Abb. 3). Der Mehrwert des Netzwerks wird durch exemplarische Nutzungsszenarien in Form von SPARQL-Abfragen konkretisiert. Frageoptionen wie „Tritt Thema x in einem bestimmten Zeitraum y gehäuft auf?“ verdeutlichen, welcher Nutzen daraus für eine datenbasierte Literaturgeschichtsschreibung entstehen kann. Exemplarisch werden Einblicke in die Infrastruktur gegeben (vgl. Abb. 4).

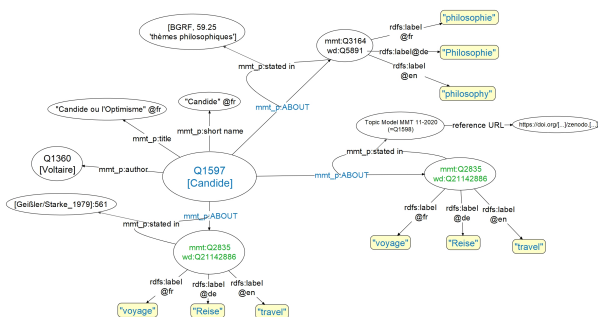


Abb. 3: Thematische Aussagen (Property „about“) über das Werk *Candide* aus den drei unterschiedlichen Informationsquellen.

Domänenspezifische Herausforderungen & Chancen

Ein Standardisierungsprozess wie er sich beispielsweise im CIDOC CRM (<http://www.cidoc-crm.org>) niederschlägt, steht für die Domäne der Literaturgeschichte noch am Anfang. Das Projekt konzentriert sich auf Aussagen über literarische Werke und Autor:innen, bisher vor allem thematische Aussagen (vgl. Schöch et al. 2022) sowie Handlungsorte (vgl. Hinzmann et al. angenommen). Dabei ist das beständig wachsende Wissensnetzwerk in der Zusammenführung von Informationen aus verschiedenen, auch wi-

dersprüchlichen Quellen möglichst offen im Hinblick auf unterschiedliche Nutzungsszenarien und Fragerichtungen. Durch die Menge der aggregierten Daten lassen sich literaturhistorische Annahmen bestätigen, revidieren oder präzisieren und neue Fragestellungen sowie eine Metaperspektive auf den literaturwissenschaftlichen Diskurs entwickeln.

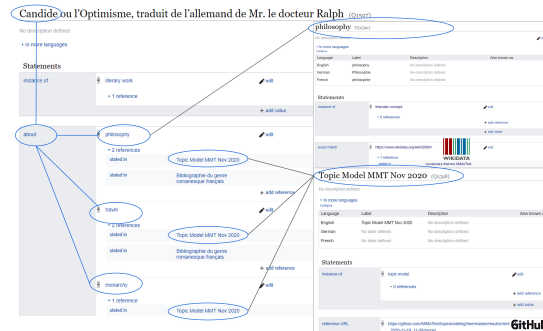


Abb. 4: Screenshot von thematischen Aussagen zu *Candide* in der MiMo-Text-Wikibase.

Präsentationsstrategie

Das Poster veranschaulicht die Integration der verschiedenen Informationsquellen in der Datenmodellierung sowie das Zusammenspiel der verschiedenen Teilprojekte und Tools im Aufbau und in möglichen Nutzungsszenarien des mehrsprachigen Wissensnetzwerks.

Fußnoten

1. Das Projekt wird im Rahmen der Forschungsinitiative Rheinland-Pfalz gefördert und vom Trier Center for Digital Humanities koordiniert, vgl. <https://mimotext.uni-trier.de>.
2. Vgl. zum Datenmodell von Wikibase/Wikidata <https://www.mediawiki.org/wiki/Wikibase/DataModel> sowie https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format.
3. Vgl. genauer zum Korpusaufbau Röttgermann (angenommen) und zu den Kodierungsprinzipien, die der European Literary Text Collection (ELTeC) folgen: Burnard et al. 2021.
4. Vgl. das GitHub-Repository unter: <https://github.com/MiMo-Text/roman18>.
5. Vgl. zum Statement-Begriff in Wikibase/Wikidata https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format sowie <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer#Statements>.

Bibliographie

- Burnard, Lou / Schöch, Christof / Odebrecht, Carolin (2021): "In Search of Comity: TEI for Distant Reading", in: *Journal of the Text Encoding Initiative* 14 <https://doi.org/10.4000/jtei.3500>.
- Erler-Fridgen, Katharina (2021a): "Die Nutzung wissenschaftlicher Ausgaben für Textanalysen", in: *IRDT PaperSeries*

1. <https://irdt.uni-trier.de/die-nutzung-wissenschaftlicher-ausgaben-fuer-textanalysen/> [letzter Zugriff 30. November 2021].

Erler-Fridgen, Katharina (2021b): "Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken", in: *IRDT PaperSeries* 2. <https://irdt.uni-trier.de/kriterien-der-urheberrechtlichen-schutzfaehigkeit-von-texten-und-sammelwerken/> [letzter Zugriff 30. November 2021].

Erler-Fridgen, Katharina (2021c): "Die Präsentation von Textteilen als Ergänzung von Textanalysen", in: *IRDT PaperSeries* 3. <https://irdt.uni-trier.de/die-praesentation-von-textteilen-als-ergaenzung-von-textanalysen/> [letzter Zugriff 30. November 2021].

Hinzmann, Maria / Röttgermann, Julia / Klee, Anne / Steffes, Moritz / Schöch, Christof (angenommen): "The French Enlightenment Novel as a Graph? Potentials and Challenges in the Construction of a Knowledge Network", angenommener Beitrag *Graphs and Networks in the Humanities 2022: Knowledge Graphs and Reasoning – Promises, Potentials, and Pitfalls*.

Klee, Anne / Röttgermann, Julia (2020): *Doing Topic Modeling on French 18th Century Novels in the Context of MiMoText Project* [data set] <https://github.com/MiMoText/topicmodeling> [letzter Zugriff 30. November 2021].

Klie, Jan-Christoph / Bugert, Michael / Boulosa, Beto / Eckart de Castilho, Richard / Gurevych, Iryna (2018): "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation", in: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* 5–9 <http://tubiblio.ulb.tu-darmstadt.de/106270/> [letzter Zugriff 30. November 2021].

Lüschow, Andreas (2020): "Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane", in: *Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, 80–84 10.5281/zenodo.3666690.

Raue, Benjamin / Schöch, Christof (2020): "Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate – Versöhnung von Urheberrecht und textbasierter Forschung", in: *RuZ – Recht und Zugang*, 1.2.: 118–27 10.5771/2699-1284-2020-2-118.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): "OCR4all -- An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings", in: *ArXiv:1909.04032* [Cs].

Röttgermann, Julia / Schöch, Christof (2020): "FAIRe Daten in den Literaturwissenschaften? Das Beispiel „Mining and Modeling Text“ und der französische Roman des 18. Jahrhunderts", in: *Romanistik-Blog. Blog des Fachinformationsdienstes* <https://blog.fid-romanistik.de/2020/11/05/faire-daten-in-den-literaturwissenschaften/> [letzter Zugriff 30. November 2021].

Röttgermann, Julia (ed.) (2021): *Collection de romans français du dix-huitième siècle (1750-1800) / Eighteenth-Century French Novels (1750-1800)* [data set]. Release v0.2.0 10.5281/zenodo.5040855.

Röttgermann, Julia (angenommen): "Établissement d'un corpus de romans français du XVIIIe siècle dans le cadre du projet Mining and Modeling Text" [angenommener Beitrag Frankoromanistentag 2021, Sektion: *Digital, global, transdisziplinär: Impulse für eine transdisziplinäre digitale Romanistik*].

Schöch, Christof / Döhl, Frédéric / Rettinger, Achim / Gius, Evelyn / Trilcke, Peer / Leinen, Peter / Jannidis, Fotis / Hinzmann, Maria / Röpke, Jörg (2020): "Abgeleitete Textformate:

Text und Data Mining mit urheberrechtlich geschützten Textbeständen", in: *Zeitschrift für digitale Geisteswissenschaften – ZfdG* 10.17175/2020_006.

Schöch, Christof (2021): "Open Access für die Maschinen", in: Kohle, Hubertus / Effinger, Maria (eds.): *Die Zukunft des kunsthistorischen Publizierens*. Heidelberg: arthistoricum.net 79–94 10.11588/arthistoricum.663.c9210.

Schöch, Christof / Hinzmann, Maria / Röttgermann, Julia / Dietz, Katharina / Klee, Anne (angenommen): "Smart Modeling For Literary History", angenommen in: *IJHAC. Linked Open Data in the Arts and Humanities* [special issue] (March 2022).

Linked Open Tafsir Rekonstruktion der Entstehungsdynamik(en) des Korans mithilfe der Netzwerkmodellierung früher islamischer Überlieferungen

Ahmed, Sajawel

sahmed@em.uni-frankfurt.de

Goethe-Universität Frankfurt am Main, Deutschland

Rehman, Misbahur

rehman@em.uni-frankfurt.de

Goethe-Universität Frankfurt am Main, Deutschland

Tischlik, Joshua

tischlik@em.uni-frankfurt.de

Goethe-Universität Frankfurt am Main, Deutschland

Kruse, Carl

ca.kruse@em.uni-frankfurt.de

Goethe-Universität Frankfurt am Main, Deutschland

Mahmutovic, Edin

mahmutovic@em.uni-frankfurt.de

Goethe-Universität Frankfurt am Main, Deutschland

Özsoy, Ömer

oezsoy@em.uni-frankfurt.de

Goethe-Universität Frankfurt am Main, Deutschland

Das Projekt *Linked Open Tafsir*¹ hat die Erstellung einer online abrufbaren Datenbank früher exegetischer Überlieferungsmaterialien auf Basis des Kommentarwerks von At-Tabari (gest. 310 n. H. / 923 n. Chr.) zum Ziel. Sein Werk *Jami' al-bayan 'an ta'wil ay al-Qur'an* (kurz: Tafsir At-Tabari) kann nach heutigem Kenntnisstand als Sammlung des Großteils aller zu Anfang des 4./10. Jahrhunderts vorliegenden exegetisch relevanten Überlieferungen gelten. In der Datenbank werden die in diesen Überlieferungen enthaltenen Informationen zu historischen Begebenheiten zur Offenbarungszeit sowie den kulturellen, religiösen, sozialen

und sprachlichen Rahmenbedingungen der Koranentstehung erfasst. Für die Erfassung der Überlieferungen bzw. Informationen in den Überlieferungen werden mit Hilfe von Künstlicher Intelligenz Daten und Programme insbesondere für das *Named Entity Recognition* entwickelt, die den Erfassungsprozess signifikant beschleunigen. Das Projekt soll eine solide Forschungsgrundlage für Zugänge zur Reflexion der Offenbarungsdynamik(en) des Korans in der frühen Exegese schaffen. Die Datenbank wird insofern "offen" sein, als dass zu einem späteren Zeitpunkt weitere exegetische Kompilationen sowie frühere Überlieferungswerke hinzugefügt werden können. Weiterhin werden die beteiligten WissenschaftlerInnen diese digitalen Zugänge in ihren Forschungsprojekten im Hinblick auf das Islamische Recht, die Systematische Theologie, die Hadithwissenschaft, die Tafsirgeschichte, die Religionspädagogik und dem Maschinellen Lernen für klassische arabische Literatur reflektieren.

Datenbank: Die Erstellung der Datenbank erfolgt in mehreren Schritten: Zunächst sammeln die Projektbeteiligten bereits digital verfügbare Überlieferungen des Kommentarwerks von At-Tabari, um daraus ein erstes digitales Textkorpus zu erstellen. Im Anschluss wird eine Datenbank aufgebaut, die alle in den exegetischen Überlieferungen erhaltenen Informationen über das Offenbarungsumfeld (Mikro-, Makro- und Sprachumfeld), Lesarten (Qira'at), intratextuelle Zusammenhänge (Wiederholung, Querreferenz, Abrogation, Spezifikation etc.) und insbesondere *Named Entities* (NEs: Ort, Zeit, Person etc.) als solche erfasst, markiert, vernetzt und auffindbar macht. Ein weiterer Arbeitsschwerpunkt liegt in der Erschließung der Überlieferungsketten (Isnade) und einzelner TradentInnen. Die Datenbank soll entsprechend verschiedener Forschungsinteressen genutzt werden können: So werden sich beispielsweise alle exegetischen Überlieferungen aus dem Werk Tafsir At-Tabari, welche verschiedene Koranverse mit einer bestimmten Person in Verbindung bringen, in einem einzigen Suchvorgang auffinden lassen. Ebenso werden sich durch die Datenbank unmittelbar Teilkorpora anzeigen lassen, die über dieselben TradentInnen überliefert worden sind. Es ist geplant, die Funktionalität der Datenbank sowie exemplarische Suchmöglichkeiten in Form von YouTube-Tutorials² vorzustellen.

Annotation von Named Entities: Unter Named Entity Recognition (NER) versteht man die computerlinguistische Aufgabe, Eigennamen (NE) in Texten zu erkennen (z.B. Mekka, Asien, Tabari, Shia). Solche Eigennamen stehen im Kontrast zu Gattungsnamen, welche eine Klasse von Eigennamen umfassen (z.B. Stadt, Kontinent, Person, Organisation). Technisch gesehen sind für NER zwei Schritte notwendig: Zuerst müssen in einem laufenden Text die Inhaltselemente gefunden werden, die zu einem Eigennamen gehören, danach können diese Eigennamen semantischen Kategorien zugeordnet werden. In unserer aktuellen Annotationsarbeit haben wir aufbauend auf *Guidelines für die Named Entity Recognition* (Benikova 2014; Ahmed 2019) fünf semantische Hauptklassen für klassische arabische Texte unterschieden (Personen, Organisationen, Orte, Zeiten und Andere).

Zurzeit annotiert unser Team aus Arabisten den arabischsprachigen Text, welcher durch einen OCR-Prozess auf historische Manuskripte der Koranexegese von At-Tabari generiert wurde und uns im XML-basierten TEI-Format vorliegt. Auf Basis dieser digitalen Textsammlung werden von unseren Annotatoren die einzelnen Eigennamen mithilfe des Tools *Oxygen XML Editor*³ identifiziert, markiert und als NE im TEI-Format abgespeichert. In der Fig. 1 bekommen wir einen Einblick in der Annotationsumgebung. Wir sehen, dass der Oxygen XML Editor in der Lage ist, den arabischen *Right-to-Left*-Text korrekt darzustellen, die markierten NEs farbig hervorzuheben und insgesamt den Annotatoren einen einfacheren Zugang zum Quelltext zu gewähren.



Abb. 1: Ausschnitt aus dem *Oxygen XML Editor* für die Annotation von *Named Entities* im klassischen arabischen Quelltext *Tafsir At-Tabari*.

Aktueller Stand der Annotationsarbeit und Ausblick: Aktuell haben wir bereits über 30.000 Sätze mit solchen NEs annotieren können, hinzu wird eine weitere solche Menge in der zweiten Phase der Annotation kommen. Zum Abschluss dieser Arbeit werden wir der DH-Fachcommunity den ersten Datensatz überhaupt für die klassische arabische Sprache von solcher Art und Dimension als Open-Source Ressource (Lizenz: CC-BY-4.0) auf *GitHub* bereitstellen, welcher ein ideales Fundament für weiterführende Anwendungen von Sprachmodellen wie *Word2vec* (Mikolov 2013), *LSTM* (Lample 2016; Ahmed 2018), *BERT* (Devlin 2019) sein wird. Begleitend hierzu werden wir eine erste Datenanalyse mit eben diesen Sprachmodellen durchführen und über die Ergebnisse (Precision, Recall und F1-Score) berichten. Unsere innovative, interdisziplinäre Annotationsarbeit legt damit die ersten Bausteine für die Analyse klassischer arabischer Texte mit modernen Verfahren des Maschinellen Lernens, so dass auch der Bereich der Islamwissenschaft und historischen Theologie von der Digitalisierungswelle profitieren kann.

Fußnoten

1. www.linkedopentafsir.de/
2. www.youtube.com/watch?v=LJODcc_Gz50
3. www.oxygenxml.com/

Bibliographie

- Ahmed, S. and Mehler, A. (2018): "Resource-size matters: Improving neural named entity recognition with optimized large corpora", in: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 919-924).
- Ahmed, S., Stoeckel, M., Driller, C., Pachzelt, A. and Mehler, A. (2019): "BIOfid Dataset: Publishing a German gold standard for named entity recognition in historical biodiversity literature", in: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 871-880).
- Benikova, D., Biemann, C. and Reznicek, M. (2014): "NoStAD Named Entity Annotation for German: Guidelines and Dataset", in: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)* (pp. 2524-2531).
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding", in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HTL)* (pp. 4171-4186).
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016): "Neural architectures for named entity recognition", in: *Proceedings of the 2016 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HTL) (pp. 260-270).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013): "Distributed representations of words and phrases and their compositionality", in: *Advances in neural information processing systems* (NIPS) (pp. 3111-3119).

Mediatheken der Darstellenden Kunst digital vernetzen

Illmayer, Klaus

klaus.illmayer@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austria

Tiefenbacher, Sara

S.Tiefenbacher@ub.uni-frankfurt.de
Universitätsbibliothek Johann Christian Senckenberg

Voß, Franziska

F.Voss@ub.uni-frankfurt.de
Universitätsbibliothek Johann Christian Senckenberg

Beck, Julia

J.Beck@ub.uni-frankfurt.de
Universitätsbibliothek Johann Christian Senckenberg

Henninger, Christine

c.henniger@iti-germany.de
Internationales Theater Institut Deutschland, Berlin / Mediathek für Tanz und Theater

Wittenbecher, Maxim

m.wittenbecher@iti-germany.de
Internationales Theater Institut Deutschland, Berlin / Mediathek für Tanz und Theater

Für die Tanz- und Theaterwissenschaft sind audiovisuelle (AV) Aufzeichnungen von Tanz- und Theateraufführung von großer Bedeutung. Sind doch AV-Materialien neben der archivalischen Dokumentation und Augenzeug*innenberichten eine der wichtigsten Quellen für die nachträgliche Auseinandersetzung mit dem flüchtigen Ereignis einer Tanz- und Theateraufführung (vgl. bspw. Fischer-Lichte 1999, S. 11). Durch die steigende Verfügbarkeit und der damit einhergehenden Kostenreduktion von Aufzeichnungsmedien – insbesondere VHS-Kassetten – wurde ab den 1980er Jahren an theaterwissenschaftlichen Instituten im deutschsprachigen Raum begonnen, systematisch solche Aufzeichnungen zu sammeln und damit Mediatheken für die Theaterforschung zu begründen (beispielhaft Fuxjäger 2020). Zugleich erstellen auch Theaterhäuser Aufzeichnungen sowie Künstler*innen und Theatergruppen zur Selbstdokumentation. Eine bedeutende Rolle nehmen zudem die öffentlich-rechtlichen Fernsehstationen ein, die

regelmäßig Tanz- und Theateraufführungen ausstrahlen und über umfangreiche Bestände verfügen. Trotz einer großen Bandbreite an Institutionen, die AV-Aufzeichnungen von Tanz- und Theateraufführungen bereithalten, gibt es weder eine einheitliche Systematik noch eine übergreifende Bestandsaufnahme dieser verstreuten Mediathekensammlungen. Prekär ist zudem meist auch die Zugriffsmöglichkeit auf die AV-Materialien.

Das in diesen Mediatheken gebündelte Potential für die Forschung ist der tanz- und theaterwissenschaftlichen Community bewusst, was auch in einer 2018 durchgeführten Umfrage unter Mitgliedern der Gesellschaft für Theaterwissenschaft bestätigt wurde. Darauf aufbauend wurde eine Bestands- und Bedarfsanalyse erstellt, die darauf drängt, die Sammlungen untereinander digital zu vernetzen und Forscher*innen und Künstler*innen einfacher zugänglich zu machen. Zugleich wurde auf die Dringlichkeit hingewiesen, die AV-Aufzeichnungen systematisch zu digitalisieren, da viele Trägermedien durch Mehrfachbenutzung und Materialverschleiß gefährdet sind.

Basierend auf diese Vorarbeiten wurde vom Fachinformationsdienst für Darstellende Kunst (FID DK) gemeinsam mit dem (ITI/MTT) der Projektantrag "Mediatheken der Darstellenden Kunst digital vernetzen" im Bereich (LIS) der DFG gestellt, der genehmigt wurde und im April 2021 gestartet ist. Das Mediatheken-Projektteam möchte im Posterbeitrag für die DHd2022 das Projekt vorstellen und über den bis dahin erfolgten Fortschritt informieren.

Ziel der ersten Phase des Mediatheken-Projekts ist es, eine digitale Infrastruktur und darauf aufbauend einen prototypischen Workflow zu entwickeln, wobei anhand der AV-Aufzeichnungsdatenbanken zweier Projektpartner*innen – des ITI/MTT und der wissenschaftlichen Audiothek und Videothek des Instituts für Theater-, Film- und Medienwissenschaft der Universität Wien – die in unterschiedlichen Formaten vorliegenden Metadaten zusammengeführt und diese auf Überschneidungen und Ergänzungen überprüft werden (z.B. Aufzeichnungen einer Inszenierung an unterschiedlichen Aufführungstagen in Berlin und in Wien). Somit wird eine Basis nicht nur für eine institutionenübergreifende Suche sondern auch für weiterführende Analysen gelegt. Zu diesem Zweck wird eine projektspezifische Ontologie entworfen und kontrollierte Vokabularen aufgebaut bzw. weiterverwendet, die die Grundlagen für die zweite Phase des Projektes legen, in der die Datensammlungen weiterer Projektpartner*innen gemapped und mittels einer Ingest-Pipeline eingepflegt werden. Für die Ontologie wird auf -Kompatibilität geachtet, wobei die Besonderheiten von Tanz- und Theateraufführungen in eigenen Erweiterungen integriert werden. Schwierigkeiten ergeben sich aus den heterogenen Sammlungskonventionen und Erfassungsstrategien der Projektpartner*innen sowie unterschiedlichen Zugriffsgenehmigungen (vgl. Klimpel et al 2015), die es zu berücksichtigen gilt. Für den öffentlich zugänglichen Großteil der Metadaten wird eine Verfügbarkeit über das Portal des FID DK hergestellt (zum Portal siehe Beck et al 2016, Voß 2017). Es gilt darauf hinzuweisen, dass das Mediatheken-Projekt zunächst nur Metadaten vernetzt und nicht digitale Materialien anbietet. Die von den Projektpartner*innen übermittelten Daten werden disambiguiert, mit Identifiern zur GND, Wikidata sowie AV-spezifischen Services wie IMDB verbunden, mit zusätzlichen Daten angereichert und zueinander in Bezug gesetzt. Ein besonderes Augenmerk besteht darin, bruchstückhafte Informationen zu den Aufführungen mit Hilfe von tanz- und theaterspezifischen "authority files" wie den im FID DK-Portal verzeichneten Ereignissen zu ergänzen. Dabei wird auch ein Rückspielen der Informationen über verfügbare AV-Aufzeichnungen in die "authority files" anvisiert. Womit

mehrere Möglichkeiten zum Auffinden von AV-Material für Forscher*innen, Künstler*innen und Interessierte hergestellt wird.

Im Poster wird neben diesem Workflow auch die technische Infrastruktur dargelegt, die die Bereitstellung von Daten basierend auf den FAIR data principles organisiert. Das Ingest sowie das Post-processing – u.a. Disambiguieren, Enrichen, Mergen – der Daten der Projektpartner*innen wird mittels eines im Projekt entwickelten REST-API-driven Backends vorgenommen, wobei die Daten in einem Triple Store abgelegt werden.

Zusätzlich wird im Poster die Motivation für die Entwicklung der projektspezifischen Ontologie dargelegt, die Schnittstellen zu tanz- und theaterwissenschaftlichen Datenmodellen aufgezeigt und auf die Verwendung von Standards und Vokabularen aus den Fernseh- und Filmwissenschaft hingewiesen. Schließlich wird noch die Bedeutung einer föderierten Datensammlung von AV-Materialien von Tanz- und Theateraufzeichnungen hinsichtlich des Tagungsthemas der DHd2021 aufgezeigt, da die im Mediatheken-Projekt angewandten digitale Verfahren für die Gedächtnispflege der Tanz- und Theaterwissenschaft von großer Bedeutung ist.

Bibliographie

Beck, Julia; Dörner, Axel; Knepper, Marko; Voß, Franziska (2016): „Neue Wege der Informationsaggregation und -vernetzung. Ein Blick hinter die Kulissen des Fachinformationsdienstes Darstellende Kunst“, in: ABI Technik, Jg. 36, Heft 4, 2016, S. 218-226.

Fischer-Lichte, Erika (1999): *Kurze Geschichte des Deutschen Theaters*. 2. Aufl. ed., Tübingen: UTB A. Francke.

Fuxjäger, Anton (2020): "Die wissenschaftliche Videothek des Instituts für Theater-, Film- und Medienwissenschaft an der Universität Wien: Geschichte, Organisation, Technik", <https://tfm.univie.ac.at/sammlungen-einrichtungen/videothek/hintergrundinformationen/>

Klimpel, Paul, König, Eva-Marie (2015): "Urheberrechtliche Aspekte beim Umgang mit audiovisuellen Materialien in Forschung und Lehre", Gutachten für die Gesellschaft für Medienwissenschaft und den Verband der Historiker und Historikerinnen Deutschlands, Berlin September 2015.

Voß, Franziska (2017): „Der Fachinformationsdienst Darstellende Kunst“, in: Die vierte Wand: Organ der Initiative TheaterMuseum Berlin e.V., Berlin: Initiative TheaterMuseum Berlin e.V., Heft 7, 2017, S. 62-63.

Mein liebster Schatz! Das Citizen Science-Projekt Gruß & Kuss stellt sich vor

Rapp, Andrea

andrea.rapp@tu-darmstadt.de
TU Darmstadt

Büdenbender, Stefan

stefan.buedenbender@h-da.de
Hochschule Darmstadt

Dietz, Nadine

nadine.dietz@tu-darmstadt.de
TU Darmstadt

Dunkelmann, Lena

dunkelmann@uni-koblenz.de
Universität Koblenz

Gnau-Franké, Birte

bcgnauf@uni-koblenz.de
Universität Koblenz

Liesenfeld, Nina

nliesenfeld@uni-koblenz.de
Universität Koblenz

Schmunk, Stefan

stefan.schmunk@h-da.de
Hochschule Darmstadt

Seltmann, Melanie E.-H.

melanie.seltmann@tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Stäcker, Thomas

thomas.staecker@ulb.tu-darmstadt.de
TU Darmstadt, Universitäts- und Landesbibliothek

Werner, Stephanie

stephanie.werner@h-da.de
Hochschule Darmstadt

Wyss, Eva L.

wyss@uni-koblenz.de
Universität Koblenz

*Gruß & Kuss – Briefe digital. Bürger*innen erhalten Liebesbriefe* ist ein innovatives Verbund- sowie Citizen-Science-Projekt, das vom Bundesministerium für Bildung und Forschung für den Zeitraum von drei Jahren mit Start April 2021 gefördert wird. Es bindet aktiv Bürger*innen in die Digitalisierung und Erforschung von aktuell über 22.000 Liebesbriefen ein. Das Projekt wird durchgeführt von einem Team aus Wissenschaftler*innen unter der gemeinsamen Leitung von Prof. Dr. Andrea Rapp vom Institut für Sprach- und Literaturwissenschaft der Technischen Universität Darmstadt (TUDa), Prof. Dr. Eva L. Wyss vom Institut für Germanistik der Universität Koblenz-Landau (UKL), Prof. Dr. Stefan Schmunk vom Fachgebiet Bibliotheks- und Informationswissenschaft der Hochschule Darmstadt (h_da) sowie Prof. Dr. Thomas Stäcker, Bibliotheksdirektor der Universitäts- und Landesbibliothek Darmstadt (ULB) (vgl. Liebesbriefarchiv 2021a).

Grundlage des Projekts sind die seit 1997 im Liebesbriefarchiv (Wyss 2000) in Zürich und seit 2013 in Koblenz zusammengetragenen Spenden authentischer privater Liebesbriefe von Bürger*innen aus insgesamt drei Jahrhunderten (vom ältesten Brief aus dem Jahr 1768 bis zum jüngsten von 2021), aus 52

Ländern und sich wandelnder medialer Formate (vom klassischen Medium Brief über E-Mails bis hin zu WhatsApp-Nachrichten) (vgl. Liebesbriefarchiv 2021b). Hierbei handelt es sich um eine „unzugängliche Quelle der Alltagskultur [...] für die bislang kein staatlicher Sammlungsantrag existiert“ (Liebesbriefarchiv 2021a), wodurch das Archiv und das hieraus entstandene Projekt weltweit einzigartig ist.

In enger Zusammenarbeit von Wissenschaftler*innen und Bürger*innen möchte Gruß & Kuss diese Liebesbriefe gemeinsam erschließen, digitalisieren und erforschen. Dabei stellt sich die Frage, wie Emotionen wie Liebe, Glück und Leid, Sehnsucht sowie Intimität in ausgewählten Konstellationen wie Krisen, Trennung oder in geheimen Beziehungen erlebt und beschrieben werden. Um das zu untersuchen, sollen Liebesbriefe in diesem Forschungsprojekt mit zivilgesellschaftlicher Teilhabe erschlossen werden. Zudem wird erstmals die dauerhafte Erforschung und Bewahrung der Liebesbriefe in (digitalen) Gedächtniseinrichtungen sichergestellt (vgl. Liebesbriefarchiv 2021a; Hastik o. J.).

Während des Projekts werden Bürgerforscher*innen von Wissenschaftler*innen methodisch begleitet, unterschiedliche text- und sprachbasierte Untersuchungs- sowie Analysepraktiken durchzuführen. Dabei sollen digitale Werkzeugkästen erarbeitet und zur Nachnutzung auch für andere (Transkriptions-)Projekte zur Verfügung gestellt werden (vgl. Liebesbriefarchiv 2021a). Die Partizipationsmöglichkeiten sind angelehnt an das vierstufige Citizen-Science-Modell von Muki Haklay (2013: 116f.).

Tab. 1: Partizipationsstufen im Projekt Gruß & Kuss

Level	Name	Teilnahmemöglichkeiten
1	Crowdsourcing	Sichtung, Transkription und Edition von Liebesbriefen
2	Distributed Intelligence	Annotation von Briefen und Textelementen
3	Participatory Science	Identifizierung von thematischen Clustern und Formen des schriftlichen Ausdrucks durch digitale Analyse der Briefe
4	Extreme Citizen Science	Bearbeitung eigener Forschungsfragen

Dabei werden folgende Meilensteine anvisiert (vgl. Liebesbriefarchiv 2021a):

- Das „ExplorationLab“ wird Bürgerwissenschaftler*innen als Liebesbrief-Freunde in drei gemeinsamen Labs an digitale Methoden der Texterschließung heranführen (z.B. transkribieren). Hieraus werden besonders beispielhafte Liebesbriefe, wissenschaftliche Beiträge und (Transkriptions-)Tutorials erarbeitet und veröffentlicht.
- Aus den Liebesbrief-Freunden werden Liebesbrief-Forscher-Teams gebildet, die in Formaten wie einem „Blind Date Café/Stelldichein“ gemeinsam eigene Themencluster (z.B. heimliche/verbotene Liebe, Liebe im Krieg) identifizieren und Forschungsideen entwickeln. Auch hier münden die Ergebnisse der (bürger-)wissenschaftlichen Forschung in Tutorials und Blog-Beiträgen. Zudem soll eine „Love Coding App“ für die weitere Projektarbeit (und darüber hinaus) entwickelt werden.
- Das „Love-Coding-Lab“ veranstaltet standortübergreifend Workshops. Die beteiligten Bibliotheken dienen als Begegnungsort zwischen Zivilgesellschaft und Wissenschaft.

Das Projekt verfolgt mehrere Ziele: Zum einen die Zusammenführung von Wissenschaft und Zivilgesellschaft durch das gemeinsame Erforschen dieses einzigartigen Kulturgutes sowie die Bewusstseinsklärung für die eigene Kulturträgerschaft; zum anderen aber auch die Ermöglichung und Sicherstellung einer di-

gitalen, datenschutzkonformen Langzeitarchivierung der Liebesbriefe. Durch die dauerhafte Verankerung dieser Alltagsquelle als Teil eines digitalen und kulturellen Gedächtnisses wird eine wissenschaftliche, zivilgesellschaftliche und technologische Nachhaltigkeit für die weitere Nachnutzung geschaffen.

Das vorgestellte Poster fokussiert hierbei insbesondere die Partizipationsmöglichkeiten am Projekt *Gruß & Kuss* und zeigt auf, inwiefern Bürgerforscher*innen aktiv in das Projekt und den Forschungsprozess eingebunden werden können. Das Poster soll als Anstoß für einen wissenschaftlichen Austausch mit der DH-Community dienen.

Bibliographie

Haklay, Muki (2013): “Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation”, in: Sui, Daniel / Elwood, Sarah / Goodchild, Michael (eds.),

Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Dordrecht: Springer Netherlands 105–122.

Hastik, Canan (o. J.): “Integration and access to heterogeneous resources of the Koblenz love letter archive.” User Story 328. In: *Text+*. <https://www.text-plus.org/en/research-data/user-story-328/>.

Liebesbriefarchiv (2021a): *Gruß & Kuss – Briefe digital. Bürger*innen erhalten Liebesbriefe*. <https://liebesbriefarchiv.wordpress.com/grus-kuss-briefe-digital-burgerinnen-erhalten-liebesbriefe/>.

Liebesbriefarchiv (2021b): *LBAKatalog*. <https://liebesbriefarchiv.de/>.

Wyss, Eva (2000): “Intimität und Geschlecht. Zur Syntax und Pragmatik der Anrede im Liebesbrief des 20. Jahrhunderts”, in: Elmiger, Daniel / Wyss, Eva Lia (eds.): *Sprachliche Gleichstellung von Frau und Mann in der Schweiz*. La féminisation de la langue en Suisse. La femminilizzazione della lingua in Svizzera. L’egualità linguistica da donna ed um en Svizra. (= Bulletin VALS/ASLA 72) 187–210.

Möglichkeiten und Grenzen eines digitalen barocken Gedächtnisses

Ein DFG-Projekt in der Rückschau

Müller, Melissa

melissa.mueller@uni-hamburg.de
Universität Hamburg, Germany

Wie kann ein literaturwissenschaftlich und linguistisch annotiertes Korpus barocker Dramen Teil eines kulturellen Gedächtnisses¹ sein? Und hat diese Form der Daten Einfluss auf das kulturelle Gedächtnis? Welche Möglichkeiten und Grenzen eröffnen sich dadurch? Diese Fragen werden anhand eines Korpus, das im Rahmen des DFG-Projekts *Interaktionale Sprache bei Andreas Gryphius – datenbankbasiertes Arbeiten zum Dramenwerk aus linguistisch-literaturwissenschaftlicher Perspektive* entstanden ist, beantwortet (Informationen zum Korpus: <https://>

Fußnoten

1. Zum Begriff des kulturellen Gedächtnisses: Assmann/Assmann 1990; Assmann 1992; 1995. Welchen Beitrag Dramen zum Gedächtnis und zu Erinnerungen leisten, legt Assmann (1999) am Beispiel von Shakespeares Dramen dar.
2. Der Persistent Identifier (PID) lautet: <https://hdl.handle.net/11022/0000-0007-F00B-E>.
3. Im DFG-Projekt wurde auf eine Digitalisierung und Annotation der Kommentare verzichtet, da der Fokus auf sprachlichen Strukturen, die als interaktional zu beschreiben sind (auf die Kommentare trifft dies nicht zu), lag.

Bibliographie

- Assmann, Aleida** (1999): *Erinnerungsräume. Formen und Wandlungen des kulturellen Gedächtnisses*. München: C.H.Beck.
- Assmann, Jan** (1992): *Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. München: C.H.Beck.
- Assmann, Jan** (1995): "Text und Kommentar". In: Assmann, J. und B. Gladigow (Hrsg.): *Text und Kommentar. Archäologie der literarischen Kommunikation IV*. München: Fink, 9-33.
- Assmann, Aleida / Assmann, Jan** (1990): "Das Gestrern im Heute. Medien und soziales Gedächtnis". *Funkkolleg Medien und Kommunikation. Konstruktionen von Wirklichkeit, Studieneinheit I, Studienbrief 5*. Weinheim/Basel: Hessischer Rundfunk, 41-82.
- Fiehler, Reinhard** (2016): "Gesprochene Sprache". In: Wöllstein, A. und P. Eisenberg (Hrsg.): *Duden - die Grammatik*. Berlin: Dudenverlag, 1181-1260.
- Gryphius, Andreas** (1991): *Dramen*. Hrsg. v. Eberhard Manneck. Frankfurt/M.: Bibliothek deutscher Klassiker.
- Hennig, Mathilde** (2006): *Grammatik der gesprochenen Sprache in Theorie und Praxis*. Kassel: Kassel Univ. Press.
- Imo, Wolfgang** (i.V.): *Interaktionale Ellipsen: Nicht-finite Prädikationskonstruktionen (NFPK) und Aposiopesen im Dramenwerk von Andreas Gryphius*.
- Imo, Wolfgang / Lanwer, Jens Philipp** (2019): *Interaktionale Linguistik. Eine Einführung*. Berlin: Metzler.
- Imo, Wolfgang / Müller, Melissa** (i.V.): *Von „Ey Pickel-haring / das ist wider Ehr und Redligkeit“ zu „ey TImo; lass_ma RISCHtisch laut (.) öh schrEIen“ – „ey“ und „ei“ gestern und heute*.
- Jeßing B.** (2020): "Gryphius, Andreas". In: Arnold H.L. (Hg.): *Kindlers Literatur Lexikon (KLL)*. J.B. Metzler, Stuttgart. https://doi.org/10.1007/978-3-476-05728-0_6573-1 [zuletzt abgerufen am 23.11.21].
- Krause, Thomas / Zeldes, Amir** (2016): "ANNIS3: A new architecture for generic corpus query and visualization." in: *Digital Scholarship in the Humanities* 2016 (31). <http://dsh.oxfordjournals.org/content/31/1/118> [zuletzt abgerufen am 23.11.2021]
- Müller, Melissa** (i.V.): *„O Himmel / ich fall über den hauffen“ – „Ach warumb sterben wir Princesse nicht zusammen?“: Die Interjektionen „ach“ und „o“ im Dramenwerk von Andreas Gryphius*.
- Schwitalla, Johannes** (2012): *Gesprochenes Deutsch*. Berlin: E. Schmidt.

MUSE4Anything Ontologiebasierte Generierung von Werkzeugen zur strukturierten Erfassung von Daten

Bühler, Fabian

fabian.buehler@iaas.uni-stuttgart.de
Institut für Architektur von Anwendungssystemen (IAAS) -
Universität Stuttgart

Barzen, Johanna

johanna.barzen@iaas.uni-stuttgart.de
Institut für Architektur von Anwendungssystemen (IAAS) -
Universität Stuttgart

Leymann, Frank

Leymann@iaas.uni-stuttgart.de
Institut für Architektur von Anwendungssystemen (IAAS) -
Universität Stuttgart

Standl, Bernhard

bernhard.standl@ph-karlsruhe.de
Institut für Informatik und digitale Bildung - Pädagogischen
Hochschule Karlsruhe

Schlomske-Bodenstein, Nadine

nadine.schlomske-bodenstein@ph-karlsruhe.de
Institut für Informatik und digitale Bildung - Pädagogischen
Hochschule Karlsruhe

Einleitung

Zur Unterstützung der Erforschung von Text und Sprache als einer der Kernbereiche der Digital Humanities (Reichert 2017: 13), haben sich bereits verschiedene Methoden und Werkzeuge etabliert, wie EU Infrastrukturprojekte wie CLARIN (Hinrichs/Krauer 2014) verdeutlichen. Um sich darüber hinaus Artefakten wie beispielsweise Filmen, Musik, Gemälden oder Architekturen, aber auch abstrakten Artefakten wie Lehr-Lern-Szenarien, mittels digitaler Ansätze zu nähern, ist oftmals die Überführung dieser Artefakte in maschinenlesbare Repräsentationen nötig, was aufwendige und allzu häufig kostspielige Individuallösungen erfordert. Dieser Problematik wollen wir mit der hier vorgestellten Werkzeugunterstützung "MUSE4Anything" begegnen. MUSE4Anything soll die strukturierte und detaillierte Erfassung von unterschiedlichen Artefakten unterstützen. Hierzu erlaubt MUSE4Anything angepasste Eingabemasken zur Erfassung der relevanten Parameter zu erstellen. Diese werden automatisch aus der von dem Benutzer erstellten Ontologie (Ontologie nach dem Verständnis von Furrer (2014: 308-309) als Werkzeug der Wissensrepräsentation) generiert.

Für die Anforderungsanalyse als Basis von MUSE4Anything haben wir die domänenspezifischen Repositorien aus unseren

langfristig laufenden Projekten MUSE¹, welches vestimentäre Kommunikation im Film untersucht (Barzen et al. 2018a, 2018b) und MUSE4Music², das sich mit symphonischer Musik des 19. Jahrhunderts befasst (Barzen et al. 2017), herangezogen.

Anforderungsanalyse: MUSE und MUSE4Music

MUSE und MUSE4Music zielen auf die Identifikation von Mustern (nach dem Musterbegriff von Alexander et al. (1977)), um zu einem besseren Verständnis des jeweiligen Untersuchungsgegenstandes beizutragen. Dazu müssen in großem Umfang Daten erfasst, analysiert und interpretiert werden. Das Vorgehen und die dazugehörige Werkzeugumgebung unterstützen dabei: (i) die Definition der potenziell relevante Parameter mittels Ontologie, (ii) die Auswahl der Kriterien zur Zusammenstellung des Untersuchungsgegenstandes, (iii) die Erfassung der Daten mittels des Repositoriums (iv) die, auf die Art und Struktur der Daten abgestimmte Analyse der Daten und (v) die Überführung der validierten Ergebnisse in Muster (Barzen et al. 2018b).

Beide Projekte nutzen Mind-Map Programme, um die Ontologie zu verwalten (Schritt (i)) und ein jeweils speziell entwickeltes Datenrepository zur Erfassung der Daten (Schritt (iii)). Der in MUSE erfasste umfangreiche Datensatz (Barzen et al. 2021b), wird aktuell mit verschiedenen Ansätzen aus den Bereichen des Machine Learning erschlossen (Barzen 2021a).

Um das hier erarbeitete Vorgehen für weitere Anwendungsfälle zu erschließen, haben wir die vorhandenen Implementierungen systematisch analysiert und Kernanforderungen an ein generisches Werkzeug extrahiert. Diese beinhalten z. B. die Möglichkeit komplex strukturierte und teilweise tief verschachtelte Eigenschaften zu erfassen und die Ontologie direkt im Werkzeug zu erstellen und zu bearbeiten (Bühler 2021: 9–18). Zusätzlich haben wir existierende Lösungen zur Verwaltung von Ontologien, wie Protégé (Musen 2015), OWLGrEd (OWLGrEd 2021) oder VocBench (Stellato et al. 2020), evaluiert (Bühler 2021: 19–24) und uns in Hinblick auf die identifizierten Anforderungen für eine eigene Lösung entschieden.

MUSE4Anything

Auf Basis der identifizierten Anforderungen haben wir MUSE4Anything entwickelt. MUSE4Anything (Bühler 2021) ist ein generisches Datenrepository und unterstützt die Schritte (i) und (iii) aus dem vorhin beschriebenen MUSE-Vorgehen (Abbildung 1). Die benutzerdefinierte Ontologie zur Definition der Domäne, sowie die ggf. enthaltenen Taxonomien, können in dem Werkzeug erstellt (Schritt (i)) und jederzeit bearbeitet werden (Abbildung 2). Aus der Ontologie werden automatisch korrespondierende Eingabemasken für die Datenerfassung (Schritt (iii)) und Detailansichten der Objekteigenschaften generiert (Abbildung 3).

Alle Funktionen sind über eine Web-Oberfläche und mit der HTTP-API nutzbar. Dabei wurde im Besonderen für die Datenerfassung auf die Benutzbarkeit des Werkzeugs ohne umfangreiche Vorkenntnisse geachtet. Um die Nachvollziehbarkeit von Änderungen zu gewährleisten, sind alle Einträge, inklusive der Ontologie, versioniert. Die Funktionen des Repositoriums wurden anhand eines Anwendungsfalls basierend auf einem Ausschnitt der MUSE4Music Ontologie (Eusterbrock et al. 2017) evaluiert. Wei-

tere Details finden sich in (Bühler 2021: 25 ff., 41 ff.). Der Prototyp ist auf GitHub unter einer Open-Source-Lizenz frei verfügbar.³

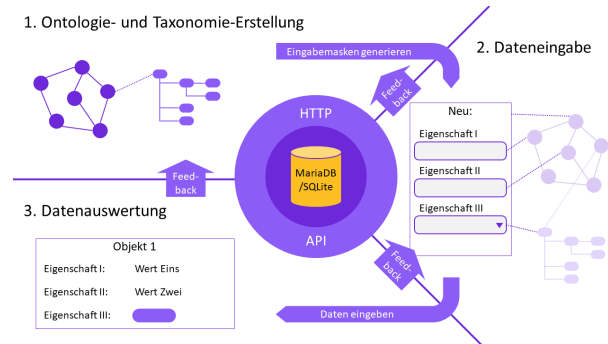


Abb. 1: Übersicht der von MUSE4Anything unterstützten Arbeitsschritte.

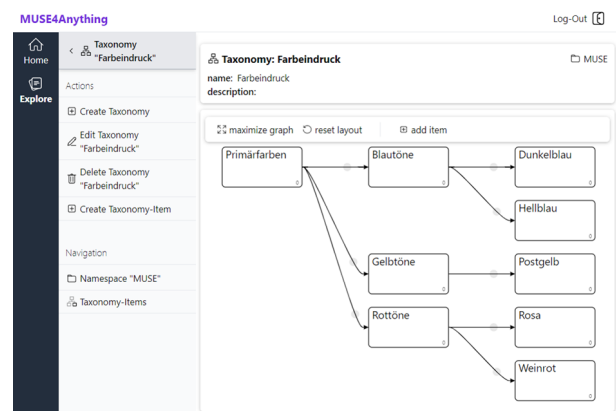


Abb. 2: Screenshot des Taxonomie-Editors mit einer stark vereinfachten Taxonomie aus der MUSE Ontologie.

Abb. 3: Screenshot einer generierten Eingabemaske.

Ausblick

MUSE4Anything befindet sich noch teils in der Entwicklung. Weitere Funktionen, darunter eine grafische Visualisierung der Ontologie, ein einfacherer Editor für die Ontologie und eine minimale Benutzerverwaltung, sind bereits in Arbeit. In Zukunft ist auch eine Anbindung an die Auswertungswerkzeuge, die für MUSE entwickelt wurden (Barzen 2021a: 35), vorgesehen.

Zusätzlich planen wir für die Evaluierung von MUSE4Anything die Anwendung in einer völlig anderen Domäne: An der Pädagogischen Hochschule Karlsruhe werden im Rahmen des von der "Qualitätsoffensive Lehrerbildung" geförderten Hochschulentwicklungsprojektes "Nachhaltige Integration von fachdidaktischen digitalen Lehr-Lern-Konzepten" (InDiKo) aus unterschiedlichen Fächern innovative digitale Lehr-Lern-Szenarien identifiziert. Um daraus fächerübergreifende didaktische Muster zu ermitteln (Standl/Schlomske-Bodenstein 2021) und für eine Wiederverwendung in der Lehre zu beschreiben, wird als Grundlage zunächst eine interdisziplinäre Taxonomie in einem Konsensvalidierungsprozess entwickelt und in MUSE4Anything abgebildet werden.

Fußnoten

1. www.iaas.uni-stuttgart.de/forschung/projekte/muse/
2. www.iaas.uni-stuttgart.de/forschung/projekte/muse4music/
3. www.github.com/Muster-Suchen-und-Erkennen/muse-for-anything

Bibliographie

Alexander, Christopher / Ishikawa, Sara / Silverstein, Murray / Jacobson, Max / Fiksdahl-King, Igrid / Angel, Shlomo (1977): „A Pattern Language: Towns, Buildings, Constructions.“ New York: Oxford University Press.

Barzen, Johanna (2018a): „Wenn Kostüme sprechen – Musterforschung in den Digital Humanities am Beispiel vestimentärer Kommunikation im Film.“ Dissertation, Universität zu Köln, <https://kups.ub.uni-koeln.de/9134/> [letzter Zugriff 6. Juli 2021].

Barzen, Johanna (2021a): „From Digital Humanities to Quantum Humanities: Potentials and Applications“ in: Miranda / E. R. (ed.): *An Introduction to Core Concepts, Theory and Applications*, Cham: Springer Nature. (Zur Veröffentlichung angenommen), Preprint <https://arxiv.org/abs/2103.11825> [letzter Zugriff 6. Juli 2021].

Barzen, Johanna / Breitenbücher, Uwe / Eusterbrock, Linus / Falkenthal, Michael / Hentschel, Frank / Leymann, Frank (2017): „The vision for MUSE4Music. Applying the MUSE method in musicology“ in: Bernhard Mitschang (eds.): *Computer Science - Research and Development. Advancements of Service Computing: Proceedings of SummerSoC 2016*, 32/3-4, 323-328.

Barzen, Johanna / Falkenthal, Michael / Leymann, Frank (2018b): „Wenn Kostüme sprechen könnten: MUSE - Ein musterbasierter Ansatz an die vestimentäre Kommunikation im Film“, in: Bockwinkel P et al (ed.) *Digital Humanities. Perspektiven der Praxis*, Frank & Timme, 223-241.

Barzen, Johanna / Bühler, Fabian / Leymann, Frank (2021b): „MUSE Datenset“. DaRUS, V1, 10.18419/darus-1805.

Bühler, Fabian (2021): „MUSE4Anything“ Masterarbeit, Universität Stuttgart 10.18419/opus-11410.

Eusterbrock, Linus / Barzen, Johanna / Hentschel, Frank (2017): „Eine Ontologie symphonischer Musik des 19. Jahrhunderts“. Technischer Bericht Nr. 2017/02, Universität Stuttgart.

Furrer, Frank J. (2014): „Eine kurze Geschichte der Ontologie. Von der Philosophie zur modernen Informatik.“ In: *Informatik Spektrum*. Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen 37 (2014), H. 4, 308-317.

Hinrichs, Erhard / Krauer, Steven (2014): „The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars.“ In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, Mai 2014, 1525-31, <http://dspace.library.uu.nl/handle/1874/307981> [letzter Zugriff 6. Juli 2021].

Musen, M.A. (2015): „The Protégé project: A look back and a look forward.“ In: *AI Matters*. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. 10.1145/2757001.2757003.

OWLGrEd (2021): <http://owlgred.lumii.lv> [letzter Zugriff 6. Juli 2021]

Reichert, Ramón (2017): „Digital Humanities als Wissenschaft“ in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): *Digital Humanities: Eine Einführung*. Stuttgart: J. B. Metzler, 13-34.

Standl, Bernhard / Schlomske-Bodenstein, Nadine (2021): „A Pattern Mining Method for Teaching Practices.“ In: *Future Internet* 13(5). 10.3390/fi13050106.

Stellato, Armando / Fiorelli, Manuel / Turbati, Andrea / Lorenzetti, Tiziano / Gemert, Willem / Dechandon, Denis / Laaboudi-Spoiden, Christine / Gerencsér, Anikó / Waniart, Anne / Costetchi, Eugeniu / Keizer, Johannes (2020): „VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons“, in: *Semantic Web* 11(5): 855-881, 10.3233/SW-200370.

Muster von “you” und “thou”

Modellierung der Anrede im englischen Sonett

Rath, Brigitte

brigitte.rath@uibk.ac.at
Universität Innsbruck, Austria

Bekanntermaßen unterscheidet das Early Modern English zwei Reihen von Pronomina der zweiten Person Singular: V-Formen (you, your etc.) und T-Formen (thou, thy etc.) Die T-Formen gehen im Zuge des Sprachwandels zum Modern English im Verlauf des 17. Jahrhunderts verloren (vgl. z.B. Lass 1999: 153), bleiben jedoch in der Lyrik erhalten. So beginnt etwa ein berühmtes, 1850 von Elizabeth Barrett Browning veröffentlichtes Sonett mit diesem Vers: “How do I love thee? Let me count the ways.” Diese wie selbstverständliche Verwendung der T-Formen in der englischsprachigen Lyrik auch weit nach dem Wandel zum Modern English ist bisher nicht systematisch untersucht.

Dieses Projekt sucht daher mit Hilfe digitaler Methoden Antworten auf folgende zwei Forschungsfragen:

- (1) Ist die Verwendung von T-Formen und V-Formen in Sonetten nach dem Sprachwandel synonym?

(2) Falls nein: Welche Muster lassen sich beschreiben?

Die Hypothese zur Frage (1) lautet, dass die Verwendung von T-Formen und V-Formen sich als nicht synonym erweisen wird, weil es gerade in der Lyrik eine gesteigerte Sensibilität für die Anrede gibt, und so zu erwarten steht, dass die linguistisch gebotenen Möglichkeiten für Differenzierung voll ausgeschöpft werden. Diese Erwartung widerspricht dem in Gedichtkommentaren häufig anzutreffenden und üblicherweise nicht weiter belegten Hinweis, "thou" sei einfach eine in Gedichten anzutreffende Version von "you".

Hypothesen für Faktoren, die eine Rolle bei der Frage (2) interessierenden Musterbildung spielen könnten, werden vor allem aus der historischen Soziolinguistik gewonnen: So kommen neben potentiellen individuellen Vorlieben von Autor:innen sowie der Entstehungszeit als plausible Faktoren nominale Anredeformen in Frage, weil Studien aus der historischen Soziolinguistik nahelegen, dass bestimmte Anreden (z.B. "terms of endearment") mit der Verwendung von T-Formen verbunden sind (vgl. Nevala 2004: 2146; Mazzon 2010), sowie die jeweilige Kategorie des:der Angesprochenen, weil Studien einen Zusammenhang zwischen bestimmten Kategorien von Angesprochenen wie etwa Kinder, Tiere oder Geister und der Verwendung von T-Formen zeigen (vgl. Yang 1991: 258; Carter/McRae 2002: 120-121).

Basis für diese Untersuchung ist ein selbsterstelltes Korpus von (bisher) 1.611 englischsprachigen, auf den britischen Inseln zwischen 1530 und 1910 publizierten Sonetten, für das die Gedichttexte manuell in TEI-5 konformem XML transkribiert und mit Metadaten (Autor:in, Titel, Entstehungsjahr, Publikationsjahr) und Annotationen zu nominalen Anredeformen, Kategorie der Adressat:innen (Gott, Mensch, Tier, Naturphänomen etc.), intertextuellen Verweisen und Reimschemata angereichert werden.

Mit diesem Korpus wurde eine Reihe von Experimenten mit Machine Learning Prediction Modellen gemacht. Mit fünf verschiedenen Machine Learning Prediction Modellen (Naive Bayes, Support Vector Machine, Decision Tree, Random Forest und XGBoost) wurde jeweils der k-fold cross validation approach (vgl. z.B. Han, Pei, Kamber 2011) durchgeführt. Die jeweiligen Trefferquoten wurden mit drei Baseline-Modellen (ZeroR sowie zwei Modellen, die alle Sonette jeweils einer Klasse zuordnen, hier: AlwaysT und AlwaysV) auf der Basis üblicher Standardwerte für Machine Learning verglichen: Precision, Recall, FMeasure, Accuracy und Area Under the ROC Curve (AUC). (vgl. z.B. Han, Pei, Kamber 2011; Mohri, Rostamizadeh, Talwalkar 2012) Es zeigt sich, dass Machine Learning Modelle, insbesondere XGBoost, bessere Ergebnisse als die Baseline-Modelle für die Vorhersage liefern können.

Tabelle: Ergebnisse für trainierte Machine Learning Prediction Models und Vergleichsmodelle

Model Name	Acc.	AUC	Macro-average			Weighted-average		
			Prec.	Rec.	F1	Prec.	Rec.	F1
ZeroR	0.797	0.5	0.399	0.5	0.444	0.636	0.797	0.707
Always Y	0.203	0.5	0.101	0.5	0.169	0.041	0.203	0.068
Always T	0.797	0.5	0.399	0.5	0.444	0.636	0.797	0.707
Naive Bayes	0.767	0.627	0.636	0.627	0.63	0.762	0.767	0.764
SVM	0.797	0.883	0.399	0.5	0.444	0.636	0.797	0.707
SVM - opt - acc	0.836	0.815	0.749	0.707	0.723	0.825	0.836	0.828
SVM - opt - f1	0.824	0.814	0.731	0.722	0.724	0.824	0.824	0.823
Decision tree	0.777	0.637	0.65	0.637	0.641	0.77	0.777	0.772
Decision tree - opt - acc	0.836	0.771	0.76	0.69	0.712	0.825	0.836	0.825
Decision tree - opt - f1	0.836	0.771	0.76	0.69	0.712	0.825	0.836	0.825
Random Forest	0.83	0.778	0.826	0.602	0.619	0.83	0.83	0.787
Random Forest - opt - acc	0.85	0.765	0.839	0.66	0.695	0.848	0.85	0.824
Random Forest - opt - f1	0.839	0.83	0.751	0.755	0.753	0.841	0.839	0.839
XGBoost	0.841	0.831	0.768	0.685	0.711	0.826	0.841	0.826
XGBoost - opt - acc	0.854	0.814	0.872	0.656	0.694	0.86	0.854	0.825
XGBoost - opt - f1	0.842	0.838	0.757	0.735	0.745	0.836	0.842	0.838

Abb. 1

Da dieses Modell einen hohen Anteil an Fällen korrekt zuordnet, folgt, dass das Modell Regeln in der Verteilung von T- und V-Formen erkennt, dass T- und V-Formen im Sonett also nicht austauschbar sind. Auf der Basis dieser Experimente kann so die erste Frage, "Ist die Verwendung von T-Formen und V-Formen in Sonetten nach dem Sprachwandel synonym?" tentativ mit nein beantwortet werden.

Für Hinweise auf mögliche Faktoren, die die Musterbildung beeinflussen, wurden mit den Machine Learning Prediction Modellen Ablation-Experimente durchgeführt: Input-Faktoren wurden individuell entfernt und die jeweilige Performanz des Modells erneut gemessen. Sinkt die Vorhersagekraft des Modells durch das Entfernen eines Faktors, so spielt dieser Faktor für die Vorhersage dieses Modells eine Rolle, was als ein erstes Indiz dafür gewertet werden kann, dass der entsprechende Faktor zur Musterbildung auch jenseits des Modells beitragen könnte. Es zeigt sich, dass dabei die Verwendung des Pronomens "ye", das bisher bei der Entwicklung von V- und T-Formen kaum beachtet wird, eine wichtige Rolle spielt; als weiterer möglicher Faktor erweist sich die Kategorie der Angesprochenen.

Das Projekt bietet für die historische Linguistik einen Beitrag zur präziseren Beschreibung der Sprachentwicklung. Für die Literaturwissenschaft erlaubt diese erstmalige systematische Beschreibung der Verteilung von T-Formen und V-Formen in englischsprachigen Sonetten bessere Gedichtinterpretationen, weil sie erstens überhaupt ein Augenmerk auf die verwendeten Pronomen der Anrede legt und zweitens die im Einzeltext gewählten Formen nun vor dem Hintergrund eines Musters gelesen werden können. Das Projekt trägt so zur aktuellen Forschungsdiskussion zur Anrede in der Lyrik bei. (vgl. z.B. Culler 1981, Culler 2015, Hedley 2009, Keniston 2006, Pollard 2012, Waters 2012)

Dieses Projekt wurde vom Vizerektorat Forschung der Universität Innsbruck mit Mitteln aus der Aktion D. Swarovski und vom Forschungszentrum Digital Humanities der Universität Innsbruck durch Mittel aus dem DI4DH Programm unterstützt und so erst ermöglicht. Die Korpuserstellung übernahmen mit einem ebenso scharfen Blick fürs Detail wie für das Gesamtprojekt Marina Höfler, Serena Obkircher und Teresa Wolf. Die Machine Learning Prediction Models wurden mit großer Umsicht von Ario Santoso und Mingzi Kong konzipiert, implementiert und trainiert.

Bibliographie

Carter, Ronald/ McRae, John (2002). "Language Note: Changing patterns of thou' and 'you'". *The Routledge History of Literature in English. Britain and Ireland*. Second Edition. Routledge. 120-121.

Culler, Jonathan (1981). "Apostrophe". Jonathan Culler. *The Pursuit of Signs. Semiotics, Literature, Deconstruction*. Routledge and Kegan Paul. 135-154.

Culler, Jonathan (2015). *Theory of the Lyric*. Harvard UP.

Han, Jiawei / Pei, Jian / Kamber, Micheline (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Hedley, Jane (2009). *I Made You to Find Me. The Coming of Age of the Woman Poet and the Politics of Poetic Address*. Ohio State UP.

Keniston, Ann (2006). *Overheard Voices. Address and Subjectivity in Postmodern American Poetry*. Routledge.

Lass, Roger (1999). "Phonology and Morphology". Roger Lass (ed.). *The Cambridge History of the English Language*. Volume III 1476-1776. Cambridge UP. 56-186.

Mazzon, Gabriella (2010). "Terms of Address". Andreas H. Jucker, Irma Taavitsainen (eds.) *Historical Pragmatics*. De Gruyter Mouton. 351-376. (Handbook of Pragmatics 8)

Mohri, Mehryar / Rostamizadeh, Afshin / Talwalkar, Ameet (2012). *Foundations of Machine Learning*. MIT Press.

Nevala, Minna (2004). "Accessing politeness axes: forms of address and terms of reference in early English correspondence". *Journal of Pragmatics* 36: 2125-2160.

Pollard, Natalie (2012). *Speaking to You. Contemporary Poetry and Public Address*. Oxford: Oxford University Press.

Waters, William (2012). Art. "Address". *The Princeton Encyclopedia of Poetry and Poetics*. Ed. Ronald Greene, Stephen Cushman. Fourth edition. Princeton UP. 6-8.

Yonglin, Yang (1991). "How to talk to the Supernatural in Shakespeare". *Language in Society* 20/2: 247-261.

NERDPool

Datenpool für Named Entity Recognition

Andorfer, Peter

peter.andorfer@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austria

Schlögl, Matthias

matthias.schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austria

Bleier, Roman

roman.bleier@uni-graz.at
Universität Graz, Austria

Bedeutung

In digitalen Editionen ist die automatische Erkennung und Annotation von Personen, Orten und Datumsangaben eine wichtige Aufgabe, die langfristig die händische Annotation ablösen wird. Machine Learning (ML) und Named Entity Recognition (NER) spielt dabei eine zentrale Rolle.¹ Historische Texte bilden noch ein Problem, da oft zu wenig Trainingsmaterial zur Verfügung steht, um entsprechende ML-Modelle zu trainieren. Andererseits werden seit über 30 Jahren digitale Editionen mit strukturierten Daten produziert, die diese Lücke füllen könnten.

Das von CLARIAH-AT finanzierte Projekt NERDPool versucht einerseits existierende (XML/TEI kodierte) Editionsdaten zu nutzen und daraus einen Pool an Trainingsdaten zu generieren, sowie andererseits Workflows zu erproben und zu implementieren, die es erlauben, einfach und effizient bestehende Korpora manuell zu annotieren. Den Schwerpunkt setzt das Projekt auf frühneuzeitliche deutsche Texte, ein Sprachstadium für die es wenig NER Material gibt. Die Datensätze werden über die Webapplikation <https://nerdpool-api.acdh-dev.oeaw.ac.at/> respektive über eine implementierte offene API veröffentlicht und können etwa mit Hilfe eines eigenen Python-Clients (<https://github.com/acdh-oeaw/nerdpool-client>, 14. Juli 2021) heruntergeladen werden. Mit

Stand Juli umfasst NERDPool rund 23.500 annotierte Datensätze. Darunter sind etwa Akten vom Regensburger Reichstag von 1576 (<https://reichstagsakten-1576.uni-graz.at>), Ministerratsprotokolle Österreichs und der österreichisch-ungarischen Monarchie 1848–1918 oder die ersten Ausgaben des Wienerischen Diariums (um 1750).

XML/TEI → Annotationen

Die in NERDPool gesammelten Daten sind stets das Resultat manueller Annotation. Die konkrete Annotationsarbeit erfolgte im Kontext der Erstellung einer XML/TEI kodierten Digitalen Edition. Hier wurden Personen, Orte, Datumsangaben mit entsprechenden TEI Tags annotiert. Die Daten werden über die GitHub API direkt von einem Repo abgerufen und dahingehend weiterverarbeitet, als die annotierten Textknoten gelesen und die Offsets der annotierten Elemente extrahiert werden (<https://github.com/acdh-oeaw/acdh-tei-pyutils>). Konkret wird ein Nodeset wie `<p><placeName>Wien</placeName>` ist eine Stadt.`</p>` in folgenden JSON-Eintrag `{ "text": "Wien ist eine Stadt.", "entities": [0, 3, "LOC"] }` konvertiert und anschließend in die Django basierte Webapplikation `nerdpool-api` importiert.

Prodigy & „custom loaders“

Ein zweiter Ansatz setzt auf das Annotationstoolkit Prodigy (<https://prodi.gy/>). Das kostenpflichtige und teilweise closed sourced Softwarepaket bietet ein äußerst effizientes Annotationsinterface und lässt sich sehr gut adaptieren, beispielsweise durch das Hinzufügen sogenannter 'custom loaders', welche Textdaten in das Annotationsinterface streamen und es so etwa erlauben Texte aus bestehenden APIs mit Prodigy zu annotieren. Mit einem solchen Loader „pr_transkribus.py“ (https://github.com/acdh-oeaw/acdh-prodigy-utils/blob/master/pr_transkribus.py) wurden etwa Texte direkt aus Transkribus über die Transkribus-API (<https://transkribus.eu/TrpServer/Swadl/wadl.html>) in Prodigy geladen.

Die Orchestrierung der einzelnen Prodigy-Instanzen, das notwendige Usermanagement der einzelnen Annotator*Innen sowie die Sicherung und Zusammenführung der Annotationsdaten erfolgt mittels Django, PostgreSQL, Nginx und dem Container Management Tool Portainer (`ptr target="https://github.com/acdh-oeaw/nerdpool/"`).

Probleme und Lösungen

In der konkreten Implementierung der obene beschriebenen Workflows bereitete vor allem die für Prodigy notwendige Tokenisierung und die darauf aufbauende Segmentierung (Sentence-Splitting) Probleme:

Die Syntax und Satzlänge historischer weicht teils massiv von jenen zeitgenössischer Texte - welche gemeinhin zum Training von NLP Modellen verwendet werden - ab.

In historischen Texten finden sich viele zum Teil heute nicht mehr gängige Abkürzungen (<https://abbr.acdh.oeaw.ac.at>) bzw. Trenn- und Satzzeichen - was sich wiederum ungünstig auf die Tokenisierung auswirkt.

Was das Problem einer automatisierten Satzsegmentierung betrifft, so wurde darauf zum Teil verzichtet und die Texte anhand von formalen Kriterien wie beispielsweise manuell annotierter (XML/TEI) oder von Layouterkennung erkannter Absätze geteilt.

Dies hat den Vorteil, dass die Annotationssamples nicht an den falschen Stellen unterbrochen werden, führt aber teilweise zu sehr langen Annotationssamples, welche das Annotieren vor allem über ein auf Effizienz ausgerichtetes System wie Prodigy erschwert.

In einem anderen Ansatz wurde das zur Tokenisierung verwendete Spacy Modell um eine Liste von Abkürzungen erweitert. Das funktioniert tendenziell gut, bringt allerdings einen (weiteren) technisch-administrativen Overhead hinsichtlich der Verwaltung der (Tokenisierungs-)Modelle mit sich.

Protokolle des österreichischen Ministerrates als Beispiel

Das Potential der gesammelten Annotationsdaten soll beispielhaft an der bereits erwähnten Edition der "Protokolle des österreichischen Ministerrates 1848-1867" (MRP) gezeigt werden. Auf Basis von rund 12.000 manuell annotierten Samples wurde ein spaCy NER Modell (Version 3.x) trainiert (ptr target="https://huggingface.co/csae8092/de_MRP_NER"/>). Während das kleine spaCy Standardmodell für Deutsch auf dem Evaluationsset der MRP Daten F1 Werte für Personen und Organisationen von rund 23 bzw. 12 Prozent erzielt, liegen die Werte beim MRP Modell bei 91 und 82 Prozent. Die Werte für die weiteren annotierten Kategorien LOC und GPE liegen bei 87 bzw. 58 Prozent (ptr target="https://github.com/csae8092/ner-tei-playgrounds"/>). Das MRP-Modell ist damit trotz historischer Sprachstufe, vielen Abkürzungen und vergleichsweise wenigen Trainingsdaten nur knapp unter aktuellen Named Entity Recognizern.

Fußnoten

1. Vgl. dazu die in der Bibliographie angeführte Literatur.

Bibliographie

Urbano, J et al. (2012): "Named Entity Recognition: Fallacies, challenges and opportunities", *Computer Standards & Interfaces* 35(5): pp. 482–489. doi: 10.1016/j.csi.2012.09.004 [letzter Zugriff 13. Juli 2021].

Kettunen, Kimmo / Mäkelä, Eetu / Ruokolainen, Teemu / Kuokkala, Juha / Löffberg, Laura (2017): "Old Content and Modern Tools: Searching Named Entities in a Finnish OCR'd Historical Newspaper Collection 1771–1910", in: *Digital Humanities Quarterly* 11(3), <http://digitalhumanities.org/dhq/vol/11/3/000333/000333.html> [letzter Zugriff 13. Juli 2021].

Kannisto, Maiju / Kauppinen, Pekka (2020): "Of Great Men and Eurovision Songs: Studying the Finnish Audio-Visual Heritage through NER-based Analysis on Metadata", in: Fridlund, Mats / Oiva, Mila / Paju, Petri (eds.) *Digital Histories: Emergent Approaches within the New Digital History*, 165–180.

Ontological modelling of the Greek Intangible Cultural Heritage for complex geo-semantic querying

Baglatzi, Alkyoni

alkyoni.baglatzi@spotin.org

Spotlight on Innovation (SPOTIN) NPCC, Greece

Velissaropoulos, Georgios

velissaropoulosg@yahoo.gr

Xorostasi NPCC, Greece

According to UNESCO¹, Intangible Cultural Heritage (ICH) includes all traditions passed over to us by our ancestors providing a sense of identity and continuity. Due to the evolution and changes of societies and the global character of our daily interactions nowadays, there is a big challenge in preserving important intangible cultural heritage assets of the past. Technology, though, provides a great opportunity to safeguard the wealth of these cultural assets and pass it over to the next generations.

Intangible cultural heritage data² is very broad including traditional dances and music, customs, health treatments etc. The current work focuses mainly on traditional music and dances of Greece. Greece has thousands of traditional dance and songs, differing a lot from place to place.

Although, a wealth of data exists in multiple forms such as videos, images, recordings, documents, physical and digital objects, it cannot be easily retrieved or interconnected. What is missing is a structured way to describe, document, formalize, visualize, and interlink this data with external resources. The current work demonstrates the use of ontologies and semantic web technologies to face this need, with particular emphasis on the spatial and temporal dimension as integrators of the information.

Maps are regarded as an enormously powerful and intuitive tool for visualizing data (Harley 2009) supporting critical thinking (Crampton 2001). The efficiency of maps has led to the development of the spatial humanities field demonstrating the power of maps for retrieving implicit knowledge of the past (Roberts 2016; Roberts et al. 2014). In the ICH domain though, little use of maps can be seen.

The importance of ontologies and linked open data in the ICH domain has already been acknowledged in various approaches (Chantas et al. 2018; Hou and Wang 2019; Ziku 2020). CultureSampo (Hyvönen et al. 2008), a flagship project introduced intelligent semantic web 2.0 technologies for cross-domain cultural heritage of the area of Finland. Europeana³, the largest EU repository of cultural heritage data, uses linked open data for providing the data in an interoperable form. Regarding Greek ICH, important projects include iTreasures (Dimitropoulos et al. 2014), Wholedance (Camurri et al. 2016) and Terpsichori (Doulamis et al. 2017) demonstrating the important contributions of semantic web technologies for ICH preservation.

In the current work, geo-semantic web technologies are being utilized in order to formalize and document all the data regarding the ICH of Greek traditional dances and songs. Ontologies are being used for the conceptualization of the information and its

provision as linked open data. For increasing interoperability and enabling the linkage with existing resources, already developed ontologies and schemas such as the DOLCE ontology (Borgo and Masolo 2010), the CIDOC CRM (Crofts et al. 2008) are being adopted. For the formalization of specialized domain concepts, the Greek Intangible Heritage Ontology (GIHO) is being developed with special focus on the spatial and temporal parameters. For enabling a better understanding of the data and providing more efficient ways of making it available to a wide range of users (either with a technical or non-technical background), a map-based web platform is being developed in which the end users will be able to pose complex queries. The Ontop-spatial (Bereta et al. 2016) and Sextant (Nikolaou et al. 2015) tools developed by the University of Athens, are being used for the processing and visualization of complex spatiotemporal thematic queries such as “Show me all places where dances with rhythms of 9/8 exist” or “Show me all the places where songs with the same text and different music exist”.

The contribution and novelty of our approach is threefold: 1) all the information about ICH currently kept in books, videos, etc. is being digitalized and formalized in an interoperable way, 2) a map based central access point is being developed enabling better overview of the information and 3) end users i.e. researchers from the social sciences are provided with an infrastructure that enables the investigation of complex queries and the retrieval of implicit knowledge (i.e. the way trade relations influenced the music and dances in the different regions of Greece)

Footnotes

1. <https://ich.unesco.org/doc/src/01856-EN.pdf>
2. <https://ich.unesco.org/doc/src/15164-EN.pdf>
3. <https://www.europeana.eu/en>

Bibliography

- Bereta, K., Xiao, G., Koubarakis, M., Hodrius, M., Bielski, C., and Zeug, G. (2016), "Ontop-spatial: Geospatial data integration using geosparql-to-sql translation", in *Proceedings of the 15th International Semantic Web Conference, Posters & Demonstrations Track (ISWC)*. Available at: <http://ceur-ws.org>, Vol. 1690.
- Borgo, S. and Masolo, C. (2010), *Ontological foundations of DOLCE*, pp. 279–295.
- Camurri, A., Sarti, A., Pietro, S., Viro, V., Whatley, S., El Raheb, K., Even Zohar, O., Ioannidis, Y., Markatzi, A., Matos, J.-M., Morley-Fletcher, E., Palacio, P. and Romero, M. (2016), *Wholodance: Towards a methodology for selecting motion capture data across different dance learning practice*, pp. 1–2.
- Chantas, G., Karavarsamis, S., Nikolopoulos, S. and Kompatsiaris, I. (2018), "A probabilistic, ontological framework for safeguarding the intangible cultural heritage", *Journal on Computing and Cultural Heritage (JOCCH)* 11(3), 1–29.
- Crampton, J. W. (2001), "Maps as social constructions: power, communication and visualization", *Progress in Human Geography* 25(2), 235–252.
- Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M. (2008), *Definition of the cidoc conceptual reference model*, ICOM/CIDOC Documentation Standards Group. CIDOC CRM Special Interest Group 5.
- Dimitropoulos, K., Manitsaris, S., Tsalakanidou, F., Nikolopoulos, S., Denby, B., Al Kork, S., Crevier-Buchman, L., Pil-

lot-Loiseau, C., Adda-Decker, M., Dupont, S., Tilmanne, J., Ott, M., Alivizatou, M., Yilmaz, E., Hadjileontiadis, L., Charisis, V., Deroo, O., Manitsaris, A., Kompatsiaris, I. and Nikos, G. (2014), *Capturing the intangible: An introduction to the i-treasures project*.

Doulamis, A., Voulodimos, A., Doulamis, N., Soile, S. and Lampropoulos, A. (2017), *Transforming intangible folkloric performing arts into tangible choreographic digital objects: The terpsichore approach*, pp. 451–460.

Harley, J. B. (2009), "Maps, knowledge, and power", *Geographic thought: a praxis perspective* pp. 129–148.

Hou, X. and Wang, X. (2019), "Modeling and representation of intangible cultural heritage knowledge using linked data and ontology", *Proceedings of the Association for Information Science and Technology* 56(1), 409–412.

Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkari, P., Laitio, J. and Nyberg, K. (2008), *Culturesampo – a collective memory of finnish cultural heritage on the semantic web 2.0*.

Nikolaou, C., Dogani, K., Bereta, K., Garbis, G., Karpathiotakis, M., Kyzirakos, K. and Koubarakis, M. (2015), "Sextant: Visualizing time-evolving linked geospatial data", *Journal of Web Semantics* 35, 35–52.

Roberts, L. (2016), *Deep mapping and spatial anthropology*.

Roberts, L., Thevenin, T., Hallam, J., Beveridge, A., Mos-tern, R., Southall, H., Cunningham, N. A., Schwartz, R. M. and Meeks, E. (2014), *Toward spatial humanities: Historical GIS and spatial history*, Indiana University Press.

Ziku, M. (2020), "Digital cultural heritage and linked data: Semantically informed conceptualizations and practices with a focus on intangible cultural heritage", *Liber Quarterly* 30(1).

Projektpräsentation "Early Medieval Glosses And The Question Of Their Genesis A Case Study On The Vienna Bede" (Gloss-ViBe)

Bauer, Bernhard

bernhard.w.bauer@gmail.com
Universität Graz, Austria

Bis heute ist das Annotieren von Texten eine gängige Praxis, deren Formen – Unterstreichen, Hervorheben, Glossieren, Kommentieren etc. – sich im Prinzip seit dem Frühmittelalter kaum verändert haben (vgl. Moulin 2009). Grundsätzlich gibt es zwei Arten von graphischen Elementen in einem Manuskript: den *principal text* und den *paratext*, der ein Manuskript als glossiert klassifiziert (vgl. Blom 2017, 10). Der Paratext lässt sich vom Primärtext in der *mise-en-page* – also der Gestaltung der Seite oder dem Layout –, durch seine Position, eine andere Schrift oder eine spezielle Markierung unterscheiden. Bei Glossen wird traditionell zwischen *interlinearen* und *marginalen* Glossen unterschieden.

Eine wissenschaftliche Beschäftigung mit (frühmittelalterlichen) Glossierungstraditionen bietet zahlreiche Anknüpfungs-

punkte für multi- und interdisziplinäre Forschungsvorhaben, etwa in Bezug auf die verschiedenen Wege intellektuellen Austauschs, Sprachkontakt, die Geschichte des Zweitspracherwerbs, der Schreib- und Lesekompetenz, der Buchproduktion sowie anderer kultureller und sozio-historischer Aspekte.

Das vorliegende Projekt Gloss-ViBe (MSCA-IF-EF-ST #101019035) beschäftigt sich im weitesten Sinn mit dem Sprach- und Kulturkontakt zwischen Irland und Kontinentaleuropa im Frühmittelalter. Es startete im September 2021 in Kollaboration mit dem *Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities* und dem *Institut für Antike* an der Universität Graz. Die zentrale Forschungsfrage bezieht sich auf die Genese der vernakulären frühmittelalterlichen keltischsprachigen Glossen: Sind die Glossen Originale oder Übersetzungen ursprünglich lateinischer Glossen? Um Antwort(en) auf diese Frage zu finden, wird eine Fallstudie an den keltischen und lateinischen Glossen der Handschrift *Wien, Österreichische Nationalbibliothek, Codex 15298 (olim Suppl. 2698)*¹, sowie den Parallelglossen in drei weiteren Manuskripten: Angers, *Bibliothèque municipale 477* (Ende 9. Jh.), Karlsruhe, *Badische Landesbibliothek, Augiensis pergamenum 167 (olim Codex Augiensis CLXVII)* (späte erste Hälfte 9. Jh.) und St. Gallen, *Stiftsbibliothek, MS 251* (erste Hälfte 9. Jh.), durchgeführt. Dieses fragmentarische Manuskript (4 Folios) stammt aus dem späten 8./frühen 9. Jahrhundert und beinhaltet zirka 200 Glossen zu Bedas *De Temporum Ratione* – wobei ungefähr ein Drittel in altirischer Sprache und der Rest in Latein verfasst ist. Forschungsgeschichtlich waren vor allem die irischen Glossen von Interesse (vgl. Stokes & Strachan, 1901–1903; Dillon, 1956; Bauer 2017). Die lateinischen Glossen sowie der Primärtext wurden bis dato noch nicht vollständig ediert, weshalb in Gloss-ViBe eine umfassende digitale Edition der Handschrift erstellt wird. Um die Genese der Glossen sowie die Texttradition näher beleuchten zu können, sind drei Forschungsziele formuliert:

- Transkription und Kollektion
- Digitale Edition
- Theoretisches Framework

Die Daten sowie das erarbeitete theoretische Framework werden unter Einhaltung der FAIR-Prinzipien für weitere Nutzung allgemein verfügbar sein. Die Transkription des Wiener Bedas sowie der Parallelglossen und die Metadaten werden in TEI/XML (Digitale Transkription und Edition) modelliert. Eine Langzeitarchivierung ist durch das *Geisteswissenschaftliche Asset Management System* (GAMS, Zentrum für Informationsmodellierung Graz) gesichert.

Die Transkription wird mit Transkribus² durchgeführt, was auch bedeutet, dass der gesamte Output des Projekts in ein zukünftiges HTR-Trainingsmodell einfließen kann. Neben einer normalisierten Transkription ist auch geplant, das Originaldokument im Sinne Pierazzos (2011) so nahe als möglich wiederzugeben. Das bedeutet, dass Abkürzungen (p/p für Lateinisch *per/pro*) oder Ligaturen (æ) nicht aufgelöst, sondern mit den jeweiligen Unicodezeichen transkribiert werden. Als Grundlage dafür dienen die Standards der *Medieval Unicode Font Initiative*.³ Das zweite Workpackage dient der Datenmodellierung und Erstellung der digitalen Edition – aufbauend auf den Frameworks von Rehbein (2014) und Monella (2019). Strukturelle Informationen wie z.B. Überschriften oder der Unterschied zwischen Primär- und Paratext werden kodiert. Weiters sind intra- und intertextuelle Links vonnöten, d.h. dass jedem glossierten Lemma sowie der Glosse selbst wird ein Unique Identifier zugewiesen. Intratextuell ist das wichtig, um etwa Marginalglossen ihrem konkreten Lemma

im Primärtext zuordnen zu können. Intertextuell sind diese Links vor allem für die Parallelglossen wichtig.

Um Antworten auf die Hauptforschungsfrage zu erlangen, wird das erstellte Korpus im dritten Workpackage analysiert. Hierzu wird ein Workflow erstellt, der Ansätze der *close* und *distant reading* verbinden soll (vgl. etwa Bauer 2019 & 2020). Die Parallelglossen werden mittels Netzwerkanalyse (Cytoscape, Gephi), Korpusanalyse Tools (AntConc, Voyant Tools), Kollations-Tools (CollateX, Juxta) aber auch „traditionellen“ Methoden der historisch-vergleichenden Sprachwissenschaft und Philologie untersucht. Durch minutiöse Analysen lässt sich die Richtung der Übersetzung bei mehrsprachigen Parallelglossen ermitteln. Mittels der Kollations-Tools können gemeinsame Fehler bzw. Abweichungen von der kanonisierten Version des Haupttextes ermittelt werden, welche auf eine engere Verbindung der Handschriften hinweisen. Dies kann mit Netzwerkanalysen visualisiert werden. Dadurch lässt sich ein Bild der Gelehrtennetzwerke im Frühmittelalter zeichnen.

Gloss-ViBe soll ein Modell für die Transkription/Edition und Analyse von (früh-)mittelalterlichen, glossierten Handschriften schaffen, das auch auf andere Epochen und Textsorten angewendet werden kann. Es soll Impulse für eine breitgefächerte fachliche Beschäftigung mit diesen reichhaltigen Fundgruben intersprachlichen und -kulturellen Wissenstransfers schaffen.

Fußnoten

1. Online unter <http://data.onb.ac.at/dtl/8650790>.
2. <https://readcoop.eu/transkribus/?sc=Transkribus>.
3. <https://folk.uib.no/hnooh/mufi/>.

Bibliographie

Bauer, Bernhard (2017): "New and corrected MS readings of the Old Irish glosses in the Vienna Bede", in: *Ériu* 67: 29–48.

Bauer, Bernhard (2019): "Venezia, Biblioteca Marciana, Zanetti lat. 349 an isolated manuscript? A (network) analysis of parallel glosses on Orosius' *Historiae adversus paganos*", in: *Études Celtiques* 45: 91–106.

Bauer, Bernhard (2020): "Distant Reading of Glossed Corpora: Stylometry and Network Analysis", Vortrag gehalten beim *Workshop: Glossing in Celtic Contexts*, Virginia Tech, Blacksburg, VA, USA, September 18–19, 2020.

Blom, Alderik H. (2017): *Glossing the Psalms*. Berlin/New York: Walter de Gruyter.

Dillon, Myles (1956): "The Vienna glosses on Bede", in: *Celtica* 3: 340–5.

Monella, Paolo (2019): „A digital critical edition model for Priscian: glosses, graeca, quotations“, in: *Analecta Romana Instituti Danici* 44, 135–149.

Moulin, Claudine (2009): "Paratextuelle Netzwerke: Kulturwissenschaftliche Erschließung und soziale Dimensionen der alt-hochdeutschen Glossenüberlieferung", in Krieger, Gerhard (ed.), *Verwandtschaft, Freundschaft, Bruderschaft. Soziale Lebens- und Kommunikationsformen im Mittelalter. Akten des 12. Symposiums des Mediävistenverbandes vom 19. bis 22. März 2007 in Trier*. Berlin: Akademie Verlag. 56–77.

Pierazzo, Elena (2011): "A Rationale of Digital Documentary Editions", in: *Literary and Linguistic Computing* 26(4): 463–477.

Rehbein, Malte (2014): "From the Scholarly Edition to Visualization: Re-Using Encoded Data for Historical Research", in: *In-*

ternational Journal of Humanities and Arts Computing 8.1: 81–105.

Stokes, Whitley / Strachan, John (1901–1903): *Thesaurus Palaeohibernicus*, vol. I and II. Cambridge: University Press.

Prosopographische Interoperabilität (IPIF) Stand der Entwicklungen

Vogeler, Georg

georg.vogeler@uni-graz.at
Universität Graz, Austria

Hadden, Richard

richard.hadden@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Schlögl, Matthias

matthias.schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Vasold, Gunter

gunter.vasold@uni-graz.at
Universität Graz, Austria

Prosopographie ist die Erhebung und Auswertung von individuellen Daten über historische Personen unter den Bedingungen einer eingeschränkten Quellenlage mit dem Ziel Aussagen über die Personen als Gruppe machen zu können. Auf der DHd2019 haben wir einen Vorschlag für eine geteilte RESTful API gemacht, mit der prosopographische Daten in ihren verschiedenen Inkarnationen leichter austauschbar gemacht werden sollen (Vogeler, Schlögl, Vasold 2018). Sie wird seitdem unter dem Titel “International Prosopographical Interchange Format” auf github weiterentwickelt (<https://prosopography.org/>). Seitdem haben wir an der Modellierung und der Implementierung von Prototypen gearbeitet.

Aus Sicht der **Modellierung** will IPIF nicht die Ausdrucksmächtigkeit von RDF basierten Modellen erreichen oder sich vollständig auf Upper-Level-Ontologies wie CIDOC-CRM abbilden lassen. Stattdessen zielt der Vorschlag der API darauf, technisch einfach zu implementieren zu sein. Es haben sich deshalb einige Modellierungsprobleme ergeben, zu denen wir Entscheidungen getroffen haben, die als exemplarisch für solche vereinfachten Lösungen gelten können:

Effiziente Filter: Das konzeptuelle Modell gruppiert sich um das von Jon Bradley in die Prosopographie eingeführte “Factoid” herum (Bradley & Short 2005; Pasin & Bradley 2015), das in der “Assertion” des Genealogie-Datenmodells (GedBase4All) 2010 eine Entsprechung gefunden hat. Davon sind die Entitäten *person* (Individuum), *statement* (Aussage über das Individuum) und *source* (Quellenbeleg für die Aussage) abhängig (Abb. 1)

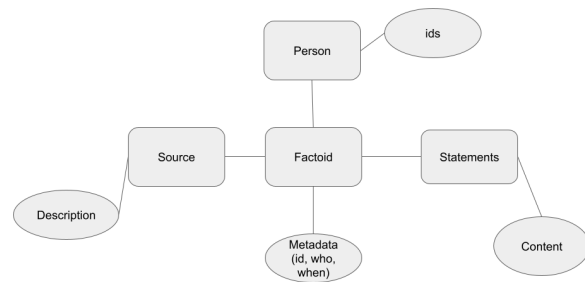


Abb. 1: Vereinfachte Darstellung des Datenmodells von IPIF

Da das Factoid aber in prosopographischer Forschung normalerweise nicht im Mittelpunkt steht, sondern Abfragen nach Personen und ihren Eigenschaften dominieren, bietet IPIF Endpunkte für genau diese Entitäten an, die leicht über ihre Beziehungen zu den anderen Entitäten gefiltert werden können: <https://example.org/ipif/v0.1/person/?name=Georg&place=Graz&from=2010&to=2015> legt für die Benutzer eine Suche nahe, die alle Personen identifiziert, über die es *statements* gibt, die ihr den Namen „Georg“ zuweisen und einen Aufenthalt in Graz zwischen 2010 und 2015. Man kann die Formulierung des Filters aber auch als Suche nach allen Personen, zu denen es je eine Aussage (*statement*) für jeden Parameter gibt, verstehen. Letztere Interpretation entspricht eher dem konzeptuellen Modell, weitet aber das Ergebnis auf schwer verständliche Art aus. IPIF verwendet deshalb die erste Interpretation als Standard, während die zweite über einen zusätzlichen Parameter `independentStatements=true` gelöst wird.

Labels für abstrakte Entitäten: Das Factoidmodell ist epistemologisch strikt und trennt das Individuum von seiner Beschreibung. So gibt es für die *person*-Identität im konzeptuellen Modell keinen menschenlesbaren Identifikator. In der Praxis ist das jedoch unrealistisch, da Listen von Personen, z.B. in *autocomplete*-Funktionen, ein menschenlesbares Label haben sollten. IPIF sieht deshalb vor, dass die jeweilige Implementation für Personenentitäten ein “*label*” erzeugt, das Informationen aus den Aussagen verwendet, die vom Standard nicht näher spezifiziert werden. Diese Label können also klassische Angaben aus Name, Lebensdaten und „Beruf“ sein, aber auch andere bei der Individualisierung helfende Daten. Sie müssen weder durchgängig konsistent mit den vorhandenen Statements sein, noch gibt die API ein Versprechen über ihre Stabilität ab.

Die derzeit umfangreichste **Implementierung** ist für das prosopographische Framework APIS in Entwicklung. APIS ist ein Entity-store, in der die aus dem Österreichischen Biographischen Lexikon (ÖBL) extrahierten Entitäten und ihre Beziehungen gespeichert werden (Schlögl & Leitowicz 2017). Das Framework ist damit ein generell für prosopographische Daten geeignetes Werkzeug und wird inzwischen in verschiedenen Projekten am ACDH-CH verwendet (APIShub: <https://apis-hub.acdh-dev.oeaw.ac.at/>). APIS bietet eine eigene RESTful API und Serialisierungen der Daten in TEI und CIDOC-CRM (<https://apis.acdh.oeaw.ac.at/apis/api2/>). Zusätzlich haben wir eine IPIF Schnittstelle für die APIS-Daten entwickelt. Die technische Lösung ist eine Solr basierte Bibliothek, die einen für APIS spezifischen Export aus APIS verwendet. Der Suchindex repräsentiert die Endpunkte als Dokumente und definiert für die von IPIF definierten Filter spezifische Indexe. Die Lösung übersetzt also nicht das konzeptuelle

elle Modell von IPIF direkt, sondern verwendete ein auf Performanz getrimmte Variante.

Einen ähnlichen Ansatz verfolgt der in Entwicklung befindliche IPIF-Python-Client, der Traversierungen zusammenfasst, um komplexere Abfragen zu vereinfachen, in denen die API in mehreren Schritten abgefragt werden müsste. Damit können Benutzungsfälle abgebildet werden, die auf Joins zwischen den drei Hauptentitäten beruhen, also z.B. Ergebnisse eines Filters auf *statements* an den *sources*-Endpoint weitergeben.

Aus diesen Fällen lassen sich folgende Ansprüche an pragmatische RESTful APIs generalisieren, die wir mit dem Poster zur Diskussion stellen wollen:

1. Bevorzuge Filter, die einen kleinen Datenausschnitt erzeugen, denn es ist einfacher, mehrere API-Aufrufe clientseitig zu kombinieren als ein zu großes serverseitiges Ergebnis clientseitig einzuschränken.
2. Keine URI ohne Label: abstrakte Konstrukte sollten immer eine menschenlesbare, semantisch besetzte Alternative besitzen, deren mangelnde semantische Präzision (es können fehlerhafte, veraltete oder umstrittene Angaben sein) hingenommen werden muss.
3. Erlaube Implementationen, die die Definitionen der API performant umsetzen, auch wenn sie nicht explizit das von der API verwendete konzeptuelle Datenmodell realisieren.

Bibliographie

Bradley, John / Short, Harold (2005): „Texts into Databases: The Evolving Field of New-Style Prosopography“, in: *Literary and Linguistic Computing* 20/Suppl 3–24. doi:10.1093/lc/fqi022.

Pasin, Michele / Bradley, John (2015): „Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach“, in: *Literary and Linguistic Computing* 30/1 86–97. doi:10.1093/lc/fqt037.

Schlögl, Matthias / Lejtovicz, Katalin (2018): „A Prosopographical Information System (APIS)“, in: Antske Fokkens, ter Braake, Serge, Sluijter, Ronald, Arthur, Paul, and Wandl-Vogt, Eveline (eds.): *BD-2017. Biographical Data in a Digital World 2017. Proceedings of the Second Conference on Biographical Data in a Digital World 2017. Linz, Austria, November 6-7, 2017*, CEUR Workshop Series 2119, [Budapest: CEUR] 53-58 < <http://ceur-ws.org/Vol-2119/paper9.pdf> > [letzter Zugriff 10.7.2021]

Vogeler, Georg / Vasold, Gunter / Schlögl, Matthias (2019). „Von IIF zu IPIF? Ein Vorschlag für den Datenaustausch über Personen“. Patrick Sahle (Hrsg.). *DHD 2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts. Universitäten zu Mainz und Frankfurt 25. bis 29. März 2019*. Frankfurt am Main. doi:10.5281/zenodo.2600812.

Zeditz, Jasper (2010): „Gedbas4all -- New Data Model for Genealogy“. In: *GenWiki* < <http://wiki-en.genealogy.net/Gedbas4all/Article> > [letzter Zugriff 10.7.2021]

Referenzierung des digitalen kulturellen (Text-)Erbes Digitale Quellenkritik und Modellierung von Metadaten

Althage, Melanie

melanie.althage@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Dreyer, Malte

malte.dreyer@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Guescini, Rolf

rolf.guescini@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Hiltmann, Torsten

torsten.hiltmann@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Lüdeling, Anke

anke.luedeling@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Odebrecht, Carolin

carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Forschungsgegenstand

Historische Textdaten wie etwa Urkunden, Briefe, Tagebücher aber auch literarische Texte sind integraler Bestandteil unseres kulturellen Erbes und insofern für viele geisteswissenschaftliche Fachbereiche wie die Sprach-, Literatur- und Geschichtswissenschaften die empirische Forschungsgrundlage. Für deren Referenzierung und Erschließung wurden insbesondere in den historischen Disziplinen und Methodiken zur Einordnung und Kritik solcher Quellen geprägt (grundlegend etwa: Droysen 1868: 16-19; Bernheim 1907: 113-134), die wiederum auch für andere wie die oben genannten Fachbereiche intellektuelle Zugänge zu Forschungsressourcen ermöglichen können. Die Anpassung der Quellenkritik an die spezifischen Eigenschaften digitalisierter und genuin-digitaler Quellen ist dagegen ein interdisziplinär zu verortendes Desiderat in dem emergierenden komplementären digitalen Forschungsparadigma „Digital Humanities“ (für die Geschichtswissenschaften etwa: Föhr 2019; Margulies 2009).¹ Mit domänen-spezifischen Zugängen, die sich Beschreibungsansätze aus den genannten Fachbereichen zu eigen machen, können (historische) Textdaten trans- und interdisziplinär kritisch eingeordnet, gefunden, erschlossen und wiederverwendet werden (vgl. FAIR Data Wilkinson et al. 2016).

In unserem Beitrag zeigen wir, wie eine domänenspezifische Datendokumentation, -kritik und -referenzierung mit standardisierten Methoden (TEI ODD Burnard 2013, UML-Modell für Metadaten Odebrecht 2019) für das LAUDATIO-Repositorium (Guescini, Schulz und Odebrecht 2021)² und damit für die historisch arbeitenden Textwissenschaften umgesetzt werden kann. Die Dokumentation und Kritik nehmen sich dabei die historische Quellenkritik als Vorbild. Entsprechend werden in LAUDATIO sowohl äußerlich beschreibende als auch aus den Inhalten extrahierte Metadaten verwendet, um eine Einordnung und Erschließung der textuellen Quellen zu ermöglichen und zugleich neue Wege zu deren korpusimmanenter wie -übergreifender Recherche und Analyse zu bieten. Ein Schwerpunkt wird hierbei zunächst auf Personen- und Ortsbezüge gesetzt (wie zum Beispiel mit GND). Dabei leiten uns drei Fragenkomplexe an:

I) Wie können wir diese multiplen Perspektiven (Kritik, Erschließung und Dokumentation) auf eine Quelle und über mehrere Quellen hinweg in einem Modell abbilden? Wie können wir Referenzen zu Personen und Orten in verschiedensten Quellen abrufbar machen? Wie gelingt die Einordnung der Quellen in etablierte Taxonomien beziehungsweise Ontologien?

II) Wir erweitern diesen Forschungsgegenstand um die folgende Perspektive: Wie ordnen sich Transkripte und Annotationen in dieses Informationsgefüge ein? Wie lassen sich deren digitale Repräsentationen und deren (gewollte) Manipulationen so beschreiben, dass auch dieser Teil der digitalen Quellenkritik sind?

III) Wie kann aus dieser Informationsdichte eine Exploration erfolgen, die den Forschenden die so versammelten Quellen als Forschungsdaten auffindbar, zugänglich, interoperabel und wiederverwendbar macht?

Um sich der Beantwortung dieser Fragen anzunähern und um im Rahmen einer nachhaltigen Infrastruktur als Weiterentwicklung des bestehenden Forschungsdatenrepositoriums LAUDATIO entsprechende Lösungsansätze umzusetzen, haben wir in einem ersten Schritt in einem interdisziplinären Team aus zentraler Infrastruktureinheit, Korpuslinguistik, Forschungsdatenmanagement und Geschichtswissenschaften 56 User Stories entwickelt, die wir in der Posterpräsentation vorstellen und mit der Community diskutieren.

Anwendungsperspektive

Diese User Stories und deren Akzeptanzkriterien stellen für uns die wesentliche Implementierungsgrundlage der Weiterentwicklung von LAUDATIO über die bisherige sprachwissenschaftliche Zielgruppe hinaus dar. LAUDATIO will explizit ein Angebot an Geisteswissenschaftler:innen verschiedenster Disziplinen zur nachhaltigen Publikation und Nachnutzung textbasierter Forschungsdaten machen. Eine breite Diskussion und ggf. Anpassung/Erweiterung ist daher essentiell. Nur so können wir LAUDATIO als Publikationsplattform in der Nutzerdomäne erweitern und einen wichtigen Beitrag für Open Science leisten sowie das Forschungsdatenmanagement unseres kulturellen Texterbes unterstützen.

User Stories für die Erschließung von Textdaten

Die User Stories repräsentieren die Bedarfe von historisch Forschenden, die das Repositorium und dessen Filter-/Suchinterfaces

zur Forschungsdatenpublikation und -recherche verwenden. Zwei entscheidende Befunde, die sich aus der Entwicklung der User Stories ergeben haben, betreffen die Umsetzung der FAIR-Prinzipien sowie die Berücksichtigung der Forschungs- und Arbeitsprozesse insbesondere im Bereich des Datenmanagements der Zielgruppe(n). Nachfolgend zeigen wir ein Beispiel für eine User Story mit Akzeptanzkriterien, die sich aus den Forschungsprozessen und Workflows der nutzenden Community ableiten:

User Story: *Als Geschichtswissenschaftler:in möchte ich eine Liste von Orten, die in den Dokumenten genannt sind, um nach für mich relevanten Dokumenten zu recherchieren.*

Akzeptanzkriterien:

- Dynamische Generierung von Ortslisten aus dem Metadatenindex
- Metadatenindex hält verschiedene Typen von Orten vor (Publikationsort, Ort genannt in Dokument)
- Orte sind, wo möglich, mit Geodaten referenziert
- Auswahl und Anzeige der Ortsliste über die Expertensuche
- Auswahl und Anzeige der Ortsliste über die Metadatenanzeige pro Korpus und pro Dokument

Der thematische Schwerpunkt liegt bei uns in der Modellierung flexibler, unterschiedlich granularer und miteinander verschränkter Referenzierungen, die auf Korpus-, Text- und Annotations-ebene ansetzen und miteinander interagieren. Personen, Orte und Werke sind Entitäten, die für eine digitale Quellenkritik entscheidende Informationen tragen. Ein umfassendes Desiderat für die Weiterentwicklung von LAUDATIO sind die Akzeptanzkriterien, welche das messbare Ergebnis und damit den Erfolg der User Story für die nutzende Community beschreib-, validierbar und im Rahmen der Entwicklung implementierbar machen. Wir diskutieren auf der Postersession nicht nur die User Stories sondern auch die Akzeptanzkriterien.

Fußnoten

1. Vgl. auch weitere Ansätze zur Diskussion einer digitalen Quellenkritik *Virtuelles BarCamp zu den theoretischen Aspekten einer digitalen Quellenkritik* : <https://vdhd2021.hypotheses.org/288> [letzter Zugriff 15.Juli 2021].
2. <https://www.laudatio-repository.org/> besucht am 12.07.2021.

Bibliographie

- Bernheim, Ernst** (1907): *Einleitung in die Geschichtswissenschaft*. Leipzig: Göschen'sche Verlagsbuchhandlung.
- Burnard, Lou** (2013): „Resolving the Durand Conundrum“, in: *Journal of the Text Encoding Initiative* 6 <http://journals.openedition.org/jtei/842> [letzter Zugriff 13.07.2021].
- Droysen, Johann Gustav** (1868): *Grundriss der Historik*. Leipzig: Veit.
- Föhr, Pascal** (2019): *Historische Quellenkritik im Digitalen Zeitalter*. Glückstadt: Verlag Werner Hülsbusch.
- Guescini, Rolf / Schulz, Konstantin / Odebrecht, Carolin**. (2021) . *Laudatio Repository - Long-term Access and Usage of Deeply Annotated Information Docker Images (Version 0.1)* . Zenodo. <http://doi.org/10.5281/zenodo.5101524>
- Margulies, Simon B.** (2009): *Digitale Daten als Quelle der Geschichtswissenschaft. Eine Einführung* (Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik, Bd. 2). Hamburg: Verlag Dr. Kovač.

Odebrecht, Carolin (2019): „A Model-to-model Approach for Developing Corpus Metadata. An “Odd” TEI Customization for Encoding Metadata“, in: Digital Humanities 2019 Conference Proceedings. Utrecht University, 9-12-7.2019, Utrecht.

Wilkinson, M. D. et al. (2016): „The FAIR Guiding Principles for scientific data management and stewardship“, in: *Scientific Data* 3, DOI: 10.1038/sdata.2016.18.

Reflexive Passagen und ihre Attribution

Varachkina, Hanna

hanna.varachkina@stud.uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Barth, Florian

barth@sub.uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Gödeke, Luisa

luisa.goedeke@uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Hofmann, Anna Mareike

annamareike.hofmann@uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Dönicke, Tillmann

tillmann.doenicke@uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Im Projekt MONA (Modes of Narration and Attribution) werden die Phänomene in fiktionaler Literatur untersucht, die mit reflexiven Passagen assoziiert sind. Reflexive Passagen kommentieren die Handlung im Text oder den Schreibprozess oder generalisieren über die fiktive und / oder reale Welt. Da das Konzept der reflexiven Passagen in der Literaturwissenschaft bisher nicht formalisiert wurde, werden diese nicht direkt annotiert. Stattdessen annotieren wir drei Phänomene, die wir für starke Indikatoren reflexiver Passagen halten: Kommentar (Bonheim 1975; Chatman 1980), nicht-fiktionale Rede (Konrad 2017; Searle 1975) und Generalisierung (Leslie et al. 2016; Dönicke et al. 2021). Darüber hinaus beschäftigt sich das Projekt mit der Zuschreibung reflexiver Passagen zu Sprechinstanzen. Für die Identifikation und Klassifikation dieser Phänomene werden Modelle entwickelt. Dafür wird ein annotiertes Korpus deutschsprachiger fiktionaler Texte erstellt, das die Entwicklung dieser Phänomene über 350 Jahre der Literaturgeschichte abbildet.

Basierend auf den bisherigen Arbeiten haben wir Definitionen für die mit reflexiven Passagen assoziierten Phänomene formuliert bzw. weiterentwickelt. Unter Generalisierungen werden quantifizierte Aussagen über angenommene Instanzen einer Klasse oder Gruppe von Objekten, Individuen oder (Zeit-)Räumen verstanden, auf die nicht kontextuell referiert wird. Kommentare schließen Textstellen ein, in denen die erzählte Zeit unterbrochen und eine ergänzende Information zu Erzählung, Figuren, Handlung oder dem Akt des Erzählens eingefügt wird (Bonheim 1975). Nicht-

fiktionale Rede bezeichnet Passagen in fiktionalen Texten, die Behauptungen bzw. Hypothesen über die reale Welt nahelegen (Konrad 2017). Generalisierung, Kommentar und nicht-fiktionale Rede können sich vollständig oder teilweise überlappen.

Und so begann der Hauptmann: »[An allen Naturwesen, die wir gewahr werden, bemerken wir zuerst, daß sie einen Bezug auf sich selbst haben. [...]]« (Goethe 2012)

In diesem Beispiel treten alle drei Phänomene auf. Die erzählte Zeit, die in der Erzählerrede fließt, wird unterbrochen und ein Kommentar über Naturwesen vorgenommen. Zugleich wird eine Aussage über angenommene Instanzen der Klasse der Naturwesen getroffen. Da auch in der realen Welt Naturwesen (jeglicher Art) vorkommen, ist die Proposition grundsätzlich auf die reale Welt übertragbar.

Für reflexive Passagen erstellen wir eine Goldannotation, auf der in einem nächsten Schritt eine Attributionsannotation vorgenommen wird. Die Attribution bestimmt, wem die in der Passage enthaltene Information zugeschrieben werden kann, wofür grundsätzlich Figuren, die Erzählinstanz oder die AutorIn in Frage kommen. Einige sprachliche Mittel im Text sind prädestiniert für bestimmte Attributionen, so markieren bestimmte Satzzeichen i. d. R. (in)direkte Rede und damit die Sprecher im Text. Dennoch gibt es Passagen, in denen sich die Sprechinstanz nicht eindeutig identifizieren lässt und sich unterschiedliche Interpretationen (Zuschreibungen) aufdecken lassen.

Zur automatischen Erkennung und quantitativen Analyse erstellen wir das Korpus MONACO (Modes of Narration and Attribution Corpus)¹, das aus deutschsprachigen fiktionalen Erzähltexten von 1600–1950 besteht. Jede AutorIn ist im Korpus nur einmal vertreten und es wird eine gleiche Verteilung der Texte über Jahrhunderte angestrebt. Die wichtigste Textquelle für MONACO ist Kolimo (Herrmann und Lauer 2017).

Die Annotation der Texte wird in CATMA 6.2² vorgenommen und basiert auf detaillierten Richtlinien, die iterativ entwickelt wurden. Dabei werden momentan lediglich die ersten 200 Sätze eines Textes annotiert, um mehr Texte zu annotieren und das Korpus möglichst divers gestalten zu können. Jedes der drei mit reflexiven Passagen assoziierten Phänomene wird von zwei studentischen AnnotatorInnen annotiert. Der Goldstandard für Generalisierung, Kommentar und nicht-fiktionale Rede wird in Kleingruppen von 2-3 DoktorandInnen erstellt. Zur Beschleunigung der Goldstandarterstellung haben wir einen „CATMA-Merger“ entwickelt, welcher eine neue Annotation als die Vereinigung mehrerer Annotationen erstellt, die dann von den DoktorandInnen überprüft und bestätigt, korrigiert oder gelöscht werden kann. Die Attribution wird im zweiten Schritt von allen (sechs) studentischen AnnotatorInnen auf dem Goldstandard für Generalisierung, Kommentar und nicht-fiktionale Rede annotiert. Für Attribution wird kein Goldstandard erstellt, um sämtliche mögliche Interpretationen zu erfassen.

Bisher wurden Goldstandards für achtzehn Texte erstellt. Der älteste Text stammt aus dem Jahr 1616, der jüngste aus dem Jahr 1930. Die annotierten Texte weisen im Durchschnitt ein moderates ($> 0,4$) oder gutes ($> 0,6$) Inter-Annotator Agreement mit κ -Werten (Fleiss et al. 1981) von 0,59 für Generalisierung, 0,44 für Kommentar und 0,66 für nicht-fiktionale Rede auf. Die Inter-Annotator Agreement-Werte für γ (Mathet et al. 2015) sind etwas höher: 0,66 für Generalisierung, 0,52 für Kommentar und 0,72 für nicht-fiktionale Rede.

Mit der zunehmenden Menge annotierter Texte werden schrittweise regelbasierte, statistische und neuronale Tagger für die einzelnen Phänomene entwickelt. Ihre Anwendbarkeit wird dabei auch für andere Textsorten wie Essays und enzyklopädische Texte

erprobt. Letzten Endes soll eine ausreichend große Menge annotierter Daten nicht nur bessere Modelle ermöglichen, sondern auch diachrone oder genreübergreifende Perspektiven auf reflexive Passagen und ihre Attribution eröffnen.

Fußnoten

1. <https://gitlab.gwdg.de/mona/korpus-public>
2. <https://catma.de/>

Bibliographie

- Bonheim, Helmut** (1975): "Theory of narrative modes ", in: *Seimiotica* 14/4: 329–344.
- Chatman, Seymour Benjamin** (1980): *Story and Discourse: Narrative Structure in Fiction and Film*. Ithaca / London: Cornell University Press.
- Dönicke, Tillmann / Gödeke, Luisa / Varachkina, Hanna** (2021): Annotating Quantified Phenomena in Complex Sentence Structures Using the Example of Generalising Statements in Literary Texts, in: *Proceedings of the 17th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* 20-32.
- Fleiss, Joseph L.** (1971): "Measuring nominal scale agreement among many raters ", in: *Psychological bulletin* 76.5: 378.
- Goethe, Johann Wolfgang** (2012): "Die Wahlverwandtschaften ", in: *TextGrid Repository, Digitale Bibliothek*, <https://hdl.handle.net/11858/00-1734-0000-0006-6A93-D> [letzter Zugriff 26. November 2021]
- Herrmann, Berenike / Lauer, Gerhard** (2017): "KOLIMO. A corpus of Literary Modernism for comparative analysis ", <https://kolimo.uni-goettingen.de/about>
- Konrad, Eva-Maria** (2017): "Signposts of Factuality: On Genuine Assertions in Fictional Literature " in: Sullivan- Bissett, Ema / Bradley, Helen / Noordhof, Paul (eds.): *Art and Belief*. Oxford: Mind Association Occasional Series 42-62.
- Leslie, Sarah-Jane / Lerner, Adam / Zalta, Edward Nouri** (2016): "Generic Generalizations ", in: *Stanford Encyclopedia of Philosophy* <https://plato.stanford.edu/entries/generics/> [letzter Zugriff 26. November 2021]
- Mathet, Yann / Widlöcher, Antoine / Métivier, Jean-Philippe** (2015): "The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment", in: *Computational Linguistics* 41.3: 437–479.
- Searle, John Rogers** (1975): "The logical status of fictional discourse ", in: *New literary history* 6(2): 319-332.

Repositorien als digitale Gedächtnisträger zwischen Evolution und Langzeitplanung

Steiner, Elisabeth

elisabeth.steiner@uni-graz.at
Karl-Franzens-Universität Graz, Austria

Vasold, Gunter

gunter.vasold@uni-graz.at
Karl-Franzens-Universität Graz, Austria

Scholger, Martina

martina.scholger@uni-graz.at
Karl-Franzens-Universität Graz, Austria

Einleitung

Seit beinahe 20 Jahren verwaltet und publiziert das Repository GAMS Forschungsdaten aus den Geisteswissenschaften und dem Kulturerbebereich. Derzeit enthält das Langzeitarchiv etwa 115.000 annotierte digitale Objekte aus mehr als 60 Projekten. Dies reicht von digitalen Editionen oder Textsammlungen (z.B. Briefe, mittelalterliche Rechnungsbücher) über Bildsammlungen (z.B. historische Fotografien und Postkarten) bis hin zu digitalisierten musealen Sammlungen (z.B. archäologische Artefakte oder andere Museumsobjekte) (GAMS 2021). Dabei liegt das Augenmerk sowohl auf der langfristigen Sicherung und Zugänglichkeit von Ressourcen wie auch auf dem nachhaltigen Umgang mit den bearbeiteten Forschungsdaten.

Domänenspezifische Repositorien müssen einen Ausgleich zwischen langzeittauglicher und somit eher konservativer Entwicklung und dem Einsatz aktueller Technologien finden. Das Poster wird den konzeptionellen Rahmen dieser Forschungsdateninfrastruktur vorstellen und Lösungsansätze aus diesem Widerspruch diskutieren, die besonders auf die Migration der Kernkomponenten fokussieren.

Grundlegende Konzepte

Das OAIS-konforme (Space Data Systems 2012) Repository folgt einer weitgehend XML-basierten Inhaltsstrategie, die domänenspezifische Datenmodelle und bewährte Standards nutzt. Alle kooperativen Forschungsprojekte werden innerhalb der gleichen Infrastruktur realisiert, wobei eine begrenzte Anzahl von bevorzugten Formaten und Technologien verwendet wird, um die Wartbarkeit über einen langen Zeitraum zu gewährleisten. Das System stützt sich ausschließlich auf Open-Source-Software (als Kernelement Fedora Commons), da es sich den Prinzipien von Open Access und Open Science verpflichtet. Die Designprinzipien zielen auf die Einhaltung internationaler Best Practices wie dem COAR Community Framework for Good Practices in Repositories (COAR 2020) und Zertifizierungsrichtlinien (das Repository ist derzeit mit dem CoreTrustSeal und als CLARIN B-Centre zertifiziert) und folgen den FAIR-Prinzipien (Force 11 2016).

Zentrales Merkmal ist die Verwaltung digitaler Ressourcen als komplexe digitale Objekte auf der Grundlage von Inhaltsmodellen (Klassen). Jedes Inhaltsmodell definiert für einen Datentyp eine beliebige Anzahl von Datenströmen (Eigenschaften), ergänzt durch teilweise automatisch generierte Metadatensätze und Methoden, über die spezielle Sichtweisen auf die Objekte realisiert werden. Die digitale Repräsentation einer mittelalterlichen Handschrift besteht beispielsweise aus deskriptiven, technischen, strukturellen und administrativen Metadaten, einer Anzahl von Faksimiles, einer TEI-kodierten Transkription und RDF-

Repräsentationen zur Unterstützung von Linked Data. An das Objekt gebundene Methoden ermöglichen synoptische Darstellungen mehrerer Varianten und bieten projektspezifische Such-, Visualisierungs- und Analysefunktionen. Alle Datenströme und Methoden werden innerhalb eines einzigen, vollständig selbstbeschreibenden Objekts gespeichert und verwaltet, das über einen persistenten Identifikator adressierbar ist, was erhebliche Vorteile für die Langzeitarchivierung bietet.

Herausforderung Langzeitplanung

Eine Herausforderung hinsichtlich der langfristigen technischen Erhaltung jedes Repositoriums liegt einerseits in der sich laufend ändernden IT-Umgebung (Betriebssysteme, neue Software-Versionen), andererseits in sich verändernden Ansprüchen an die Nutzbarkeit eines solchen Systems aus BenutzerInnen-sicht; aus organisatorischer Sicht erweist sich die dauerhafte Finanzierung oft als schwierig.

Die technische Evolution machte den Austausch der zentralen Komponente des Systems notwendig. Der Schritt von Fedora 3 auf Fedora 6 war kein simples Upgrade, sondern bedeutete einen grundlegenden Wechsel vom objektorientierten Paradigma zu einer Linked-Data-Architektur auf Basis der Linked-Data-Plattform (LDP)-Spezifikation (Speicher / Arwe / Malhotra 2015). Die Abwendung vom objektorientierten Paradigma und damit das Verschwinden der für uns zentralen Inhaltsmodelle (DuraSpace 2016) stellte uns vor erhebliche Herausforderungen, da der Verzicht auf dieses Prinzip eine massive Überarbeitung aller bestehenden Daten und Zugriffsmethoden bedingt hätte. Als Lösungsstrategie wurde eine Kompatibilitätsschicht eingezogen, die die bestehende objektbasierte API auf die neue graphenbasierte API abbildet. Diese Schicht wurde als Service implementiert, wie auch allgemein darauf geachtet wurde, das neue System als Sammlung einzelner, Docker-basierter Services zu realisieren, die nun in einem Kubernetes-Cluster laufen.

Diese Kompatibilitätsschicht ermöglicht die Übernahme aller bestehenden Projekte und Forschungsdaten ohne Änderungen an diesen notwendig zu machen. Dies wurde ausschließlich durch die einheitliche Abwicklung in den Jahren davor ermöglicht und zeigt die Notwendigkeit dieser Begrenzung mit Hinblick auf die nachhaltige Verfügbarkeit. Zurückblickend auf das Beispiel der mittelalterlichen Handschrift bedeutet dies, dass sich an der grafischen Oberfläche des Projektes für NutzerInnen nichts verändert, alle Links und Disseminatoren weiterhin funktionieren und auch die persistente Identifikation gewährleistet bleibt, obwohl im dahinter liegenden System kaum ein Stein auf dem anderen geblieben ist.

Fazit

Der Betrieb jedes Repositoriums erfordert die periodische Erneuerung des Technologiestacks. Um eine solche Migration reibungslos zu ermöglichen, müssen die enthaltenen Daten und Projekte einem eng spezifizierten Rahmen genügen. Zusätzlich zu diesem praktischen Erfordernis muss das Repositorium als organisatorische Einheit ausreichend finanzielles und institutionelles Commitment beweisen, um eine solche massive und personalintensive Umstellung bewältigen zu können. Das Ersetzen von Kernkomponenten der Software-Infrastruktur kann damit als ultimativer Beweis für eine erfolgreiche Langzeitstrategie des Betreibers angesehen werden. Dies unterstreicht, dass ein Repositorium in erster Linie keine technische Lösung, sondern eine organisato-

rische Einheit ist, die bereit ist, aktiv Verantwortung für die enthaltenen Ressourcen zu übernehmen. Vertrauenswürdige Repositorien tragen diese Verantwortung nicht nur für die Wissenschaft, sondern für das kollektive kulturelle Gedächtnis und bilden somit den Grundstein eines "digitalen Gedächtnisses".

Bibliographie

- Geisteswissenschaftliches Asset Management System (GAMS).** <https://gams.uni-graz.at> [letzter Zugriff 14. Juli 2021].
- Brickley, D. / Guha, R. V.** (2014): RDF Schema 1.1. <https://www.w3.org/TR/rdf-schema> [letzter Zugriff 14. Juli 2021].
- Confederation of Open Access Repositories (COAR)** (2020): COAR Community Framework for Good Practices in Repositories. <https://doi.org/10.5281/zenodo.4110829> [letzter Zugriff 14. Juli 2021].
- DuraSpace** (2016): Design - API Extension Architecture. <https://wiki.duraspace.org/display/FF/Design++api+extension+architecture> [letzter Zugriff 14. Juli 2021].
- Fielding, R.T.** (2000): Architectural Styles and the Design of Network-based Software Architectures. Ph.D. thesis, University of California, Irvine.
- Force 11** (2016): The Fair Data Principles. <https://www.force11.org/group/fairgroup/fairprinciples> [letzter Zugriff 14. Juli 2021].
- Lagoze, C. / Payette, S. / Shin, E.** (2005): Fedora: An architecture for complex objects and their relationships. Berlin/Heidelberg. <http://arxiv.org/ftp/cs/papers/0501/0501012.pdf> [letzter Zugriff 14. Juli 2021].
- Speicher S. / Arwe J. / Malhotra, A.** (2015): Linked Data Platform 1.0. <https://www.w3.org/TR/ldp/> [letzter Zugriff 14. Juli 2021].
- TEI Consortium** (2021): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.2.2. <http://www.tei-c.org> [letzter Zugriff 14. Juli 2021].
- The Consultative Committee for Space Data Systems** (2012): Reference Model for an Open Archival Information System (OAIS). <https://public.ccsds.org/pubs/650x0m2.pdf> [letzter Zugriff 14. Juli 2021].

Semantische Suche mit Word Embeddings für ein mehrsprachiges Wörterbuchportal

Tu, Ngoc Duyen Tanja

tu@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Germany

Meyer, Peter

meyer@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Germany

Das Lehnwortportal Deutsch (LWPD) (Meyer/Eppinger 2019; lwp.ids-mannheim.de) ist ein Online-Informationssystem zu Ent-

lehnungen von Wörtern aus dem Deutschen in andere Sprachen. Es beruht auf einer wachsenden Zahl von lexikographischen Ressourcen zu verschiedenen Sprachen und bietet eine einfache ressourcenübergreifende Suchfunktion an. Das Poster präsentiert eine derzeit in Entwicklung befindliche onomasiologische Suchfunktion für das LWPd.

Ähnliche Projekte, z.B. van der Sijs (2015), nutzen für die Implementierung ihrer semantischen Suche eigens für ihre Datenbasis erstellte Taxonomien von semantischen Feldern. Eine sehr komplexe Open-Source-Taxonomie findet sich beispielsweise auf semdom.org/. Solche Klassifikationen ziehen häufig folgende Probleme nach sich: (a) Aufgrund der inhärenten Vagheit von Definitionen für semantische Felder beruht die Zuordnung von Einzelbedeutungen einer lexikalischen Einheit zu Feldern immer auf einer subjektiven Annotationspraxis, die (b) von dem:der Nutzer:in gewissermaßen rekonstruiert werden muss; (c) bezogen auf die Taxonomie ist es schwierig, einen guten Kompromiss zwischen einfacher Handhabung und Detailgenauigkeit zu finden; (d) grundsätzliche Änderungen an der Taxonomie sind mit einem hohen Aufwand verbunden.

Wir haben einen alternativen Ansatz implementiert, um die oben genannten Probleme anzugehen. Die technische Umsetzung unserer Methode basiert auf den ConceptNet NumberBatch Word Embeddings (CN) (Speer/Chin/Havasi 2017), die auf multilingualen Daten sowie semantischen Beziehungen zwischen Wörtern trainiert sind. Ein im Grunde ähnlicher Lösungsansatz wurde erfolgreich zur Optimierung von Suchmaschinen genutzt (Castro Fernandez et al. 2018; Kuzi/Shtock/Kurland 2016). Da wir keinen Zugriff auf die Korpusdaten haben, die den lexikographischen Ressourcen des LWPd zugrunde liegen, ist es uns nicht möglich, selbst Word Embeddings zu trainieren.

Für die Implementierung der semantischen Suche werden zunächst jedem Wort im LWPd (darunter deutsche Etyma, Lehnwörter, etc.) mindestens ein Wort sowie der/die entsprechende/n Vektor/en aus CN zugeordnet. Für jedes zugeordnete CN-Wort wird angegeben, in welcher semantischen Beziehung (z.B. Synonym, Hyperonym) es zu dem LWPd-Wort steht. Die Zuordnung wird folgendermaßen durchgeführt:

(1) Wenn ein monosemes LWPd-Wort in CN enthalten ist, wird ihm als Default-Wert automatisch dieses CN-Wort zugeordnet.

(2) Wenn ein LWPd-Wort nicht in CN enthalten ist, aber ein LWPd-Wort, das in einer etymologischen oder Derivationsbeziehung zu ihm steht, dann wird ihm als Default-Wert dieses CN-Wort zugeordnet.

(3) Wenn ein LWPd-Wort polysem ist, wird jeder Bedeutung manuell ein CN-Wort zugeordnet.

(4) Wenn ein LWPd-Wort nicht in CN enthalten ist, wird ihm manuell ein semantisch ähnliches CN-Wort zugeordnet.

Ebenfalls können die Default-Werte manuell geändert werden. Homonymen LWPd-Wörtern werden manuell eindeutige CN-Wörter zugeordnet.

Einem LWPd-Wort bzw. einer Bedeutung eines Wortes können auch mehrere CN-Wörter zugeordnet werden, u.a. um Word Embeddings polysemer Wörter zu disambiguieren. Es wird dann eine gewichtete und normierte Summe der Vektoren der einzelnen zugeordneten CN-Wörter zugrunde gelegt. Das Gewicht eines CN-Wortes ergibt sich aus der angegebenen semantischen Beziehung, die auf eine Ganzzahl abgebildet wird.

Die semantische Suche im LWPd läuft aus Nutzer:innensicht folgendermaßen ab: Es steht eine große Anzahl an häufig verwendeten deutschen Wörtern (im Folgenden: Suchschlüssel) zur Auswahl, die mit automatischer Vervollständigung eingegeben werden können. Mit diesen kann der:die Nutzer:in beliebige Aspekte lexikalischer Bedeutung beschreiben. Alle Suchschlüssel sind in

CN enthalten. Somit berechnet sich die semantische Ähnlichkeit der Suchschlüssel und der, den Bedeutungen der LWPd-Wörter zugeordneten, CN-Wörter aus ihrer Kosinus-Ähnlichkeit. Wenn die Kosinus-Ähnlichkeit über einem bestimmten Schwellenwert liegt, werden die entsprechenden LWPd-Wörter in der Suchergebnisliste angezeigt. Die Kosinus-Ähnlichkeiten zwischen den Suchschlüsseln und den CN-Wörtern liegen vorberechnet im LWPd vor.

Um die Qualität der Suchergebnisse zu evaluieren, wurde eine Vorstudie durchgeführt:

(1) Die Bedeutungsangabe (z.B. für das Etymon *Riss*: „*Spalte, Einschnitt, Einriss*“) von jedem Etymon aus dem LWPd wird lemmatisiert und POS-getagged.

(2) Für jedes Etymon E und jedes Lemma L aus einer zugehörigen Bedeutungsangabe wird mit Hilfe von GermaNet (Hamp/Feldweg 1997; Henrich/Hinrichs 2010) ihre semantische Ähnlichkeit gemäß dem Maß in Lin (1998) ermittelt. Budanitsky/Hirst (2006) zeigen, dass die von menschlichen Annotatoren vergebenen semantischen Ähnlichkeitsscores bei englischen Wortpaaren mit den berechneten Scores des Maßes in Lin (1998) eine hohe Korrelation aufweisen. Wir unterstellen, dass Etyma mit den Lemmata ihrer Bedeutungsangaben typischerweise in einer engen semantischen Relation stehen. Daher werden jeweils die Lin-ähnlichsten Synsets von E und L zugrunde gelegt.

(3) Die Kosinus-Ähnlichkeit zwischen dem Vektor von E (z.B. *Riss*) und dem von L (z.B. *Spalte*) wird berechnet.

(4) Die Kosinus-Ähnlichkeit wird mit dem Ergebnis des Ähnlichkeitsmaßes in (2) verglichen.

Es ergibt sich, dass in unserem LWPd-Datensatz die Kosinus-Ähnlichkeit mit dem Ähnlichkeitsmaß von Lin (1998) positiv korreliert ist ($r=0,52$) und für Lin-Ähnlichkeit größer als 0,9 auf fast 0,65 ansteigt.

Diese Evaluation gibt allerdings nur einen ersten Hinweis dazu, dass unser Ansatz vielversprechend ist. Sobald die hier präsentierte semantische Suche implementiert und im LWPd verfügbar ist, soll anhand einer Benutzungsstudie die Qualität der Suchergebnisse untersucht werden.

Bibliographie

Budanitsky, Alexander / Hirst, Graeme (2006): "Evaluating WordNet-based Measures of Lexical Semantic Relatedness", in: *Computational Linguistics* 32(1): 13-47.

Castro Fernandez, Raul / Mansour, Essam / Qahtan, Abdulhakim A. / Elmagarmid, Ahmed / Ilyas, Ihab / Madden, Samuel / Ouzzani, Mouhrad / Stonebraker, Michael / Tang, Nan (2018): "Sleeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery", in: *Proceedings of the 34th International Conference on Data Engineering* 989-1000.

Hamp, Birgit / Feldweg, Helmut (1997): "GermaNet - a Lexical-Semantic Net for German", in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* 9-15.

Henrich, Verena / Hinrichs, Erhard (2010): "GernEdit - The GermaNet Editing Tool", in: *Proceedings of the Seventh Conference on International Language Resources and Evaluation* 2228-2235.

Kuzi, Saar / Shtok, Anna / Kurland, Oren (2016): "Query Expansion Using Word Embeddings", in: *Proceedings of the 25th ACM International Conference on Information and Knowledge Management* 1929-1932.

Lin, Dekang (1998): “An Information-Theoretic Definition of Similarity”, in: *Proceedings of the Fifteenth International Conference on Machine Learning* 296–304.

Meyer, Peter (2019): “Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Builder für lexikografische Property-Graphen“, in: *Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. Konferenzabstracts* 312–314.

Meyer, Peter / Eppinger, Mirjam (2018): “fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data“, in: *Proceedings of the XVIII EURALEX International Congress* 1017–1022.

Speer, Robyn / Chin, Joshua / Havasi, Catherine (2018): “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge“, in: *arXiv:1612.03975 [cs]*.

van der Sijs, Noline (2015): Uitleenwoordenbank, uitleenwoordenbank.ivdnt.org, hosted by the Instituut voor de Nederlandse Taal. Accessed at: <http://uitleenwoordenbank.ivdnt.org/>. (06 April 2021)

Semi-automatische Erschließung von Rechnungsbüchern am Beispiel des Stadtarchivs Leuven

Bigalke, Jan

jbigalke@smail.uni-koeln.de
Universität zu Köln, Deutschland

Drach, Sviatoslav

s.drach@uni-koeln.de
Universität zu Köln, Deutschland

Neuefeind, Claes

c.neuefeind@uni-koeln.de
Universität zu Köln, Deutschland

Historische Rechnungsbücher stellen eine reichhaltige Quelle für geschichtswissenschaftliche Fragestellungen dar. Sie geben einen detaillierten Einblick in die Alltagsgeschichte, indem sie offenlegen, was in welchem Zeitraum gehandelt wurde und von wem.¹ So können bspw. Überschüsse und Mängel im Warenverkehr rekonstruiert, Handelsbeziehungen nachgezeichnet oder auch Rückschlüsse auf die Lebensumstände in bestimmten Zeiten gezogen werden. Z.B. lassen sich aus klösterlichen Rechnungsbüchern Informationen über die Besitzungen eines Klosters gewinnen, etwa wie viele Hörige das Kloster zu einer bestimmten Zeit hatte und welche Abgaben diese zu leisten hatten (siehe z.B. Bruch 2013, Lübbers 2009). Vergleicht man dann die verzeichneten Abgabemengen über Zeiträume hinweg, können so bspw. reiche Ernten oder Missernten identifiziert werden (Gleba 2016: 53ff).

In unserem Beitrag präsentieren wir einen Prototyp zur semi-automatischen Erschließung einer großen Sammlung von Rech-

nungsbüchern des Stadtarchivs Leuven, den wir derzeit im Kontext des Projekts “Itinera Nova” entwickeln.² Im Rahmen des Projekts, das sich seit 2009 der Erschließung der im Stadtarchiv der belgischen Stadt Leuven gesammelten 1127 Schöffengerichte aus den Jahren 1362–1795 mit einem Umfang von ca. 1.000.000 Seiten widmet, wurden seit 2019 auch die dort ebenfalls vorliegenden 457 Rechnungsbücher digitalisiert. Die Erschließung der Schöffengerichte wird derzeit von einer Community mit rund 50 Freiwilligen betrieben und von Archiv-MitarbeiterInnen gemanagt. Das Cologne Center for eHumanities (CCeH) stellt dabei u.a. eine spezifisch für das Projekt entwickelte Web-Plattform für die kollaborative manuelle Erschließung der Schöffengerichte bereit, die Rechnungsbücher hingegen sind bislang noch nicht erschlossen.

Um mit den in Rechnungsbüchern enthaltenen Informationen auch quantitativ arbeiten zu können, müssen sie in strukturierter Form erfasst werden.³ Dabei müssen die enthaltenen Kennzahlen nicht nur transkribiert, sondern auch nach ihrem jeweiligen Typ unterschieden werden, etwa nach verschiedenen Ausgabe- und Einnahmearten, etc. Nicht zuletzt aufgrund des sehr großen Umfangs von mehreren hunderttausend Seiten ist eine vollständige manuelle Transkription kaum zu leisten. Jedoch können Methoden der Digital Humanities hier eine starke Unterstützung bieten. Dabei lässt sich der Umstand nutzen, dass Rechnungsbücher im Vergleich zu anderen historischen Quellen über einen relativ hohen Grad an Strukturierung verfügen (Pollin 2020: 5ff), insbesondere bei Kostenaufstellungen, die in Form von Tabellen festgehalten sind. Hierbei denkt man in der Regel zuerst an vorgedruckte bzw. vorgezeichnete Tabellen mit sichtbaren Linien. Neben diesen gibt es aber auch Rechnungsbücher, die rein konzeptionelle Tabellen ohne diese Linien beinhalten – so auch die Rechnungsbücher des Stadtarchivs Leuven. Ein Workflow zur Informationsextraktion aus Rechnungsbüchern muss demnach neben der Digitalisierung und Transkription auch eine Erkennung der Tabellen und eine anschließende Entity Recognition beinhalten. Für die Transkription existieren schon diverse etablierte Services wie beispielsweise Transkribus.⁴ Diese bieten teilweise auch bereits Lösungen zur Layouterkennung, die auf Tabellen mit Linien ausgerichtet sind. Rein konzeptionelle Tabellen können dagegen nicht effektiv mit diesen Tools erkannt werden.

Für diese Problematik wurde von uns ein Prototyp entwickelt, mit dem auch konzeptionelle Tabellen automatisch erfasst werden können. Genutzt werden hierfür die Koordinaten der einzelnen Zeilen, wie sie auch für die automatische Texterkennung erfasst werden müssen. Die Koordinaten müssen hierbei in XML im PAGE-Schema⁵ vorliegen. Erzeugt werden können diese Daten zum Beispiel durch eine Layoutanalyse mit dem Tool P2PaLA.⁶ Bei den Tabellenzeilen wird davon ausgegangen, dass sie mindestens zweispaltig sind und dass die Spalten einen messbaren Abstand zueinander haben. Daher wird zur Erkennung der einzelnen Zeilen ein speziell trainiertes P2PaLA Modell genutzt, das die zwei Spalten einer Tabellenzeile als zwei einzelne Zeilen erkennt. Das Tool vergleicht hierfür sämtliche benachbarten Textzeilen miteinander. Zuerst werden die Positionen der Textzeilen auf der y-Achse miteinander verglichen. Unterschreitet die Distanz der Textzeilen einen bestimmten Wert, wird davon ausgegangen, dass die Textzeilen auf der gleichen Zeile stehen. Als nächstes werden nun die Positionen der Textzeilen auf der x-Achse miteinander verglichen. Wird hier eine festgelegte Differenz überschritten, ist dies ein Zeichen dafür, dass es sich bei den Textzeilen um zwei Spalten einer Tabelle handelt. Wiederholt sich dieses Phänomen, ist dies ein starker Hinweis darauf, dass es sich bei den Textzeilen um Zeilen einer Tabelle handelt.

Dieses Vorgehen stellt eine robuste Methode dar, um rein konzeptionelle Tabellen eines Rechnungsbuches zu erkennen. Durch das Anpassen der Regeln ist dieses Vorgehen nicht nur auf Rechnungsbücher aus einer Quelle beschränkt, sondern kann leicht an die Gegebenheiten anderer Rechnungsbücher angepasst werden.

Fußnoten

1. Unter anderem beschreibt Paul Kirn in seiner Einführung in die Geschichtswissenschaften Quellen zur Wirtschaftsgeschichte als typisch mittelalterliche Quellenart (Kirn 2019: 33).
2. Das Projekt Itinera Nova ist über den Link <https://www.itinera-nova.be> zu erreichen. Dort können auch die bereits digitalisierten Objekte eingesehen werden.
3. Vgl. dazu v.a. auch die Überlegungen von Georg Vogeler (2015) zu digitalen Rechnungsbüchern.
4. Siehe <https://readcoop.eu/transkribus>
5. PAGE steht für "Page Analysis and Ground-Truth Elements" (Pletschacher/Antonacopoulos 2010).
6. P2PaLA steht für "Page to PAGE Layout Analysis", siehe <https://github.com/lquiroso/P2PaLA>. Eine Dokumentation findet sich bei Quiròs (2018).

Bibliographie

- Bruch, Julia** (2013): Die Zisterze Kaisheim und ihre Tochterklöster. Studien zur Organisation und zum Wirtschaften spätmittelalterlicher Frauenklöster mit einer Edition des Kaisheimer Rechnungsbuches. Berlin.
- Gleba, Gudrun** (2016): Rechnen. Wirtschaften. Aufschreiben. Vernetzte Schriftlichkeit - Wirtschaft und Rechnungsbücher als Quellen klösterlicher Alltagsgeschichte. In: Pätzold, Stefan und Stumpf, Marcus (Hg.): Mittelalterliche und frühneuzeitliche Rechnungen als Quellen der landesgeschichtlichen Forschung. Münster. S. 51-64.
- Kirn, Paul / Leuschner, Joachim** (2019): Einführung in die Geschichtswissenschaften. Berlin / Boston: De Gruyter.
- Lübbers, Bernhard** (2009): Die Ältesten Rechnungen des Klosters Alderbach (1291 - 1373/1409). Analyse und Edition. In: Quellen und Erörterungen zur bayerischen Geschichte: NF 47.3. München: Beck.
- Pletschacher S. / Antonacopoulos A.** (2010): The PAGE (Page Analysis and Ground-truth elements) Format Framework", in: ICPR2010, Istanbul. S. 257-260.
- Pollin, Christopher** (2020): Digitale, formale Methoden und Modelle in den Geschichtswissenschaften. Am Beispiel digital editierter historischer Rechnungsbücher. https://chpollin.github.io/paper/Pollin_MA_2020_Geschichte.pdf [letzter Zugriff 14. Juli 2021].
- Quiròs, Lorenzo** (2018): Multi-Task Handwritten Layout Analysis <https://arxiv.org/pdf/1806.08852.pdf> [letzter Zugriff 14. Juli 2021].
- Vogeler, Georg** (2015): Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert? In: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum / Thomas Stäcker (Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). DOI: 10.17175/sb001_007.

Spotlights

Wie das OpenMethods-Metablog Digital Humanities-Methoden, -Tools und -Toolmaker ins Scheinwerferlicht rückt

Wuttke, Ulrike

ulrike.wuttke@gmx.net
Fachhochschule Potsdam, Germany

Tóth-Czifra, Erzsébet

erzsebet.toth-czifra@dariah.eu
DARIAH-EU, France

Testori, Marinella

testorimarabella@gmail.com
CIRCSE at the Catholic University in Milan, Italy

Horvath, Aliz

aliz.horvath06@gmail.com
Eötvös Loránd University, Hungary

Spence, Paul

paul.spence@kcl.ac.uk
King's College London, UK

Katsiadakis, Helen

hkatsiad@academyofathens.gr
Academy of Athens, Greece

Einführung

Forschungsmethoden und -werkzeuge sind Grundlagen der Forschung in den Digital Humanities. Sie sind genuine Formen der Wissenschaft und als solche nicht-neutrale wissenschaftliche Güter (van Zundert, Antonijević und Andrews 2020), die in der Regel in eine bestimmte epistemische Kultur eingebettet sind und mit spezifischen Forschungsprojekten verbunden sind. Dennoch bleiben ihre Schöpfer sowie die Entscheidungen, die diese im Laufe der Entwicklung treffen, oft in wissenschaftlichen Arbeiten sowie in formalen akademischen Belohnungskriterien unsichtbar (Eve 2020). Ein Kernanliegen von OpenMethods ist es, dies zu verbessern, indem es Werkzeugen und ihren Schöpfer*innen in den Digital Humanities (kurz DH) mehr Anerkennung zukommen lässt und die wissenschaftliche Diskussion um sie stärkt.

Was ist OpenMethods?

Das OpenMethods-Metablog ist eine Plattform, die es ermöglicht, Open-Access-Inhalte zu DH-Methoden und -Werkzeugen in

verschiedenen Formaten und Sprachen zusammenzuführen, um das Wissen um die selbigen zu verbreiten und ihre Anerkennung in der DH-Community und darüber hinaus zu erhöhen. Neben wissenschaftlichen Artikeln und Buchkapiteln umfasst dieser Ansatz verschiedenste Arten von Publikationen im weitesten Sinne und schließt auch solche Inhalte mit ein, die in der formalen Forschungsbewertung meist unsichtbar bleiben, wie z. B. Blogartikel und Preprints oder multimediale Inhalte wie Tutorials, Videos oder Podcasts (Eve 2020).

Der Metablog-Ansatz beinhaltet, dass Mitglieder des OpenMethods-Redaktionsteams bereits veröffentlichte Inhalte, die von *Community Volunteers* vorgeschlagen wurden, sowie Materialien ihrer eigenen Wahl auswählen, um sie auf OpenMethods besonders hervorzuheben. Zu den Themen gehören Beschreibungen von Methoden und Werkzeugen, Werkzeug- und Methodenkritik sowie praktische und theoretische Überlegungen dazu, wie und warum geisteswissenschaftliche Forschung digital betrieben wird und wie der zunehmende Einfluss digitaler Methoden und Werkzeuge die wissenschaftliche Grundhaltung und Praxis der geisteswissenschaftlichen Forschung verändert.

Die OpenMethods-Plattform ist bewusst interdisziplinär und mehrsprachig angelegt, um den Reichtum der DH-Diskurse und wie sie in verschiedenen regionalen, nationalen und sprachlichen Communities Gestalt annehmen aufzuzeigen (Tóth-Czifra / Moranville 2018). Die Gruppe der DH-Expert*innen, bekannt als das OpenMethods Editorial Team, umfasst derzeit 30 Editor*innen aus 14 Ländern, die gemeinsam fast 20 Sprachen abdecken.

Die Nominierung von Inhalten steht jedem offen (über Twitter oder über das Nominierungs-Tool auf der OpenMethods-Plattform) und externe Mitstreiter*innen, wie z. B. Studierende der DH, sind herzlich willkommen, auf der OpenMethods-Website als solche genannt zu werden. In einem zweiten Schritt kommentieren, filtern und kuratieren die Mitglieder des Redaktionsteams die Nominierungen entsprechend der *Guidelines for the Editorial Team* (OpenMethods o. J.). Erfolgreiche Beiträge werden nicht nur auf der Plattform wiederveröffentlicht, sondern auch mit der *Taxonomy of Digital Research Activities in the Arts and Humanities* (TaDiRAH) kategorisiert (Borek et al. 2016, Borek et al. 2021) und durch eine kurze englische Einleitung ergänzt, in der ein*e OpenMethods-Editor*in den Wert und die Relevanz des Beitrags erläutert.

Ins Rampenlicht mit einem Spotlight!

Die Hervorhebung auf OpenMethods ist ein offizielles Zeichen der Anerkennung durch die Expert*innen des Redaktionsteams. Das Korpus der OpenMethods-Inhalte bildet eine kuratierte und kontextualisierte Zusammenstellung von DH-Werkzeugen und -Methoden sowie des Diskurses um diese herum, unabhängig davon, ob die Beiträge Teil der etablierten Kanäle der wissenschaftlichen Kommunikation sind oder nicht.

Als jüngste Weiterentwicklung der Plattform hat das OpenMethods-Redaktionsteam beschlossen, eine neue Serie mit dem Namen *Spotlights* zu starten (OpenMethods 2020). *Spotlights* sind längere Originalbeiträge im Metablog (z. B. Horváth 2020). Diese Beiträge in Interviewform zielen darauf ab, die Menschen hinter den Werkzeugen und Methoden besser sichtbar zu machen und Gespräche über wissenschaftliche Zusammenhänge zu ermöglichen, die in der Regel in bestimmte epistemische Kulturen eingebettet sind und nach Entscheidungen und Wahlmöglichkeiten gestaltet wurden, die für die breitere wissenschaftliche Ge-

meinschaft oft unsichtbar bleiben. Ein zentrales Ziel der *Spotlights*-Reihe ist es, einige dieser epistemischen Überlegungen aufzudecken und den Tool-Machern in den DH mehr Aufmerksamkeit und Anerkennung zu schenken.

In der Posterpräsentation werden folgende Themen angesprochen:

- Wie es die neue *Spotlights*-Serie ermöglicht, die Menschen und epistemischen Überlegungen hinter den Werkzeugen und Methoden der DH sichtbar zu machen
- Welchen Herausforderungen sich das OpenMethods-Team in Bezug auf die unterschiedlichen Ebenen der Nachhaltigkeit gegenübersteht (soziale, infrastrukturelle Elemente sowie angemessene Belohnungs- und Anreizmechanismen für alle Mitwirkenden für ihre Beiträge zur Plattform im Besonderen und zur DH-Methodik im Allgemeinen) (Grant et al. 2020)
- Wie Beiträge, neueste Entwicklungen, Community-Praktiken etc. aus dem deutschsprachigen DH-Diskurs besser auf der Plattform dargestellt werden können

Ziel der Posterpräsentation ist ein breites Feedback von und ein reger Austausch mit den Teilnehmer*innen der DHd-Konferenz. Zu diesem Zweck werden in ihrem Rahmen nicht nur die Ziele und Strategien von OpenMethods erläutert, sondern auch eine interaktive Demo eingeschlossen. Außerdem sollen die Konferenzteilnehmer*innen dazu angeregt werden, dem OpenMethods-Netzwerk beizutreten und es zu erweitern, seine Potenziale für die Weiterentwicklung der eigenen Forschungsmethoden zu erkunden und sich an der Weiterentwicklung der Plattform zu beteiligen.

OpenMethods wurde in Zusammenarbeit mit der DARIAH-Community als ein Ergebnis des DARIAH-Projekts "Humanities at Scale" (Engelhardt et al. 2017) entwickelt.

Bibliographie

Borek, Luise / Dombrowski, Quinn / Perkins, Jody / Schöch, Christof (2016): "TaDiRAH: A Case Study in Pragmatic Classification" in: *DHQ* 10: 1 <http://www.digitalhumanities.org/dhq/vol/10/1/000235/000235.html> [letzter Zugriff 23.11.2021].

Borek, Luise / Hastik, Canan / Khramova, Vera / Illmayer, Klaus / Geiger, Jonathan D. (2021): "TaDiRAH Revised, Formalized and FAIR", in: Schmidt, Thomas / Wolff, Christian (eds.): *Information between Data and Knowledge*. Information Science and its Neighbors from Data Science to Digital Humanities - Proceedings of the 16th International Symposium of Information Science (ISI 2021) Regensburg, Germany, 8th - 10th March 2021 (= Schriften zur Informationswissenschaft 74). Glückstadt: Hülsbusch 321-332. <https://epub.uni-regensburg.de/44951/> [letzter Zugriff 23.11.2021].

Engelhardt, Claudia / Leone, Claudio / Larrousse, Nicolas / Montoliu, Delphine / Moranville, Yoann / Mounier, Pierre / Oltersdorf, Jenny / Ribbe, Paulin / Wuttke, Ulrike (2017): Open Humanities Methods Review Journal (Research Report). DARIAH; TGIR Huma-Num (UMS3598); Göttingen State and University Library. <https://hal.archives-ouvertes.fr/hal-01685852> [letzter Zugriff 23.11.2021].

Eve, Martin Paul (2020): "Violins in the Subway: Scarcity Correlations, Evaluative Cultures, and Disciplinary Authority in the Digital Humanities", in: Jennifer Edmond (eds.): *Digital Technology and the Practices of Humanities Rese-*

arch. Cambridge: Open Book Publishers 105-122. <https://doi.org/10.11647/OBP.0192> [letzter Zugriff 23.11.2021].

Grant, Kaitlyn / Dombrowski, Quinn / Ranaweera, Kamal / Rodriguez-Arenas, Omar / Sinclair, Stéfan / Rockwell, Geoffrey (2020): "Absorbing DiRT: Tool Directories in the Digital Age", in: *Digital Studies / Le Champ Numerique* 10: 1. <https://doi.org/10.16995/dscn.325> [letzter Zugriff 23.11.2021].

Horváth, Alíz (2020): "OpenMethods Spotlights #1: Interview with Hilde De Weerd about MARKUS", in: *OpenMethods* (Blogartikel) <https://openmethods.dariah.eu/2020/10/13/openmethods-spotlights-1-interview-with-hilde-de-weerd-about-markus/> [letzter Zugriff 23.11.2021].

Nyhan, Julianne (2020). "The Evaluation and Peer Review of Digital Scholarship in the Humanities: Experiences, Discussions, and Histories" in: Edmond, Jennifer (eds.): *Digital Technology and the Practices of Humanities Research*. Cambridge: Open Book Publishers 163-182. <https://doi.org/10.11647/OBP.0192> [letzter Zugriff 23.11.2021].

OpenMethods (o. J.), "Guidelines for the Editorial Team", in: *OpenMethods* (Blogartikel) <https://openmethods.dariah.eu/guidelines-for-editorial-team/> [letzter Zugriff 23.11.2021].

OpenMethods (2020), "OpenMethods Spotlights", in: *OpenMethods* (Blogartikel) <https://openmethods.dariah.eu/openmethods-spotlights/> [letzter Zugriff 23.11.2021].

Tóth-Czifra, Erzsébet / Moranville, Yoann (2018): "Leveraging on the power of expert content curation: the OpenMethods metablog: Conference abstract presented at EADH 2018", in: *EADH 2018: Data in Digital Humanities*. EADH, Dec 2018, Galway, Ireland. <https://hal.inria.fr/halshs-02010915/> [letzter Zugriff 23.11.2021].

Zundert, Joris J. van / Antonijević, Smiljana / Andrews, Tara L. (2020). "Black Boxes" and True Colour - A Rhetoric of Scholarly Code", in: Jennifer Edmond (eds.): *Digital Technology and the Practices of Humanities Research*. Cambridge: Open Book Publishers 123-162. <https://doi.org/10.11647/OBP.0192.06> [letzter Zugriff 23.11.2021].

Strukturen und Impulse zur Weiterentwicklung der DHd-Abstracts

Busch, Anna

annabus@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Cremer, Fabian

cremer@ieg-mainz.de
Leibniz-Institut für Europäische Geschichte (IEG)

Lordick, Harald

lor@steinheim-institut.org
Steinheim-Institut (STI), Germany

Mischke, Dennis

dennis.mischke@uni-potsdam.de
Netzwerk Digitale Geisteswissenschaften, Universität Potsdam

Steyer, Timo

t.steyer@tu-braunschweig.de
Universitätsbibliothek Braunschweig

Die DHd-Abstracts

Das Book of Abstracts ist seit 2015 ein fester Bestandteil der DHd-Jahrestagungen und hat sich seither stetig weiterentwickelt. Es spiegelt die Diskurse und Aktivitäten der deutschsprachigen DHd-Community wider und ist daher nicht nur für die Tagung selbst relevant, sondern gleichzeitig ein „Schaufenster“ der aktuellen Forschung in den digitalen Geisteswissenschaften“ (Schöch 2020: V). Die wissenschaftliche Relevanz der Books of Abstracts der DHd-Jahrestagungen bekräftigte zuletzt noch einmal Sahle in seiner Einführung des vorletzten Konferenzbandes (Sahle 2019: V): „Books of Abstracts als durch peer review-Verfahren gefilterte und qualitätsgesicherte Summen der aktuellen Forschungen definieren das Feld, sind ein äußerst nützliches Instrument der Fachkommunikation und wertvolle Dokumente zum Beleg der Entwicklung über die Zeit.“ Trotz der gestiegenen Bedeutung und Weiterentwicklung bleiben die DHd-Abstracts Gegenstand der Diskussion in der DHd-Community, wie zuletzt die Veranstaltung „Die DHd-Abstracts im Zukunftslabor“ auf der vDHd2021 zeigte (Andorfer et al. 2021a).

Die DHd-Abstracts-Community

Eine Gruppe der DHd-Community, zu der die Autor:innen gehören, setzt seit 2018 kontinuierlich Impulse zur Weiterentwicklung der DHd-Abstracts vor allem im Hinblick auf zwei Bereiche: 1) Das Abstract als eigenständige und der Reputation förderliche Publikation und 2) das Abstract als Datenquelle selbstreflexiver Untersuchungsansätze in den DH. Neben konzeptionellen Impulsen (Cremer 2018, Andorfer et al. 2020) werden auch Diskussions- und Hands-On-Formate (Andorfer et al. 2019) organisiert sowie Datenbereitstellung und Softwareapplikationen (Andorfer 2019 und Lordick 2020) umgesetzt. Gemeinsam mit dem DHd-Data Steward und dem Organisationsteam der DHd2022 wurden im Rahmen des „Zukunftslabors“ auf der vDHd2021 die Herausforderungen und Potentiale konkret diskutiert und in Empfehlungen für zukünftige Jahrestagungen übersetzt (Andorfer et al. 2021b).

Die DHd-Abstracts-Task-Force

Die aus den bisherigen Aktivitäten entstandenen Ergebnisse, Ideen und Ziele möchten die Autor:innen durch die Bildung einer festen Task-Force innerhalb des DHd-Verbandes weiter voranbringen. Dabei wird auf eine offene, community-getriebene Arbeitsform, die enge Zusammenarbeit mit dem DHd-Data Steward, eine Anbindung an die DHd AGs Digitales Publizieren und Datenzentren sowie die Kooperation mit dem lokalen Organisationsteam der jeweiligen DHd-Jahreskonferenzen geachtet.¹ Im Fokus stehen die (1) standardisierte und persistente Veröffentlichung von einzelnen Beiträgen zur Jahreskonferenz, die (2) nachhaltige Verfügbarmachung der Beiträge als Daten, die (3) technische und organisatorische Anreicherung der XML-Dateien mit weiteren Informationen sowie deren (4) Standardisierung. Damit verbundene

Herausforderungen sind u.a. die Normdatenversorgung, die Nachnutzung der Daten durch bibliografische Ansetzungen, die Stärkung der Sichtbarkeit sowie der interdisziplinären und internationalen Anschlussfähigkeit.

Die DHd-Abstracts-Entwicklung

Die Abstracts der vergangenen DHd-Jahrestagungen wurden mittlerweile durch den Data Steward des Verbands als einzelne und persistent referenzierbare Beiträge in der *DHd-Community* auf Zenodo sowie als Jahrespakete via Github publiziert (Helling 2021).² Davon ausgehend skizziert dieses Poster die bisherigen Initiativen einer Task-Force, die sich der kurz-, mittel- und langfristigen Aufwertung der Abstracts (einschließlich weiterer Beitragsformen wie Poster) widmet. Zu den aktuell identifizierten Aufgaben zählen:

- Betreuung und Weiterentwicklung von Publikationsprozessen: Die entwickelten Workflows und Pipelines zur Datenvorverarbeitung, zum automatisierten Upload und zur Publikation auf Zenodo soll durch die Task-Force organisatorisch und technisch weiterentwickelt werden. Dadurch werden auch Organisator:innen kommender DHd-Jahrestagungen beim Umgang mit der Publikation von Konferenzabstracts unterstützt. Hierzu gehört u.a. die Aufwertung und nachhaltige Verfügbarmachung von Posterbeiträgen, die im Rahmen der DHd2022-Jahrestagung erstmals in die entwickelten Publikationspipelines integriert und unter einer CC-BY-Lizenz in der *DHd-Community* auf Zenodo publiziert werden.³
- Normdaten-Annotation (ORCID und GND): Dass bei der Einreichung für die DHd2022 erstmals die ORCID-iDs der Autor:innen eingetragen werden können, ist ein erster Schritt. Die konsequente Nutzung von Normdaten, PID-Systemen sowie die Integration der Gemeinsamen Normdatei (GND) können gelingen, wenn entsprechende technische und organisatorische Voraussetzungen im Einreichungs-, Verarbeitungs- und Publikationsprozess geschaffen werden. Gleichzeitig bedarf es der Aufmerksamkeit und der Bereitschaft der DHd-Community zu deren Nutzung, zu der auch dieses Poster beitragen soll.
- Steigerung der Sichtbarkeit von DHd-Abstracts: Durch die Publikation der DHd-Abstracts als einzeln referenzierbare Beiträge in der DHd-Community auf Zenodo wurden Metadateninformationen automatisiert zu aggregierenden Portalen wie bspw. OpenAIRE⁴ weitergegeben. Auch eine Indexierung durch die *dblp computer science bibliography*⁵ ist mittlerweile erfolgt. Eine zentrale Herausforderung stellt noch die Anpassung des TEI-XML-Schemas der DHd-Abstracts dar, damit diese auch im *Index of Digital Humanities Conferences* (Weingart et al. 2019) erfasst und indexiert werden können.

Das DHd-Abstracts-Poster

Mit diesem Poster soll die DHd-Community über die aktuellen Entwicklungen der DHd-Abstracts und die Arbeit der geplanten Task-Force informiert sowie zur Mitgestaltung und Mitarbeit eingeladen werden. Die Postersession dient auch als inhaltlicher Diskussionsraum für Ideen und Bedürfnisse der Community.

Fußnoten

1. Dieses Poster versteht sich als komplementärer Beitrag zur Postereinreichung seitens des DHd Data Steward.
2. Zenodo-Community der Digital Humanities im deutschsprachigen Raum (DHd), <https://zenodo.org/communities/dhd/>.
3. Vgl. Jahrestagung des Verbands „Digital Humanities im deutschsprachigen Raum e.V.“, DHd2022, Call for Papers, <https://www.dhd2022.de/cfp/>.
4. OpenAIRE (Open Access Infrastructure for Research in Europe), <https://www.openaire.eu/>.
5. Vgl. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum bei der dblp computer science bibliography: <https://dblp.org/db/conf/dhd/index.html>.

Bibliographie

Andorfer, Peter (2019): *dhd-boas-app*. <https://dhd-boas-app.acdh-dev.oeaw.ac.at/>.

Andorfer, Peter / Busch, Anna / Cremer, Fabian / Helling, Patrick / Henrich, Andreas / Henrich / Lordick, Harald / Mischke, Dennis / Steyer, Timo (2021a): *Die DHd-Abstracts im Zukunftslabor*. vDHd2021 – Experimente (blog). <https://vdhd2021.hypotheses.org/137> [letzter Zugriff 6. Juli 2021].

Andorfer, Peter / Busch, Anna / Cremer, Fabian / Helling, Patrick / Henrich, Andreas / Henrich / Lordick, Harald / Mischke, Dennis / Steyer, Timo (2021b): *Bericht zur vDHd2021-Veranstaltung: Zukunftslabor DHd-Abstracts*. DHd-Blog (blog). <https://dhd-blog.org/?p=15980> [letzter Zugriff 6. Juli 2021].

Andorfer, Peter / Cremer, Fabian / Steyer, Timo (2019): „DHd 2019 Book of Abstracts Hackathon“, in: *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*. Frankfurt am Main. <https://doi.org/10.20375/0000-000B-D512-0> [letzter Zugriff 6. Juli 2021].

Andorfer, Peter / Cremer, Fabian / Steyer, Timo (2020): „Abstract Enhancement. Potentiale der DHd-Konferenzabstracts als Daten/Publikation“, in: *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*. Paderborn. <https://doi.org/10.5281/zenodo.3666690> [letzter Zugriff 6. Juli 2021].

Cremer, Fabian (2018): „Nun sag, wie hältst Du es mit dem Digitalen Publizieren, Digital Humanities?“. *Digitale Redaktion* (blog). <https://editorial.hypotheses.org/113> [letzter Zugriff 6. Juli 2021].

Helling, Patrick (2021): „DHd-Konferenzen 2014-2020 – einzelne Abstracts in DHd-Community auf Zenodo publiziert | DHd-Blog“. *DHd-Blog* (blog). <https://dhd-blog.org/?p=15599> [letzter Zugriff 6. Juli 2021].

Lordick, Harald (2020): *DH(d) Konferenzbeiträge*. <http://www.steinheim-institut.de/dhd/> [letzter Zugriff 6. Juli 2021].

Sahle, Patrick (2019): *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*. Frankfurt am Main: Zenodo. <https://doi.org/10.5281/zenodo.2596095> [letzter Zugriff 6. Juli 2021].

Schöch, Christof (2020): *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*. Paderborn: Christof Schöch. <https://doi.org/10.5281/zenodo.3666690> [letzter Zugriff 6. Juli 2021].

Weingart, Scott B. / Eichmann-Kalwara, Nickoal / Lincoln, Matthew (2020): *The Index of Digital Humanities Conferences*.

Carnegie Mellon University. <https://dh-abstracts.library.cmu.edu/> [letzter Zugriff 6. Juli 2021].

Studying the ephemeral, cultures of digital oblivion Identifying patterns in Instagram Stories.

Achmann, Michael

michael.achmann@sprachlit.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg,
Germany

Hampel, Lisa

Lisa.Hampel@stud.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg,
Germany

Asabidi, Ruslan

Ruslan.Asabidi@stud.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg,
Germany

Wolff, Christian

Christian.Wolff@sprachlit.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg,
Germany

Introduction

While some argue that due to technology remembering was the default today, forgetting the exception (Mayer-Schönberger, 2011: 2) and, in a subtle revolution, remembering swaps with forgetting (Assmann, 2016: 205–208), Instagram stories turn this concept upside down, being available for 24 hours only. As 2021 marks the year of Germany's federal election we're curious to see how politicians and political parties use this ephemeral medium in a total-recall world of social media.

While Instagram use for past elections, e.g. the 2014 Swedish (Filimonov et al., 2016), 2016 U.S (Muñoz and Towner, 2017; Towner and Muñoz, 2020), 2017 German (Voigt and Seidenglanz, 2019) and 2018 Swedish election (Grusell and Nord, 2020), and similarly Justin Trudeau's use of Instagram (Lalancette and Raynauld, 2019), have been investigated, stories have mostly been omitted.

We present results from our ongoing project, collecting and analyzing ephemeral Instagram stories. We collect stories by political parties, their front-runners and, additionally, stories by influencers and members of the public as baseline material of Instagram use. Bainotti et al. (2020) were able to outline two grammars for Instagram stories, one for interaction, one for documentation. Further they observed users to follow certain aesthetics and norms of influencer marketing. Their work serves as a guideline for our pro-

ject to identify typical patterns of Instagram stories for each group, curious to see the aesthetics and norms used by politicians and parties. Our project is guided by the following research questions:

1. How do the prototypical Instagram stories of candidates and parties of the 2021 German federal election during the final phase look like?
2. How do their stories compare to other user groups?
3. How do ephemeral stories of politicians and parties compare to their permanent posts?

Data Collection

To answer ethical challenges (Bainotti et al., 2020; Franzke et al., 2020; McCrow-Young, 2021; Zimmer, 2010), we decided to acquire our corpus employing different methods for different target groups: a) political actors and influencers whose profiles are public, b) members of the public, whose profiles may not be public, and thus their stories are intended for a small group of people only, while they expect the ephemeral content to expire. To answer this challenge, we collected our corpus: a) using a custom-made scraper, b) crowdsourcing codings of stories through a browser plugin.

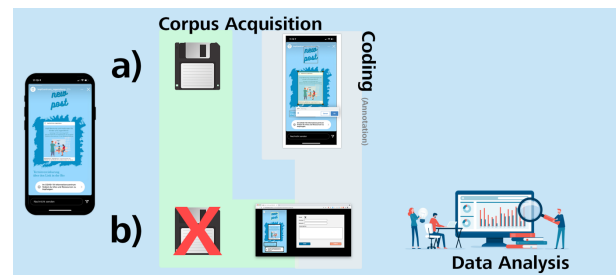


Abb. 1: The two corpus collection methods. Source: own work; floppy disk: pixabay.com; phone frame & data analysis: adobe stock; story: @impfzentrum_regensburg on instagram.

The first method is similar to Bainotti et al.'s (2020): We access Instagram using a custom script to emulate a user accessing the platform. For method b) we created a plugin for the chrome browser, allowing participants to code stories of the profiles they follow from within the browser. While method a) allows us to save and collect the corpus of stories as image or video files and their meta data, method b) only saves the users' codings and some meta data. Scraping took place from Sep 13 through 26, 2021, counting 2208 stories. A first batch of crowd-annotations was collected in November, more codings using an improved plugin are planned.

Data Analysis

We are going to conduct a content analysis as outlined by Rose (2016: ch. 5), and used in previous work (e. g. Bainotti et al., 2020; Towner and Muñoz, 2020): The first step is going to be an initial open coding of the collected images by the authors, supported by codes from the literature. Inspired by Bainotti et al. coding will focus on the content as in "what does the image show", as well as the composition, as in "which Instagram specific affordances were used" (e.g. polls, questions, countdown, gif, ...). Further narrative style and context of use are added. Coding of stories by members

of the public will be the most challenging, as our approach means we must rely on our crowd supporters' annotations using predefined codes. They will be evaluated by coding predefined public profiles. Data analysis will start in early 2022.

Discussion

Our research expands the methodological framework for collecting ephemeral content on social media platforms. By introducing the annotation plugin as a bypass for ethical concerns collecting private Instagram stories, we hope to inspire future research analyzing personal content on social media, especially ephemeral formats. If successful, our analysis will improve our understanding of Instagram stories and their use by political protagonists during the hot phase of an election campaign. Future work could compare the content of stories and posts to the users' content on other platforms like Twitter. Further we would like to train a model for automated coding and large-scale analysis.

Bibliography

Assmann, Aleida (2016): *Formen des Vergessens*. Göttingen: Wallstein Verlag.

Bainotti, Lucia / Caliandro, Alessandro / Gandini, Alessandro (2020): "From archive cultures to ephemeral content, and back: Studying Instagram Stories with digital methods", in: *New Media & Society* 23(12): 3656-3676.

Filimonov, Kirill / Russmann, Uta / Svensson, Jakob (2016): "Picturing the Party: Instagram and Party Campaigning in the 2014 Swedish Elections", in: *Social Media + Society* 2(3) 10.1177/2056305116662179.

franzke, alineshakti / Bechmann, Anja / Zimmer, Michael / Ess, Charles / Association of Internet Researchers (2020): "Internet research: ethical guidelines 3.0: association of internet researchers." <https://aoir.org/reports/ethics3.pdf>. [last access: 12-01-2021]

Grusell, Marie / Nord, Lars (2020): "Not so Intimate Instagram: Images of Swedish Political Party Leaders in the 2018 National Election Campaign", in: *Journal of Political Marketing* 10.1080/15377857.2020.1841709.

Lalancette, Mireille / Raynauld, Vincent (2019): "The Power of Political Image: Justin Trudeau, Instagram, and Celebrity Politics", in: *The American behavioral scientist* 63(7): 888-924.

Mayer-Schönberger, Viktor (2011): *Delete: The Virtue of Forgetting in the Digital Age*. Princeton, N.J.: Princeton University Press.

McCrow-Young, Ally (2021): "Approaching Instagram data: reflections on accessing, archiving and anonymising visual social media", in: *Communication Research and Practice* 7(1): 21-34.

Muñoz, Caroline Lego / Towner, Terri L (2017): "The Image is the Message: Instagram Marketing and the 2016 Presidential Primary Season," *Journal of Political Marketing* 16(3-4): 290-318.

Rose, Gillian (2016): *Visual Methodologies: An Introduction to Researching with Visual Materials*. London: SAGE Publications.

Towner, Terri / Muñoz, Caroline Lego (2020): "Instagramming Issues: Agenda Setting During the 2016 Presidential Campaign", in: *Social Media & Society* 6(3) 10.1177/2056305120940803.

Voigt, Mario / Seidenglanz, Rene (2019): "Trendstudie Digital Campaigning in der Bundestagswahl

2017 - Implikationen für Politik und Public Affairs" <https://www.medianet-bb.de/wp-content/uploads/2018/01/quadrige-digital-campaigning-studie-btw2017.pdf> [last access 12-01-2021].

Zimmer, Michael (2010): "'But the Data is Already Public': On the Ethics of Research in Facebook", in: *Ethics and information technology* 12(4): 313-325.

Szenario-basierte Planung eines semantischen Digitalisierungsvorhabens in der digitalen Geschichtswissenschaft

Scheltjens, Werner

werner.scheltjens@uni-bamberg.de

Digitale Geschichtswissenschaften, Otto-Friedrich-Universität Bamberg, Germany

Schlieder, Christoph

christoph.schlieder@uni-bamberg.de

Kulturinformatik, Otto-Friedrich-Universität Bamberg, Germany

Die Retrodigitalisierung von nicht urheberrechtlich geschützten Bibliothekssammlungen hat zahlreiche gedruckte Texte als digitale Quellen verfügbar gemacht und fordert HistorikerInnen heraus, sich mit neuen Methoden der Nutzung dieser Quellen vertraut zu machen (Paju et.al. 2020). Für die historisch metrologische Forschung sind Nachschlagewerke und Lexika zu Handel und Gewerbe von besonderem Interesse, die im 18. und 19. Jahrhundert versuchten, das Wissen ihrer Zeit zu systematisieren. Insbesondere in der Übergangszeit zum metrischen System (bis etwa 1870) waren diese Werke weit verbreitet. Sie kombinierten eine positivistisch anmutende Neugier auf Maße, Gewichte und Münzen in aller Welt mit dem Versuch den neuen Anforderungen einer zunehmend auf Standardisierung und Systematisierung hinarbeitenden Gesellschaft gerecht zu werden (Kramper 2019). Wie alle historischen Texte sind auch metrologische Nachschlagewerke und Lexika Zeugen einer Epoche. In digitalisierter Form liegen sie nunmehr als neue Quellen für historische Untersuchungen vor.

Ein bekanntes Beispiel ist das angesehene und bis heute häufig zitierte *Vollständige Taschenbuch der Münz-, Maass- und Gewichtsverhältnisse* von Christian und Friedrich Noback (Noback & Noback 1850; Denzel 2002; Witthöft 2018). Die Bayerische Staatsbibliothek hat dieses Nachschlagewerk digitalisiert und stellt zusammen mit den Scans im PDF-Format auch eine Textdatei mit dem Volltext aus der OCR für die nicht-kommerzielle Nutzung zur Verfügung. Die Retrodigitalisierung hat eine digitale Version des Vollständigen Taschenbuchs hervorgebracht, die den lesenden Zugriff auf einzelne Artikel des Lexikons erheblich vereinfacht. Erkenntnisse über die Zusammensetzung historischer metrologischer Systeme ergeben sich aber vor allem aus der vergleichenden Auswertung einer großen Zahl von Lexikonartikeln, etwa aus allen Artikeln zu Handelsorten einer Wirtschaftsregion mit den darin aufgeführten Getreidemaßen. Solche eine Auswertung lässt sich durch Blättern oder Volltextsuche im Digitalisat nur

sehr mühsam durchführen. Auch wenn Nachschlagewerke über die historische Metrologie, wie das Vollständige Taschenbuch, in digitaler Form vorliegen, steht die semantische Erschließung ihrer Inhalte noch aus.

Wir argumentieren, dass für bestimmte, konkret identifizierbare Forschungsfragen der historischen Metrologie eine zweite oder semantische Retrodigitalisierung unabdingbar ist. Diese ergänzt die erste, größtenteils automatisch realisierte Retrodigitalisierung und strebt die Extraktion und explizite Modellierung der semantischen Struktur des enzyklopädischen Wissens an. Für die Planung der auf die Explizierung semantischer Beziehungen zielenden Digitalisierung historischer Quellen steht im Prinzip das allgemeine Methodeninventar der ontologischen Modellierung zur Verfügung. An erster Stelle sind hier szenario-basierte Methoden zu nennen, die den Planungsprozess an sogenannten Kompetenzfragen orientieren, d.h. einem Katalog derjenigen Fragen, die FachanwenderInnen anhand der Modellierung untersuchen und beantworten wollen (Kendall, McGuinness, 2019). Dieses bewährte szenario-basierte Vorgehen haben insbesondere Lodi et al. (2017) sowie Carriero et al. (2021) auf Fragestellungen der Digital Humanities übertragen. Konkret waren Metadaten italienischer Gedächtnisinstitutionen semantisch zu modellieren, was durch Abbildung der Kompetenzfragen auf Ontologieentwurfsmuster gelöst wurde.

Wir zeigen, dass sich dieser Ansatz zwar grundsätzlich, im Detail aber eben doch nur begrenzt, auf die beschriebene Problemstellung der digitalen Geschichtswissenschaft anwenden lässt. Eine erste Schwierigkeit ergibt sich aus dem Unterschied zwischen Modellierungsvorhaben, die sich auf Metadaten beziehen und solchen, die sich auf (Primär-)Daten stützen. Am Beispiel der Planung der semantischen Modellierung des Vollständigen Taschenbuchs der Nobacks führen wir vor, dass anhand der Kompetenzfragen zunächst beurteilt werden muss, welche semantischen Beziehungen explizit zu repräsentieren sind und welche durch Anfragen abgeleitet werden sollen. Auch stellt sich heraus, dass Domänenontologien, die nicht dem Umfeld der DH-Forschung entstammen, z.B. die von Martín-Recuerda et al. (2020) beschriebenen metrologischen Ontologien für die Naturwissenschaften, kaum direkt verwendet werden können. Wir stellen einen Prozess für die Planung semantischer Modellierungsvorhaben vor, der den von Lodi et al. (2017) beschriebenen in mehreren Punkten variiert bzw. detailliert.

Anhand eines Beispiels aus dem Bereich der historischen Metrologie werden Arbeitsabläufe für das Explizieren von semantischen Beziehungen in historischen Texten vorgestellt und diskutiert. Ziel des semantischen Digitalisierungsvorhabens ist es, einen Beitrag zur Erforschung von historischen Texten zu liefern, indem zwischen der Planung der „ersten“ und der Planung der „zweiten“ oder semantischen Retrodigitalisierung unterschieden wird und Vorschläge für die systematische Erschließung der semantischen Ebene digitalisierter historischer Texte formuliert werden.

Bibliographie

Carriero, Valentina Anita / Gangemi, Aldo / Mancinelli, Maria Letizia / Nuzzolese, Andrea Giovanni / Presutti, Valentina / Veninata, Chiara (2021): “Pattern-based Design Applied to Cultural Heritage Knowledge Graphs”, in: *Semantic Web* 12: 313 – 357. 10.3233/SW-200422

Denzel, Markus A. (2002): “Handelspraktiken als wirtschaftshistorische Quellengattung vom Mittelalter bis in das frühe 20. Jahrhundert. Eine Einführung” in: Denzel, Markus A. / Hocquet, Jean-Claude / Witthöft, Harald (eds.): *Kaufmannsbücher*

und Handelspraktiken vom Spätmittelalter bis zum beginnenden 20. Jahrhundert — Merchant's Books and Mercantile Pratiche from the Late Middle Ages to the Beginning of the 20th Century. Stuttgart: Steiner Verlag 11-45.

Kendall, Elisa F. / McGuinness, Deborah L. (2019): *Ontology engineering*. (= Synthesis Lectures on The Semantic Web: Theory and Technology, Lecture 18). [California]: Morgan and Claypool. 10.2200/S00834ED1V01Y201802WBE018

Kramper, Peter (2019). *The Battle of the Standards. Messen, Zählen und Wiegen in Westeuropa, 1660-1914*. Berlin / Boston: De Gruyter.

Lodi, Giorgia / Asprino, Luigi / Nuzzolese, Andrea Giovanni / Presutti, Valentina / Gangemi, Aldo / Recupero, Diego Reforgiato / Veninata, Chiara / Orsini, Annarita (2017): “Semantic Web for Cultural Heritage Valorisation”, in: Hai-Jew, Shalin (ed.): *Data Analytics in Digital Humanities*. Multimedia Systems and Applications. Springer: Cham 3-37. https://doi.org/10.1007/978-3-319-54499-1_1

Martín-Recuerda, Francisco / Walther, Dirk / Eisinger, Siegfried / Moore, Graham / Andersen, Petter / Opdahl, Per-Olav / Hella, Lillian (2020): “Revisiting Ontologies of Units of Measure for Harmonising Quantity Values – A Use Case”, in: Pan, Jeff Z. / Tamma, Valentina / d'Amato, Claudia / Janowicz, Krzysztof / Fu, Bo / Polleres, Axel / Seneviratne, Oshani / Kagal, Lalana (eds.): *The Semantic Web – ISWC 2020*. (= Lecture Notes in Computer Science, vol. 12507). Springer: Cham 551-567. https://doi.org/10.1007/978-3-030-62466-8_34

Noback, Christian / Noback, Friedrich (1850): *Vollständiges Taschenbuch der Münz-, Maass-, und Gewichtsverhältnisse, der Staatspapiere, des Wechsels- und Bankwesens, und der Usanzen aller Länder und Handelsplätze*. Leipzig: F.A. Brockhaus.

Paju, Petri / Oiva, Mila / Fridlund, Mats (2020): “Digital and distant histories: Emergent approaches within the new digital history”, Fridlund, Mats / Oiva, Mila / Paju, Petri (eds.): *Digital histories: Emergent approaches within the new digital history*. Helsinki: Helsinki University Press 3-18. <https://doi.org/10.33134/HUP-5-1>

Witthöft, Harald (2018): “Numerical Communication in Intercontinental Trade and Monetary Matters: Coins and Weights in China and East Asia in Merchants' Pocketbooks and Commercial Guides (16th–19th Centuries)”, in: Theobald, Ulrich / Cao, Jin (eds.): *Southwest China in a Regional and Global Perspective (c. 1600-1911)*. Leiden / Boston 225-290. https://doi.org/10.1163/9789004353718_009

Text-Bild-Gefüge

Digital Humanities und der Diskurs der Moderne

Klemstein, Franziska

f.klemstein@gmail.com

Bauhaus-Universität Weimar, Germany

Die DH gelten als interdisziplinäres Feld par excellence. Sie bringen geisteswissenschaftliche Disziplinen, wie etwa die Linguistik oder die Geschichtswissenschaft, in enge Verbindung mit Disziplinen wie Computer Science und Information Science. Das daraus entstehende Potential zur übergreifenden Zusammenarbeit zwischen einzelnen geisteswissenschaftlichen Disziplinen wird bereits vielfach genutzt. Dennoch stehen textfokussierte und bild-

zentrierte Projekte oft unvermittelt nebeneinander. Gleichwohl haben multimodale Untersuchungsmethoden derzeit Konjunktur (Barman et al 2021).

Zumeist findet dort jedoch eine Fokussierung auf sehr spezifische Einzelaspekte statt. So fokussierten Barman et al auf das Training weniger, exemplarisch ausgewählter Klassen, um ihre Netzwerke auf das Erkennen von Todesanzeigen innerhalb eines schweizerisch-luxemburgischen Korpus historischer Zeitungen zu trainieren. Bilder wurden hierbei als auffällige Layout-Bereiche (areas of interests) innerhalb eines strukturierten und sich wiederholenden Gesamtlayouts verstanden, nicht jedoch als eigenständige Bilder innerhalb eines Gesamtgefüges aus Text und Bild.

Unser Forschungsprojekt nimmt dies zum Anlass um Texte und Bilder als gleichwertige Elemente sowohl im Bereich CV als auch im Bereich des NLP zu untersuchen. Innerhalb unseres ersten Projektabschnitts wird jedoch vor allem auf die Layout-Analyse und die Untersuchung des Verhältnisses von Text und Bild zueinander fokussiert. Konkret sollen dabei im ersten Teilprojekt u.a. folgende Fragen beantwortet werden:

- 1.) Wie entwickelt sich in unserem Zeitschriftenkorpus die strukturelle Anordnung von Text und Bild?
- 2.) Welche Rolle spielen dabei Bildunterschriften und wie wird im Text auf die Bilder Bezug genommen?
- 3.) Welche Arten von Bildern und welche Bildinhalte werden in die Texte einbezogen und wann dominieren welche Bildarten in welchen Textgattungen bzw. Untersuchungsdomänen?

Diese hier formulierten Fragen sollen eine systematische Untersuchung größerer Text-Bild-Korpora ermöglichen und damit das Text-Bild-Verhältnis innerhalb der sogenannten Moderne untersuchen. Unsere Hypothese lautet, dass Veröffentlichungen, wie zum Beispiel die Bauhausbücher als Ausdruck und Zeugnis des Diskurses der Moderne verstanden werden, da sie sich in ihrem Text-Bild-Verhältnis in einem Spektrum bewegen, das zwischen einer Verwissenschaftlichung künstlerischer Publikationen bzw. der Kunstpraxis einerseits steht und andererseits experimentelle Layoutversuche offenbart, die die Zweidimensionalität der Seite grundlegend in Frage stellt. Ob diese Hypothese zutrifft und tatsächlich als charakteristisch oder prägend für den Zeitraum der Moderne bezeichnet werden kann, muss durch eine möglichst breite und dennoch tiefgehende Analyse verifiziert werden.

Um diese Hypothese nicht allein in Bezug auf die Bauhausbücher, sondern in einem größeren Kontext mit Blick auf die Veränderungen von Text-Bild-Gefügen innerhalb des Untersuchungszeitraums (1880–1930) untersuchen zu können, umfasst unser Korpus Publikationen (Zeitschriften, Monographien u.v.m.) aus verschiedenen Themenbereichen. Exemplarisch sind hier zu nennen: die Bauhausbücher, die Zeitschrift „Das Kunstgewerbe“, die „Fliegenden Blätter“, die „Zeitschrift für Psychologie und Physiologie der Sinnesorgane“ oder auch das „Centralblatt der Bauverwaltung“.

Neben der Vorstellung der Untersuchungsdomänen und Textkorpora soll ebenso das methodische Vorgehen erläutert werden. So soll unter anderem der Annotationsprozess zur Erstellung des Datensatzes erläutert werden, auf dessen Grundlage dann zu einem späteren Zeitpunkt mithilfe des PubLayNet-Datensatzes (Zhong et al 2019) ein Mask R-CNN-Transferlernen (He et al 2017) für ein 19-Klassen-Problem (siehe Abb.1) durchgeführt werden soll. Da der Annotationsprozess selbst jedoch bereits zahlreiche nicht-triviale Entscheidungen in sich birgt, soll dieser hier besonders ausführlich beschrieben werden. Hierbei soll einerseits der Entscheidungsprozess für das Computer Vision Annotation Tool (CVAT) (Sekachev et al 2019) und gegen das Aletheia Document Analysis System (Clausner et al 2011) dargelegt werden

sowie die Definition der zu annotierenden Klassen, die sich auch im Inter Annotator Agreement widerspiegeln und andererseits erste Analyseergebnisse vorgestellt werden, die auf der Grundlage der ersten Annotationen bereits möglich sind.

Die in CVAT annotierten Digitalisate können in verschiedenen Formaten exportiert und auf diese Weise für verschiedene weitere Anwendungsmöglichkeiten nutzbar gemacht werden. Innerhalb unseres Projektes wird das Format „COCO 1.0“ genutzt, d.h. das nach dem Export ein Ordner mit den Digitalisaten sowie einer json-Datei vorliegen. Diese Datei bildet die Grundlage für erste oberflächliche Analysen, die jedoch bereits erste aufschlussreiche Anhaltspunkte für die Untersuchung verschiedener Textgattungen im Zeitraum von 1880 bis 1930 zulassen. Zu diesen gehören unter anderem folgende Abfragen:

- a) Welche Annotationsklassen sind auf welchen Seiten zu finden?
- b) Welche Seiten haben wie viele Textfelder bzw. Textspalten?
- c) Welche Seiten haben wie viele Abbildungen?
- d) Welche Seiten haben Bildunterschriften?

Ebenso kann das Verhältnis von der spezifischen Klasse zur Gesamtseite oder im Verhältnis zu anderen Klassen ausgewertet werden und vieles mehr.

Durch die Beantwortung dieser Fragen können nicht nur erste Annahmen zur strukturellen Anordnung von Text und Bild systematisch analysiert und hinterfragt werden, sondern zugleich eine Vergleichsebene zwischen verschiedenen Untersuchungsdomänen vorgenommen werden. Im Zentrum steht dabei die Frage, inwiefern sich die von Daston und Galison vorgenommene Ausweitung der Foucaultschen Diskursanalyse auf die (wissenschafts-)historische Untersuchung von Bilddiskursen (Daston/Galison 2007) durch den Gebrauch von verschiedenen digitalen Technologien konkretisieren und damit präzisieren lässt.

Labelclass	Short-Definition
image	is described by 2 categories: a) media type, b) content type
logo	is a graphically designed sign
editorial note	is every information that originates from the editor/publisher
decoration	is a artistic or structuring element
frame	is the border of an advertisement, image, text or page
footer	is a specific area of a page and follows the main text, or body
footnote	is a note or further explanation by the author
page number	is a consecutive numbering
header	is detached from the main text at the top of the text page
dropped capital	is a letter at the beginning of a word that is larger than the rest of the text and is often ornately decorated
heading	is the name for a book or an article within a journal.
subheading	is the name for a section or the sub-head of the heading
author	is the creator or originator of any written work
text	is a thematically and / or functionally oriented, coherent linguistic or linguistic-figurative complex
noise	is an element subsequently added to the published work
column title	is a text usually placed at the top - in the header - of the page or column
advertisement	is a page-region and can include decorations, text, images, logos, frames
caption	is either an explanatory text nearby the image/table eventually including a reference number; a reference number nearby the image/table without any explanatory text; a reference number within the main text or a reference number within the main text, referencing to multiple images
table	is an ordered arrangement of information or data, typically in rows and columns, or possibly in a more complex structure

Abb. 1.

Bibliographie

Barman, R. / Ehrmann, M. / Clematide, S. / Oliveira, S. A. / Kaplan, F.: Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers, in: Journal of Data Mining & Digital Humanities, DOI: 10.46298/jdmhdh.6107 , arXiv:2002.06144.

Clausner, C. / Pletschacher, S. / Antonacopoulos, A.: Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments, in: International Con-

ference on Document Analysis and Recognition, 2011, pp. 48-52, doi: 10.1109/ICDAR.2011.19.

Daston, L. / Galison, P.: Objektivität, Frankfurt/Main 2007.

He, K. / Gkioxari, G. / Dollár, P. / Girshick, R.: Mask R-CNN, arXiv:1703.06870.

Sekachev, B. / Zhavoronkov, A. / Manovich, N.: Computer Vision Annotation Tool: A Universal Approach to Data Annotation, 2019, <https://software.intel.com/content/www/us/en/develop/articles/computer-vision-annotation-tool-a-universal-approach-to-data-annotation.html> [letzter Zugriff: 13.07.2021].

Zhong, X. / Tang, J. / Yepes, A. J.: PubLayNet: largest dataset ever for document layout analysis, arXiv:1908.07836.

The Digitized minutes of the Habsburg Governments, 1848-1918

Fischer-Nebmaier, Wladimir

wladimir.fischer@oeaw.ac.at

Österreichische Akademie der Wissenschaften, Austria

Kurz, Stephan

stephan.kurz@oeaw.ac.at

Österreichische Akademie der Wissenschaften, Austria

Lein, Richard

richard.lein@oeaw.ac.at

Österreichische Akademie der Wissenschaften, Austria

Dieses Poster beschreibt den Stand der Hybridedition „Edition der Ministerratsprotokolle Österreichs und der Österreich-Ungarischen Monarchie 1848-1918“, die ein traditionsreiches Editionsprojekt ins Web gebracht hat.

1) Die Ausgangslage

Die Habsburgermonarchie hatte von 1848 bis 1918 vier verschiedene Regierungsinstitutionen, drei davon existierten parallel ab dem Jahr 1867. Davor gab es eine Regierung, den österreichischen Ministerrat (ÖMR), danach gab es einen ungarischen Ministerrat (UMR) und einen Ministerrat für die restlichen Länder (CMR), sowie einen gemeinsamen Ministerrat (GMR). Große Teile der Textgrundlage des CMR sind Brandakten und nur teilweise überliefert (Justizpalastbrand 1927).

Das erklärt einen Teil der Komplexität des Unterfangens, die Ministerratsprotokolle der Habsburgermonarchie nach 1848 edieren zu wollen.

Hinzu kommt, dass die Dokumente dieser Regierungen seit den späten 1960er Jahren von verschiedenen Institutionen historisch-kritisch ediert wurden und werden: von der ungarischen Akademie der Wissenschaften und von einem österreichischen Institut, das in unterschiedlichen Rechtsformen existierte und inzwischen im Bereich Digitale Historiographie und Editionen des Institute for Habsburg and Balkan Studies der Österreichischen Akademie der Wissenschaften integriert ist.

Daraus ergaben sich drei unterschiedliche Textbasen, die sich in drei verschiedenen Zuständen befanden, als die TEI-Digitalisierung begonnen wurde:

a) Die Printeditionen der ÖMR 1848–1867 und der GMR 1867–1918 lagen in Form von PDF-Digitalisaten vor, die Worddateien dazu waren nur noch teilweise vorhanden, der Rest wurde extern transkribiert, sodass die Textbasis schließlich minder zuverlässiges XML war.

b) die CMR 1867–1918 sind derzeit im Entstehen begriffen. Deshalb kann dieser Teil laufend als born-digital TEI-Edition erscheinen (Band I bis III befinden sich in unterschiedlichen Übergangsstadien).

d) nachträglich kam durch Kooperation mit der Ungarischen Akademie der Wissenschaften noch deren Publikation navigierbarer Images der Originalprotokolle auf Ungarisch hinzu, die kurzerhand mit in das Projekt einbezogen wurden.

2) Die Digitaledition: Stand der Dinge

Die digitale Edition ist in einer web-app auf mrp.oeaw.ac.at verfügbar. Sie umfasst derzeit 36 Bände, sie enthalten 2816 Protokolle mit 11825 Tagesordnungspunkten aus 61 Regierungsjahren.

2.1) Alles an einem Ort

Als Prinzip sind alle Protokolle über dieselben Logiken gleichberechtigt zugänglich. Jedes Protokoll ist über sein Datum im Kalender ansteuerbar. Mit Ausnahme der Images der UMR ist die Grundeinheit der Tagesordnungspunkt. Auf den Tagesordnungspunkten beruhen alle Abfragen und Darstellungen (Protokollansicht, Bandansicht, Suchergebnisse etc.).

Um Lesegewohnheiten entgegenzukommen, sind alle Protokolle auch als Bände wie in der Printedition darstellbar. Soweit vorhanden ist jedes XML-Protokoll seitenweise mit einem PDF der Print-Edition verlinkt. Da einige Bände der UMR-Edition erschienen sind, wird auf diese zumindest verwiesen.

Die Edition zeichnet sich durch einen ausführlichen Kommentar und instruktive historische Einleitungen aus. Diese werden durch eigene Verlinkungen in allen Protokollansichten präsent und zugänglich gemacht.

2.2) Angereicherte Daten

Alle Texte werden je nach Zustand mit strukturierten Daten angereichert. Das aus den Print-Bänden generierte TEI-XML verfügt über Fußnoten und editorische Anmerkungen zu Textbereichen, Querverweise zu anderen Tagesordnungspunkten mit Mouse-over-Vorschau, strukturierte bibliografische und archivalische Einheiten (tei:bibl), jene Zeitungszitate, zu denen es Digitalisate auf anno.onb.ac.at gibt, sind dorthin verlinkt, alle Kalenderdaten sind mit ISO-Daten vorbereitet.

Das born digital TEI-XML wird darüber hinaus mit Links zu existierenden und vom Projekt oder dessen Partnern betreuten eigenen Online-Datenbanken verlinkt. Named entities werden mit mrp.acdh.oeaw.ac.at verbunden, wobei Personennamen mit GND-IDs und geographische Bezeichnungen mit Georeferenzen über geonames.org referenziert sind, Institutioneneinträge verweisen großteils auf Digitalisate von Einträgen im Staatshandbuch der Österreichisch-Ungarischen Monarchie 1917 auf alex.onb.ac.at,

wo auch deren Mitarbeiter und verbundene Institutionen verzeichnet sind. Bibliographische Einheiten, inklusive Gesetzblätter, sind in einer offenen Zotero-group library erfasst.

Diese Anreicherungen sollen erweitert (Beispiel Parlamentsprotokolle auf ALEX) und sukzessive auch auf die retrodigitalisierten TEI-XML ausgeweitet werden.

Sämtliche Daten werden begleitend auf Zenodo zugänglich gemacht.

2.3) Workflow: pragmatische Lösungen

Die Digitalisierung hat den ohnehin schon komplexen Workflows der Edition eine neue Dimension der Komplexität verliehen und zwingt zur Reflexion derselben. Besonders das Erstellen, Betreuen und Einbinden der peripheren Datenbanken erfordert zusätzliche Arbeitskraft und Planung.

Die Originaldokumente werden fotografiert und (automatisch oder händisch) transkribiert und in MS Word kollationiert sowie wissenschaftlich kommentiert. Parallel dazu wird der Text mit Links und Formatvorlagen angereichert. Das angereicherte Worddokument wird dann über eine XSL-T-Pipeline in TEI-XML umgewandelt und über die parallel betriebene Datenbank der Auxiliardaten mit `tei:listEvent/Org/Person/Place/Bibl-Elementen` angereichert.

Die Motivation des Workflow ist, a) den Geboten einer historisch-kritischen Edition gerecht zu werden (Übertragung der Originaldokumente und kritischer Kommentar), b) den Arbeitsgewohnheiten der Editoren entgegenzukommen (Bearbeiten und Kollationieren in MS-Word), c) den Standards offener Online-Editionen zu entsprechen (XML-TEI, Auxiliardaten), d) große Textmengen gleichzeitig und auch retrospektiv für Print und online formatieren zu können (XSLT).

2.4) Hybridedition: ein exploratives Feld

Da aus dem TEI-XML nicht nur die Webversion generiert wird, sondern auch eine Druckvorlage, die der Editionstradition möglichst folgen soll, muss aus den angereicherten Elementen auch der Apparat (Liste der Teilnehmenden, Abkürzungen, Bibliografie, Register etc.) erstellt werden. Hier überschneiden sich teilweise Logiken, was zu Problemen führen kann. Die Arbeit an der Lösung solcher Probleme zwingt zur Ausformulierung bisher implizit gebliebener Editionslogiken. Da es nicht viele Hybrideditionen dieser Art und in diesem Umfang gibt, bietet dieses Ausformulieren Gelegenheit, Einblick sowohl in mediumsbedingt unterschiedliche Möglichkeiten der zwei Publikationsarten zu nehmen und Editionsprinzipien zu überdenken. Bspw. scheint ein Problem aber auch ein Vorteil des traditionellen Edierens zu sein, inkonsequent sein zu können.

Die Herkunft der Textgrundlagen war bisher relativ einheitlich und daher ergaben sich kaum Fragen dazu. Mit Edition der Brandakten ergeben sich neue Fragen der Textgrundlage, welche das digitale Editionsmodell immer wieder zur Anpassung zwingen.

3) Zusammenfassung/Ausblick

Eine umfangreiche Ressource zur politischen, aber auch sozialen, wirtschaftlichen und rechtlichen Geschichte nicht nur der Habsburger Monarchie, die in über 40 Bänden in Print verfügbar sein wird, ist weltweit zugänglich und auf ganz neue Art verwendbar für mehr Fachrichtungen und Usergruppen als bisher.

Bibliographie

Kurz, Stephan / Fischer-Nebmaier, Wladimir / Kampkasper, Dario / u. a. (2019): "Die Edition der Ministerratsprotokolle 1848–1918 digital: Workflows, Möglichkeiten, Grenzen" in: Zeppezauer-Wachauer, Katharina / Hinkelmanns, Peter / Ernst, Marlene (eds.), 5. *digital humanities austria konferenz DHA 2018 conference proceedings*, Salzburg 86–93.

O. A. (2019): "Durch Krisen zum Sozialstaat" in: *scilog*, [online] <https://scilog.fwf.ac.at/kultur-gesellschaft/8985/durch-krisen-zum-sozialstaat> [10.11.2020].

Rumpler, Helmut (1970): *Einleitungsband. Ministerrat und Ministerratsprotokolle 1848–1867. Behördengeschichtliche und aktenkundliche Analyse*, Wien: Verlag der Österreichischen Akademie der Wissenschaften (Die Protokolle des österreichischen Ministerrates 1848–1867).

Towards a Computational Study of German Book Reviews

A Comparison between Emotion Dictionaries and Transfer Learning in Sentiment Analysis

Rebora, Simone

simone.rebora@uni-bielefeld.de
University of Bielefeld, Germany

Messerli, Thomas

thomas.messerli@unibas.ch
University of Basel, Switzerland

Herrmann, J. Berenike

berenike.herrmann@uni-bielefeld.de
University of Bielefeld, Germany

This poster reports on the groundwork for the computational study of evaluative practices in German language book reviews. We trained classifiers for evaluation and sentiment at sentence level on the LOBO corpus, comprising ~1.3 million book reviews downloaded from the social reading platform LovelyBooks.

For the two classification tasks, we compared performance of dictionary-based and transfer-learning (TL) based sentiment analysis (SA). To use dictionary-based SA systematically, a repository of twelve open-source German-language SA lexicons was created (see Table 1). Lexicon formats were uniformed to automatically annotate reviews for sentiment in a processing pipeline. For the TL approaches, we chose BERT and FastText, both of which based on distributional representations of natural language (see Devlin et al., 2019; Mikolov et al., 2017).

Lexicon	Length (words)	Dimensions (categories)	Reference
ADU	26,832	12	(Hölzer et al., 1992)
AffectiveNorms	351,617	4	(Köper and Schulte im Walde, 2016)
BAWLr	2,902	3	(Vö et al., 2009)
Ekman	4,293	7	(Klinger et al., 2016)
Germanlex	8,693	1	(Clematide and Klenner, 2010)
LANG	1,000	3	(Kanske and Kotz, 2010)
MorphComp	9,256	3	(Ruppenhofer et al., 2017)
NRC	4,622	10	(Mohammad and Turney, 2013)
Plutchik	951	8	(Stamm, 2014)
PolarityCues	10,790	3	(Waltinger, 2010)
SentiArt	116,313	7	(Jacobs, 2019)
sentiWS	3,471	1	(Remus et al., 2010)

Tab. 1: Overview of German-language sentiment dictionaries

The dictionary-based and TL approaches were evaluated on two manually annotated datasets, working with two annotators: in the first dataset (~21,000 sentences), the annotation task was that of identifying evaluative language (vs. descriptive language); in the second dataset (~13,500 sentences), the task focused on the distinction between positive and negative sentiment. These two classification tasks form the basis for a large-scale analysis of the LOBO corpus, which segments reviews into evaluative and descriptive passages, to describe differences in evaluation practices across genres (e.g., romance, science fiction) and ratings (1-5 stars).

For the creation of the Gold Standard of Task 1 (evaluation classification), manual annotation reliability was evaluated on a subset of 250 reviews (~4,000 sentences). Cohen's *Kappa* (0.76) indicated a strong agreement between annotators. Overall, 66% of the total sentences were annotated as "evaluation". Training an SVM classifier on the features generated by the 12 sentiment dictionaries rendered a macro *F1* score of 0.75 (see Table 2 for details).

	Precision	Recall	<i>F1</i>	Support
Evaluation	0.854	0.777	0.813	2852.6
Other	0.636	0.746	0.687	1494.4
Accuracy			0.766	4347
Macro	0.745	0.761	0.750	4347
Weighted	0.779	0.766	0.770	4347

Tab. 2: Efficiency of dictionary-based SVM on Task 1

To compare the efficiency of the dictionaries, the same classifier was trained separately with the single dictionaries. Fig. 1 shows the results, with AffectiveNorms as the best-performing dictionary (macro *F1* score of 0.67).

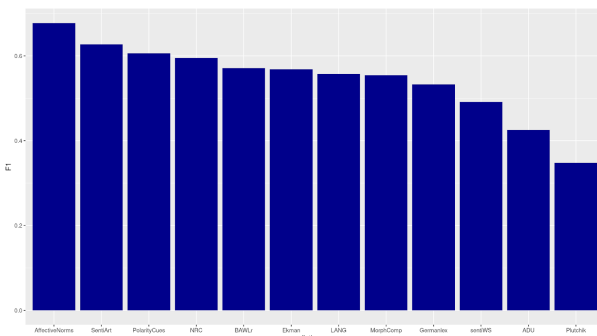


Fig. 1: Efficiency of single German-language SA dictionaries (Task 1)

Yet, by contrast, TL-methods proved substantially more efficient, with macro *F1* scores of 0.83 for FastText and of 0.89 for BERT (results obtained via a 5-fold cross validation, repeated five times to average variance, see Table 3 for details on BERT).

	Precision	Recall	<i>F1</i>	Support
Evaluation	0.9206	0.9273	0.9239	2828.52
Other	0.8595	0.8473	0.8533	1482.6
Accuracy			0.8998	4311.12
Macro	0.89	0.8873	0.8886	4311.12
Weighted	0.8996	0.8998	0.8996	4311.12

Tab. 3: Efficiency of BERT on Task 1

The evaluation procedure was repeated on Task 2 (positive vs. negative sentiment). Again, inter-annotator agreement was strong for manual annotation of the Gold Standard (Cohen's *Kappa* = 0.79). Annotation percentages are shown by Fig. 2 (where the "other" category indicates both mixed feelings and the absence of evaluation).

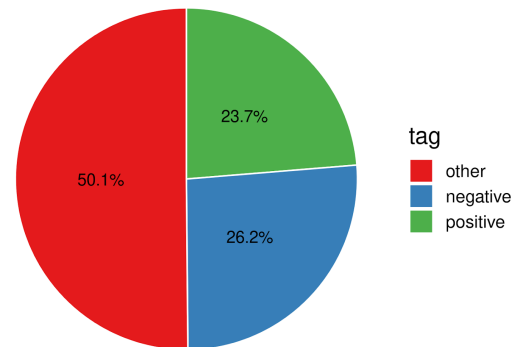


Fig. 2: Percentages of annotations for task 2

The dictionary-based SVM classifier reached a macro *F1* score of 0.64, while the best performance was obtained by SentiArt (see Fig. 3). Efficiency was again higher for FastText (macro *F1* score = 0.72) and best for BERT (macro *F1* score = 0.83). However, the learning curve for BERT shows how there is still room for improvement, with efficiency not fully reaching a plateau (see Fig. 4).

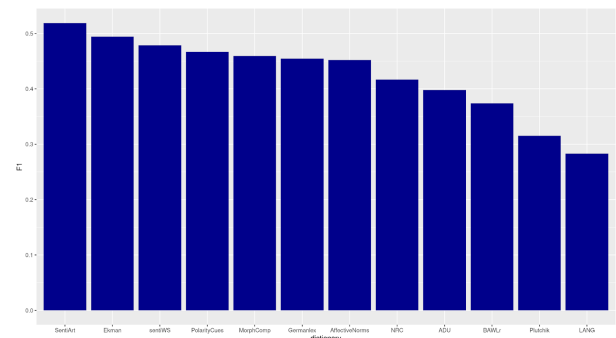


Fig. 3: Efficiency of single German-language SA dictionaries on Task 2

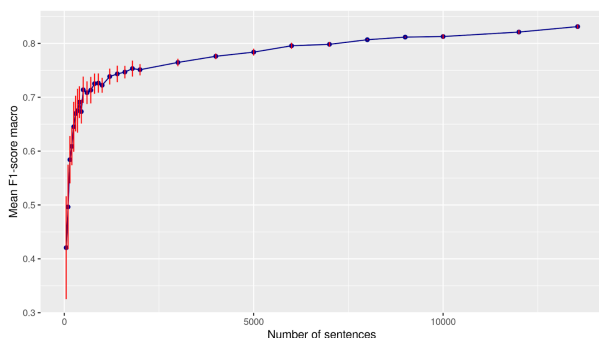


Fig. 4: Learning curve (with increasing amount of training material) of BERT for Task 2

Task	Lexicon-based sentiment analysis	TL-based sentiment analysis
Evaluative language (21,735 sentences)	SVMs trained on features generated by SA dictionaries: macro F1-score .75	BERT: macro F1-score .89 FastText: macro F1-score .83
Valence (13,552 sentences)	SVMs trained on features generated by SA dictionaries: macro F1-score .64	BERT: macro F1-score .85 FastText: macro F1-score .72

Tab. 4: Overview of results for lexicon-based and TL-based approach

Our results highlight the higher efficiency of TL-methods (see Table 4) and of dictionaries based on vector space models (like SentiArt and AffectiveNorms). They show that computational methods can reliably identify sentiment of book reviews in German. In order to fruitfully use similar methodology to identify types of engagement by reviewers with literature beyond the descriptive/evaluative and positive/negative dichotomies, a useful next step will be to attempt the design of TL-tasks for the identification of more fine-grained evaluative practices. These include the construction of and orientation to particular evaluative scales (e.g. reading pleasure, literary quality) and particular subjects of evaluation (e.g. novels, authors, characters).

Bibliography

Clematide, S. and Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. s.n. doi:10.5167/UZH-45506. <https://www.zora.uzh.ch/id/eprint/45506> (accessed 12 July 2021).

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]* <http://arxiv.org/abs/1810.04805> (accessed 13 July 2021).

Hölzer, M., Scheytt, N. and Kächele, H. (1992). Das „Affektive Diktionär Ulm“ als eine Methode der quantitativen Vokabularbestimmung. In Züll, C. and Mohler, P. Ph. (eds), *Textanalyse: Anwendungen der computerunterstützten Inhaltsanalyse. Beiträge zur 1. TEXTPACK-Anwenderkonferenz*. (ZUMA-Publikationen). Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 131–54 doi:10.1007/978-3-322-94229-6_7. https://doi.org/10.1007/978-3-322-94229-6_7 (accessed 12 July 2021).

Jacobs, A. M. (2019). Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI*, 6 doi:10.3389/frobt.2019.00053. <https://www.frontiersin.org/article/10.3389/frobt.2019.00053/full> (accessed 8 September 2019).

Kanske, P. and Kotz, S. A. (2010). Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, 42 (4): 987–91 doi:10.3758/BRM.42.4.987.

Klinger, R., Suliya, S. S. and Reiter, N. (2016). Automatic Emotion Detection for Quantitative Literary Studies: A case study based on Franz Kafka's 'Das Schloss' und 'Amerika'. *DH2016 Book of Abstracts*. Kraków: ADHO <https://dh2016.adho.org/abstracts/318>.

Köper, M. and Schulte im Walde, S. (2016). Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2595–98 <https://aclanthology.org/L16-1413> (accessed 12 July 2021).

Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C. and Joulin, A. (2017). Advances in Pre-Training Distributed Word Representations. *ArXiv:1712.09405 [Cs]* <http://arxiv.org/abs/1712.09405> (accessed 13 July 2021).

Mohammad, S. M. and Turney, P. D. (2013). CROWDSOURCING A WORD-EMOTION ASSOCIATION LEXICON. *Computational Intelligence*, 29 (3): 436–65 doi:10.1111/j.1467-8640.2012.00460.x.

Remus, R., Quasthoff, U. and Heyer, G. (2010). SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA) http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf (accessed 12 July 2021).

Ruppenhofer, J., Steiner, P. and Wiegand, M. (2017). Evaluating the Morphological Compositionality of Polarity. *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*. Incoma Ltd. Shoumen, Bulgaria, pp. 625–33 doi:10.26615/978-954-452-049-6_081. <http://www.acl-bg.org/proceedings/2017/RANLP%202017/pdf/RANLP081.pdf> (accessed 12 July 2021).

Stamm, N. (2014). Klassifikation und Analyse von Emotionswörtern in Tweets für die Sentimentanalyse.

Vö, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J. and Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, 41 (2): 534–38 doi:10.3758/BRM.41.2.534.

Waltinger, U. (2010). GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA) http://www.lrec-conf.org/proceedings/lrec2010/pdf/91_Paper.pdf (accessed 12 July 2021).

Training the Archive Von der maschinellen Exploration musealer Sammlungsdaten zur Curator's Machine

Bönisch, Dominik

dominik.boenisch@mail.aachen.de

Ludwig Forum für Internationale Kunst, Germany

Die Digitalisierung in Kunstmuseen verspricht einen erweiterten Zugriff auf Sammlungsobjekte sowohl für die Wissenschaft als auch für eine interessierte Öffentlichkeit und das bestenfalls online ortsunabhängig und jederzeit (Glinka / Dörk 2018: 236). Dabei reicht es nicht aus, die hiesige Archivlogik eins zu eins in den digitalen Raum zu übertragen und die Suche in Datenbanken auf eng gedachte Stichworte zu limitieren. Vielmehr sollen über spezielle Interfaces und Visualisierungen eine Exploration von digitalen Beständen, sowie ein *Schlendern* durch die Online-Sammlung möglich gemacht werden, ohne dabei zwingend einem vorgegebenen Suchbegriff folgen zu müssen (Brüggemann et al. 2016: 227). Beispielhafte Lösungen wie PixPlot¹, iArt² oder imgs.ai³ ermöglichen bereits heute mittels künstlicher Intelligenz die systematische und strukturierte Aufbereitung von Sammlungsdaten. Denn durch maschinelles Lernen können Zusammenhänge und Verbindungen zwischen Kunstwerken offenbart werden, die Kurator*innen nicht (mehr) mit bloßem Auge wahrnehmen können (Bell / Ommer 2016: 68). Ein maschinengestütztes, exploratives Aufdecken von Verknüpfungen innerhalb der eigenen musealen Sammlung ist Untersuchungsgegenstand des Forschungsprojekts Training the Archive am Ludwig Forum für Internationale Kunst Aachen im Verbund mit dem HMKV Hardware MedienKunst Verein, Dortmund und dem Visual Computing Institute der RWTH Aachen University.

Training the Archive (2020-2023) möchte ein Software-Tool entwickeln, welches hilft, die kuratorische Praxis und Recherche am Museum zu unterstützen. Dabei sollen Kurator*innen befähigt werden, beispielsweise über sinnvolle Nearest-Neighbor-Vorschläge die eigenen musealen Sammlungsdaten neu zu entdecken. Die zugrundeliegende Mustererkennung soll hierbei nicht nur auf visuelle Embeddings zugreifen, sondern auch semantische Verbindungen einbeziehen. Training the Archive entwickelt dazu ein kollaboratives Konzept, welches in die sogenannte Curator's Machine mündet, die im Poster vorgestellt werden soll. Durch einen effektiven Prozess der Mensch-Maschine-Interaktion soll auch das historische, stilistische und objektbasierte Kontextwissen von Expert*innen Beachtung finden, indem neuronalen Netzen ein *kuratorischer Blick* trainiert wird.

Da vortrainierte Netze nicht uneingeschränkt off-the-shelf für kunsthistorische Korpora verwendet werden können (vgl. Hunger 2021a), untersucht Training the Archive über Prototypen⁴, wie Embeddings aus visuellen Bild-Features und Text-Informationen (Metadaten, sprachliche Konzepte, semantische Kontexte) für Kunstsammlungen verbunden sowie nutzbar gemacht werden können. Dafür werden automatisierte Cluster um verdeckte Beziehungsmuster zwischen Kunstwerken oder die persönliche Intuition sowie den subjektiven Geschmack von Kurator*innen erweitert (vgl. Bönisch 2021) sowie ein Modell – ähnlich dem CLIP-Algorithmus (Radford et al. 2021), welcher semantische Suchen über Bild-Text-Zusammenhänge ermöglicht (Saglan 2021) – mit den Annotationen zu Werkbeschreibungen aus dem ARTigo-Projekt⁵ trainiert. Dies soll besonders variable Suchen für Kurator*innen ermöglichen und gleichzeitig deren Expert*innenwissen in den maschinellen Lernprozess einbeziehen. Eine im Forschungsprojekt durchgeführte Studie über das Kuratieren mit international tätigen Kurator*innen gibt dazu Aufschluss über den Prozess der Werkauswahl oder die Kontextualisierung von Kunstwerken in einer Ausstellung und soll helfen, die nächste prototypische Entwicklung zu schärfen. Dabei wird auch verhandelt, wie der Begriff des *Kuratorischen* im heutigen algorithmischen Zeitalter einzuordnen ist (Hunger 2021b). Zusammenfassend möchte das Poster das Forschungsprojekt Training the Archive, dessen Zielstellungen und Prototypen sowie erste Lessons Learned zur Mitte der Projektlaufzeit vorstellen.

Fußnoten

1. Bereits 2017 veröffentlichte Douglas Duhaime über das Yale Digital Humanities Lab eine Visualisierung für das Clustering von Bilddaten im hochdimensionalen Feature-Raum. Verfügbar unter: <https://dhlabs.yale.edu/projects/pixplot/>.
2. Ein digitales Research-Werkzeug, mit dessen Hilfe große Bilddatenmengen in den Geisteswissenschaften nutzbar gemacht werden sollen. Ein DFG-Projekt der Ludwig-Maximilians-Universität München, der Technischen Informationsbibliothek Hannover sowie des Heinz-Nixdorf-Instituts. Erreichbar unter: <https://www.iart.vision/>.
3. Eine visuelle Suchmaschine für die digitale Kunstgeschichte, die auf Embeddings neuronaler Netzwerke basiert. Entwickelt von Fabian Offert mit Unterstützung von Peter Bell und Oleg Harlamov. Abrufbar unter: <https://imgs.ai/>.
4. Ein Proof of Concept ist veröffentlicht unter: <https://github.com/DominikBoenisch/Training-the-Archive>. Ein weiterer Prototyp kann über die RWTH Aachen University genutzt werden: <https://vci.rwth-aachen.de/annotation-tool/>.
5. ARTigo ist ein Forschungsprojekt der Ludwig-Maximilians-Universität München, welches mittels Crowdsourcing über ein Online-Spiel das Ziel verfolgt, Kunstwerke mit Tags zu versehen und durch Schlagworte zu beschreiben. Siehe: <http://www.artigo.org>.

Bibliographie

Bell, Peter / Ommer, Björn (2016): "Visuelle Erschließung. Computer Vision als Arbeits- und Vermittlungstool", in: *Konferenzband EVA Berlin 2016*. Elektronische Medien & Kunst, Kultur und Historie. EVA Berlin, Band 23. Heidelberg: arthistoricum.net 67-73.

Bönisch, Dominik (2021): "The Curator's Machine: Clustering of Museum Collection Data through Annotation of Hidden Connection Patterns Between Artworks", in: *International Journal for Digital Art History*, 5 (Mai): 5.20-5.35. doi:10.11588/dah.2020.5.75953.

Brüggemann, Viktoria / Kreiseler, Sarah / Dörk, Marian (2016): "Museale Bestände im Web: Eine Untersuchung von acht digitalen Sammlungen", in: *Konferenzband EVA Berlin 2016*. Elektronische Medien & Kunst, Kultur und Historie. EVA Berlin, Band 23. Heidelberg: arthistoricum.net 227-236.

Glinka, Katrin / Dörk, Marian (2018): "Zwischen Repräsentation und Rezeption – Visualisierung als Facette von Analyse und Argumentation in der Kunstgeschichte", in: *Computing Art Reader: Einführung in die digitale Kunstgeschichte*. Computing in Art and Architecture, Band 1. Heidelberg: arthistoricum.net 234-250.

Hunger, Francis (2021a): "Why so many windows?" – Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst. Training the Archive – Working Paper, Aachen/Dortmund, Juni. doi:10.5281/zenodo.4742621.

Hunger, Francis (2021b): *Kuratieren und dessen statistische Automatisierung mittels Künstlicher 'Intelligenz'*. Training the Archive – Working Paper, Aachen/Dortmund, Oktober. doi:10.5281/zenodo.5589930.

Radford, Alec / Kim, Jong Wook / Hallacy, Chris / Ramesh, Aditya / Goh, Gabriel / Agarwal, Sandhini / Sastry, Girish / Askell, Amanda / Mishkin, Pamela / Clark, Jack / Krueger, Gretchen / Sutskever, Ilya (2021): "Learning Transferable Vi-

sual Models from Natural Language Supervision ", in: *arXiv pre-print*. arxiv:2103.00020.

Saglani, Vatsal (2021): "What is CLIP (Contrastive Language – Image Pre-training) and how can it be used for semantic image search?", *Towards AI* , 9. Februar <https://pub.towardsai.net/what-is-clip-contrastive-language-image-pre-training-and-how-it-can-be-used-for-semantic-image-b02ccf49414e> [letzter Zugriff 22.11.2021].

Volltexterkennung für historische Sammlungen mit OCR4all-libraries iterativ und partizipativ gestalten

Klaes, Jan Sebastian

klaes@leibniz-gei.de

Georg Eckert Institute for International Textbook Research.
Member of the Leibniz Association

Korwisi, Kristof

kristof.korwisi@uni-wuerzburg.de

University of Würzburg Human-Computer Interaction, Germany

Krüger, Katharina

katharina.krueger@gei.de

Georg Eckert Institute for International Textbook Research.
Member of the Leibniz Association

Reul, Christian

christian.reul@uni-wuerzburg.de

University of Würzburg Centre for Philology and Digitality,
Germany

Towara, Nadine

towara@leibniz-gei.de

Georg Eckert Institute for International Textbook Research.
Member of the Leibniz Association

Mit der Initiierung und Durchführung von Massendigitalisierungsprojekten haben Bibliotheken eine wesentliche Grundlage für den Zugang und die Nutzung digitaler Quellen geschaffen. Vor dem Hintergrund der Weiterentwicklung digitaler Forschungsmethoden u.a. im Bereich der Digital Humanities stellt sich nun zunehmend die Frage nach der Qualität der Volltexterkennung (OCR). Die Volltextqualität in den digitalen Sammlungen ist dabei nicht nur abhängig von Materialbesonderheiten, sondern auch von dem eingesetzten Texterkennungssystemen und deren Weiterentwicklungen.

Die Zusammenarbeit zwischen dem Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung (GEI), dem Zentrum für Philologie und Digitalität "Kallimachos" (ZPD) und dem Lehrstuhl für Mensch-Computer-Interaktion (HCI) der Universität Würzburg zielt darauf ab, das Web-GUI-basierte Open-Source-Werkzeug OCR4all (Reul et. al 2017; 2019) so zu

erweitern und anzupassen, dass Bibliotheken und Archive bei ihrer Massendigitalisierung die im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten OCR-D-Verbundprojekts (Engl, 2020) erarbeiteten Lösungen niederschwellig, flexibel und eigenständig einsetzen können. Als Use Case fungiert die Forschungsbibliothek des GEI mit ihrer digitalisierten Schulbuchbibliothek GEI-Digital.

Die digitale Schulbuchbibliothek umfasst historische deutsche Schulbücher der Fächer Geschichte, Geographie und Politik, Religion/Werteerziehung sowie Realien- und (Erst-)Lesebücher von den Anfängen der Schulbuchproduktion im 17. Jahrhundert bis zum Ende des Ersten Weltkriegs (Hertling/Klaes 2018a ; 2018b) . Mit GEI-Digital ist für die Digital Humanities ein einzigartiges Korpus mit über 6.100 digitalisierten Schulbüchern entstanden, dass die gesamte Epoche der deutschen Schulbücher von deren Entstehung bis 1918 mit hoher Vollständigkeit virtuell zusammenführt. Die Digitalisate und Daten werden in zahlreichen Digital-Humanities-Projekten bereits nachgenutzt, wie z.B. im Projekt „Welt der Kinder“, in dem das Korpus mit Topic Modeling-Verfahren untersucht wurde (Nieländer / Weiß 2018).

Die besonderen Bedarfe nach hochwertiger Texterkennung unter der Gruppe der Forscher*innen und der digitalen Schulbuchbibliothek wurden 2014 in Form einer Befragung ermittelt. Darauf aufbauend erfolgte eine Studie zur Qualität der Texterkennung der seit 2009 eingesetzten kommerziellen und Open-Source-Softwarelösungen für Texterkennungsverfahren bei der Massendigitalisierung. Die Ergebnisse der Studie zeigten auf, dass Massenverfahren nicht die Bedarfe der Forscher*innen decken. Der digitale Bestand weist erhebliche Unterschiede in der OCR-Qualität auf, auch weil ein komplexes Layout und uneinheitliche Typographien noch immer große Hürden für eine hochwertige Volltexterkennung darstellen. Um die OCR-Qualität gezielt zu verbessern, soll ausgehend vom konkreten Use Case der Forschungsbibliothek des GEI ein möglichst generisch anwendbares Verfahren implementiert werden, das eine nach Sammlungen mit jeweils ähnlicher Materialgrundlage organisierte Volltexterkennung erlaubt.

Um zunehmende Komplexitäten der so entstehenden OCR-Lösung nutzerorientiert aufzufangen, wird die bestehende grafische Benutzerschnittstelle in enger Kooperation und unter Anleitung der HCI angepasst und weiterentwickelt. Eine zusätzliche visuelle Erklärungskomponente soll darüber hinaus Unterstützung bei der Erstellung und Konfiguration optimaler OCR-Workflows bieten. Alle im Projekt erarbeiteten Lösungen werden schritt haltend mittels umfassender Nutzerstudien evaluiert. Um sicher zu stellen, dass auch nicht-technische Anwender*innen in Bibliotheken und Archiven komfortabel und selbstständig auf OCR-D-Lösungen zugreifen können, fließen die Evaluationsergebnisse stetig in die Weiterentwicklung ein. Dafür werden im Rahmen des Projekts Workshops mit interessierten Anwendern stattfinden, um deren Bedarfe in die Weiterentwicklung der Benutzerführung einfließen zu lassen. Ein besonderer Fokus wird dabei auf die intuitive und selbsterklärende Bedienbarkeit durch ein breites Nutzerspektrum gelegt.

Forscher*innen als wichtige Zielgruppe von Forschungs- und Spezialbibliotheken profitieren mit der Weiterentwicklung von OCR4all von der Möglichkeit OCR-Prozesse *iterativ* und *partizipativ* mitgestalten zu können. Die im Rahmen des OCR-D-Verbundprojekts entwickelten Komponenten können von Forscher*innen dynamisch eingesetzt und bei Bedarf sind Konfigurationen, Trainingsdaten und Modelle zwischen Institutionen und Individuen flexibel austauschbar, was die Optimierung von Texterkennungsprozessen verbessert. OCR-Prozesse und deren Komponenten sollen damit nachnutzbar werden. Ebenso sollen für die Forscher*innen Möglichkeiten geschaffen werden, die Qualität

der Texterkennung bewerten zu können und selbst OCR-Prozesse für bestimmte Bestände steuern zu können, da deren spezifischen Bedarfe oftmals an bestimmte Forschungsfragen gebunden sind.

Bibliographie

Engl, Elisabeth. "OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative". *Bibliothek Forschung und Praxis*, vol. 44, no. 2, 2020, pp. 218-230. <https://doi.org/10.1515/bfp-2020-0024>

Hertling, Anke; Klaes, Sebastian (2018): "Historische Schulbücher als digitales Korpus für die Forschung: Auswahl und Aufbau einer digitalen Schulbuchbibliothek". In: Maret Nieländer und Ernesto William De Luca (Hg.): *Digital Humanities in der internationalen Schulbuchforschung*. Göttingen: V&R unipress, S. 21–44. DOI: 10.14220/9783737009539.21

Hertling, Anke; Klaes, Sebastian (2018): "'GEI-Digital' als Grundlage für Digital-Humanities-Projekte: Erschließung und Datenaufbereitung". In: Maret Nieländer und Ernesto William De Luca (Hg.): *Digital Humanities in der internationalen Schulbuchforschung*. Göttingen: V&R unipress, 45-68. DOI: 10.14220/9783737009539.45

Nieländer, M., Weiß, A. (2018): "Schönere Daten - Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten". In: Maret Nieländer und Ernesto William De Luca (Hg.): *Digital Humanities in der internationalen Schulbuchforschung*. Göttingen: V&R unipress, 91-116. DOI: 10.14220/9783737009539.91

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F. (2019). "OCR4all - An Open-Source Tool Providing a (Semi-) Automatic OCR Workflow for Historical Printings". In: *Applied Sciences* 9 (22) 4853. <https://doi.org/10.3390/app9224853>

Reul, C., Springmann, U., Puppe, F. (2017). "LAREX: A Semi-automatic Open-Source Tool for Layout Analysis and Region Extraction on Early Printed Books". In: *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage, DATeCH2017*, 137–142, New York, NY, USA. ACM. <https://doi.org/10.1145/3078081.307809>

Webservice correspSearch subVersion 2

Dumont, Stefan

dumont@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Grabsch, Sascha

grabsch@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Müller-Laackman, Jonas

jonas.mueller-laackman@fu-berlin.de
Freie Universität Berlin

Briefe sind wichtige Quellen für eine Vielzahl geisteswissenschaftlicher Fragestellungen.¹ Sie versprechen und ermöglichen

Einblicke in die Lebens- und Gedankenwelt der Korrespondent:innen, in ihnen wird eine Vielzahl von Themen, Ereignissen, Personen usw. angesprochen oder kommentiert (Schmid 2001, 38). Darüber hinaus bilden historische Korrespondenzen ein – eigentlich unbegrenztes – Briefnetzwerk (Bunzel 2013). Schon lange werden Briefe deshalb in wissenschaftlichen Editionen erschlossen. Die Anzahl edierter Briefe wächst mittlerweile allerdings so stark, dass ihre Menge kaum noch zu überschauen ist (Bunzel 2013). Darüber hinaus sind Briefeditionen in der Regel an einer oder zwei Korrespondent:innen orientiert. Die Erschließungsform einer Briefedition wird also (gerade in gedruckter Form) einem zentralen Charakteristikum der Textsorte „Brief“ nur schlecht gerecht: Briefe Dritter, denen keine eigene Ausgabe gewidmet ist, lassen sich nur mit Mühe auffinden, erweiterte Korrespondenznetzwerke, Arbeits- oder Freundeskreise werden durch eine solche Editionspraxis aufgespalten oder gar unsichtbar. Für systematische, nicht personenzentrierte Fragestellungen werden so Hürden aufgebaut.

Um diese methodischen Probleme im Umgang mit Briefen zu lösen, wurde bereits 2014 der Webservice *correspSearch* (<https://correspSearch.net>) entwickelt, der Briefmetadaten aus gedruckten und digitalen Editionen aggregiert und zur Recherche bereitstellt (Dumont 2016). Stand *correspSearch* bisher nur als Prototyp mit eingeschränkten Suchfunktionalitäten zur Verfügung, wurde im Juni 2021 die neu entwickelte Version 2.0 mit vielen neuen Recherchemöglichkeiten relaunched.² Das Poster wird die maßgeblichen Neuerungen dieser Version sowie einige zugrundeliegende theoretische und technische Entscheidungen vorstellen und beispielhaft Nutzungsszenarien aufzeigen.

Mit dem Relaunch wird bei der Suche eine umfangreiche Auswahl an Facetten angeboten, die eine explorative Erkundung und die Filterung der Suchergebnisse erlauben. So gibt ein Histogramm einen schnellen Überblick über die nachgewiesene Korrespondenz pro Jahr; Korrespondent:innen und Orte lassen sich anhand der Trefferanzahl oder alphabetisch sortieren, usw. Die Normdaten-basierte Architektur von *correspSearch* ermöglicht die Anreicherung der aus den Briefverzeichnissen übernommenen Daten mit Informationen aus Normdateien. Ein erstes Ergebnis dieser Anreicherung ist die Suche anhand des Geschlechts der Korrespondent:innen, die natürlich mit weiteren Merkmalen kombiniert werden kann. In Zukunft werden auch Suchen nach Briefen von bestimmten Berufsgruppen (Musiker:innen, Philolog:innen etc.) ermöglicht. Darüber sollen auch die Inhalte von Briefen schlagwortartig recherchierbar gemacht werden (Dumont u. a. 2019).

Eine völlig neue Rechercheoption der jetzt veröffentlichten Version 2 des Webservices bietet die Karten-basierte Suche: hier kann auf einer Karte eine Region frei eingezeichnet werden, in der alle Orte als Schreib- oder Empfangsort enthalten sind – natürlich in Kombination mit Datumsangaben. Dadurch werden ereignisbasierte Suchen möglich.³ Wichtige Grundlage für eine solche Suche ist die oben bereits erwähnte Nutzung von Normdaten. Die in den indizierten Briefverzeichnissen angegebenen Ortsangaben werden bei der Verarbeitung und Aufbereitung für die Indizierung in Elasticsearch mit Koordinaten aus dem Datenbestand von *geonames.org* angereichert. Darüber hinaus können in der Karten-basierten Suche neben frei gezeichneten Regionen auch vordefinierte historische Staatsgebiete verwendet werden. Deren geographische Angaben werden von HistoGIS (Andorfer u. a. 2019), einem Webservice der Österreichischen Akademie der Wissenschaften, als Shapefiles bezogen. *CorrespSearch* folgt hier wie auch bei Personen und Orten dem Ansatz einer „lightweight infrastructure“, bei dem auf briefspezifische Charakteristika fokussiert wird und für andere Aspekte (Lebensdaten zu Personen, Geoko-

dinaten, etc.) die Daten anderer einschlägiger Informationsinfrastrukturen und Normdatengeber nachgenutzt werden.

CorrespSearch weist derzeit rund 140.000 Briefe nach – von der Reformation bis weit ins 20. Jahrhundert. Die Daten werden dabei von Editionsprojekten und Institutionen aus den verschiedenen Fachcommunities in einem standardisierten Austauschformat – dem Correspondence Metadata Interchange Format (TEI Correspondence SIG 2015; Dumont u. a. 2019) – bereitgestellt. Um Datenbereitstellungen so einfach wie möglich zu machen, erlaubt die browserbasierte Eingabeoberfläche des CMIF Creators⁴ die einfache Erfassung von Briefmetadaten inkl. Normdaten-IDs.

Als Webservice bietet correspSearch die aggregierten, frei-lizenzierten Daten nicht nur in einer Rechercheoberfläche an, sondern auch zum automatisierten Abruf über eine API. Dadurch können die Briefmetadaten umfassend nachgenutzt werden, etwa zur Analyse mit Methoden der Historischen Netzwerkforschung oder zur Integration in digitale Editionen – vgl. zum Beispiel die Alexander von Humboldt-Chronologie in der edition humboldt digital (Schwarz 2017). Zur automatischen Vernetzung von Briefeditionen bietet der Webservice mit Version 2 das Javascript-Widget csLink an.⁵ csLink ermöglicht digitalen Briefedition den automatisierten Verweis von einem edierten Brief auf Briefe und Korrespondenzpartner:innen in anderen Editionen.

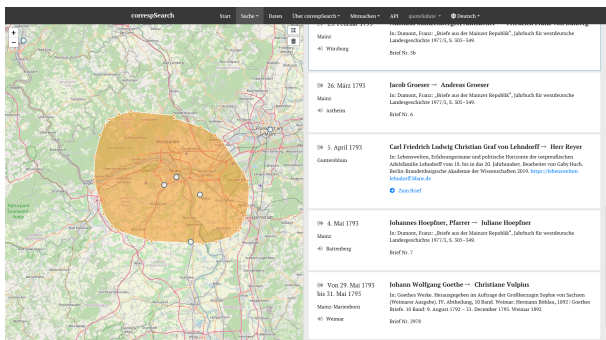


Abb. 1

Fußnoten

1. Vgl. die vielfältigen Beiträge zum *Brief als Forschungsfeld* im Handbuch Brief (Matthews-Schlinzig u. a. 2020).
2. <https://correspsearch.net/de/suche.html>
3. Eine beispielhafte Suche für diese Suchfunktion: die Region um Mainz (als Schreibort) zur Zeit der Mainzer Republik, d.h. von Oktober 1792 bis Juli 1793. Das Ergebnis bringt zahlreiche Treffer in verschiedenen Editionen hervor, darunter auch ein Brief, der wohl sonst nicht gefunden worden wäre: in dem Brief aus dem rheinhessischen Dorf Guntersblum schildert der preußische Offizier Carl Friedrich Ludwig Graf von Lehndorff die Tage unmittelbar vor der Einschließung der Festung Mainz Guntersblum im April 1793. Vgl. Abb. 1
4. <https://correspsearch.net/de/cmif-creator.html>
5. <https://github.com/correspSearch/csLink>

Bibliographie

Andorfer, Peter, Matthias Schlögel, Antonia Dückelmann, Peter Paul Marckhgott-Sanabria, und Anna Piechl (2019):

HistoGIS. A Geographical Information System, Workbench and Repository to Retrieve, Collect, Create, Enrich and Preserve Historical Temporalized Spatial Data Sets [Webservice]. Wien: Austrian Centre for Digital Humanities and Cultural Heritage / Österreichische Akademie der Wissenschaften. <https://histogis.acdh.oeaw.ac.at/>.

Bunzel, Wolfgang (2013): „Briefnetzwerke der Romantik. Theorie - Praxis - Edition.“ In: *Brief-Edition im digitalen Zeitalter*, herausgegeben von Anne Bohnenkamp und Elke Richter, 109–32. Beihefte zu editio 34. Berlin/Boston: De Gruyter.

Dumont, Stefan (2016): „CorrespSearch – Connecting Scholarly Editions of Letters“. *Journal of the Text Encoding Initiative* 10 (Dezember). <https://doi.org/10.4000/jtei.1742>.

Dumont, Stefan, Ingo Börner, Dominik Leipold, Jonas Müller-Laackman, und Gerlinde Schneider (2019): „Correspondence Metadata Interchange Format“. In: *Encoding Correspondence. A Manual for TEI-XML-Based Encoding of Letters and Postcards in TEI-XML and DTABf*, herausgegeben von Stefan Dumont, Susanne Haaf, und Sabine Seifert, 1. Aufl. Berlin: Berlin-Brandenburg Academy of Sciences and Humanities. <https://encoding-correspondence.bbaw.de/v1/CMIF.html>.

Handbuch Brief. Von der Frühen Neuzeit bis zur Gegenwart., herausgegeben von Marie Isabel Matthews-Schlinzig, Jörg Schuster, Gesa Steinbrink, und Jochen Strobel (2020). Berlin/Boston: De Gruyter.

Schmid, Irmtraud (2001): „Anforderungen an die Kommentierung von Briefen und amtlichen Schriftstücken“. In: *„Ich an Dich“. Edition, Rezeption und Kommentierung von Briefen*, 35–45. Innsbrucker Beiträge zur Kulturwissenschaft. Germanistische Reihe 62. Innsbruck: Universität Innsbruck.

TEI Correspondence SIG, Hrsg. (2015): „Correspondence Metadata Interchange Format (CMIF)“. <https://github.com/TEI-Correspondence-SIG/CMIF>.

What was Theoretical Biology?

A Topic-Modelling Analysis of a Multilingual Corpus of Monographs and Journals, 1914-1945

Böhm, Alexander

alexander.boehm@rub.de
Department of Philosophy I, Ruhr University Bochum

Reiners-Selbach, Stefan

stefan.reiners-selbach@hhu.de
Faculty of Arts and Humanities, Heinrich-Heine-University Düsseldorf

Baedke, Jan

jan.baedke@rub.de
Department of Philosophy I, Ruhr University Bochum

Fábregas Tejeda, Alejandro

Alejandro.FabregasTejeda@ruhr-uni-bochum.de
Department of Philosophy I, Ruhr University Bochum

Nicholson, Daniel J.

dnicho@gmu.edu
Department of Philosophy, George Mason University

Over the course of the twentieth century, theoretical biology changed beyond all recognition. Although the field today is synonymous with mathematical biology, when it first emerged it had a drastically different agenda: to critically analyze the conceptual foundations of biology in order to resolve long-standing theoretical disputes, abstract from the ‘burden of details,’ and bring about the epistemic unification of biological science. The field began acquiring its now familiar mathematical character in the 1940s, as formal models became increasingly applied in different areas of biology, such as growth, ecology, genetics, and evolution. Regrettably, the early ‘philosophical’ period of theoretical biology has been almost completely forgotten and its existence is seldom acknowledged—let alone carefully examined (but see Nicholson & Gawne 2015, Baedke 2019). Much of this early discourse took place in a handful of book series, monographs, and journals, the majority of which were published in German (at least initially). Hence, it is perhaps not surprising that Anglophone scholars remain almost completely ignorant of this large, and surprisingly rich, body of literature.

Our aim is to analyze this forgotten corpus and rescue it from the dustbin of history. Our guiding question is: What did theoretical biology look like in the early 20th century? More specifically, we ask: (i) What were the central debates and topics? (ii) Who were the central authors and how international was the scientific community at the beginning? (iii) Can distinct language-(of-origin)-specific camps be identified in terms of the kinds of topics they addressed? (iv) What, where, and when did transitions occur in networks of authors and topics? (v) When, how, and why did the discipline develop its emphasis on formal modeling? At this early exploratory stage of the project, we operationalize these central questions mainly as a topic-modelling problem: (1) Which central topics can be identified and how does their ‘share’ of the documents develop? Which topic clusters can be identified? (2-3) Are certain topics dominated by particular authors, languages (of origin), and nationalities? (4) Can certain ‘turning points’ be identified? Additionally: (5) How steadily does the proportion of publications that use mathematical formulas increase over time? Is it gradual or rather discontinuous?

After (a) preparing and selecting documents for the corpus on a historical basis (encompassing monographs, book series and journal articles)—digitizing, and OCR-ing with tesseract where necessary—we (b) machine translate the non-German texts into German using the Google Translate API. As de Vries, Schoonvelde, and Schumacher (2018) argue for topic-modelling in general, and Malaterre (2021) for the special use-case of history of science, modern machine translations deliver useful results that are reliable for multilingual topic-modelling. Additionally, we plan to assess our translation accuracy with Malaterre’s proposed “Semantic Topology Preservation Test” (2021). Then, we (c) preprocess the corpus: Following a general cleaning of common OCR-errors and stop words, we reduce the corpus to lemmatized adjectives and nouns via spaCy’s POS tagging and lemmatization algorithms. We assume that the conceptual topics we aim to explore are mostly expressed in nouns and adjectives (see Jockers 2013, Malaterre et

al. 2020). The preprocessed documents are then (d) analyzed with LDA topic-modelling, using gensim’s MALLET-wrapper and (e) analyzed with top2vec, to cluster the documents thematically – enabling a different granularity and perspective, since top2vec does not treat the documents as bags-of-words and tends to generate few more general topics (see Angelov 2020). Finally, (f) we calculate document embeddings using UMAP and (g) visualize the embedding as an interactive scatter plot (with the option of time-period slices) with Bokeh, since the heterogeneity of our corpus does not allow for a simple linear visualization. We enrich the scatter plot with metadata for a mouse-over pop-up window, generated from the most important topics for each document, and color the documents by their top2vec cluster, complementing the visual clustering and topological distribution the document embedding shows. Thus, we create an interactive tool for exploration, hoping to motivate future research.

Moreover, we plan to utilize tesseract’s equ language data to detect mathematical equations in documents. We take the use of mathematical formulas as a signal of affiliation with the mathematical side of the discourse on theoretical biology. This way, each document is assigned a gradual mathematization score. To model the mathematization of theoretical biology, we then analyze the mathematization scores per year and the scores’ correlations with topics. The score can in turn be used for visual classification in the visualization by choosing different symbols for documents in the scatter plot based on their score.

Bibliography

- Alt, W.; Deutsch, A.; Kamphuis, A.; Lenz, J. and Pfister, B. (1996). "Zur Entwicklung der Theoretischen Biologie: Aspekte der Modellbildung und Mathematisierung", in: *Jahrbuch für Geschichte und Theorie der Biologie* 3, pp. 7-59.
- Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics*, in: arXiv:2008.09470v1. <https://arxiv.org/abs/2008.09470v1>
- Baedke, J. (2019). "O Organism, Where Art Thou? Old and New Challenges for Organism-Centered Biology", in: *J Hist Biol* 52, pp. 293–324. <https://doi.org/10.1007/s10739-018-9549-4>
- Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003) "Latent Dirichlet allocation", in: *J Mach Learn Res* 3 (March), pp. 993–1022.
- Bokeh Development Team (2018). *Bokeh: Python library for interactive visualization*. URL: <http://www.bokeh.pydata.org>
- De Vries, E.; Schoonvelde, M. and Schumacher, G. (2018). "No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications", in: *Political Analysis*, 26 (4), pp. 417 – 430. <https://doi.org/10.1017/pan.2018.26>
- Honnibal, M.; Montani, I.; Van Landeghem, S. and Boyd, A. (2020). *spaCy 3.1: Industrial-strength Natural Language Processing in Python*. <https://spacy.io/>
- Jockers, M. (2013). "Secret" recipe for topic modeling themes. <https://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/>
- Laubichler, M. (2001). "Mit oder ohne Darwin? Die Bedeutung der darwinschen Selektionstheorie in der Konzeption der Theoretischen Biologie in Deutschland von 1900 bis zum Zweiten Weltkrieg", in: Hoßfeld U, Brömer R (eds): *Darwinismus und/als Ideologie. Verhandlungen zur Geschichte und Theorie der Biologie*, Band 6. VWB, Berlin, pp. 229–262.
- Malaterre, C. (2021). "Topic-modeling of multilingual non-parallel corpora: Applying machine-translation to a philosophy of

science corpus". Talk at the *DS² 2021 online Conference*, March 16, 2021. <https://youtu.be/FTzmpNYZs3E>

Malaterre, C.; Chartier, J.-F. and Pulizzotto, D. (2019). "What is this thing called philosophy of science? A computational topic-modeling perspective 1934–2015", in: *HOPOS*, 9 (2), pp. 215–249. <https://doi.org/10.1086/704372>.

Malaterre, C.; Lareau, F.; Pulizzotto, D. and St-Onge, J. (2020). "Eight journals over eight decades: a computational topic-modeling approach to contemporary philosophy of science. Synthese." <https://doi.org/10.1007/s11229-020-02915-6>

McCallum, A. K. (2002). "MALLET: A Machine Learning for Language Toolkit." <http://mallet.cs.umass.edu>.

McInnes, L. and Healy, J. (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction", in: *ArXiv e-prints* 1802.03426. <https://arxiv.org/abs/1802.03426v3>

Nicholson, D.J. and Gawne, R. (2015). "Neither logical empiricism nor vitalism, but organicism: what the philosophy of biology was", in: *HPLS* 37, pp. 345–381. <https://doi.org/10.1007/s40656-015-0085-7>

Noichl, M. (2019). "Modeling the structure of recent philosophy. Synthese." <https://doi.org/10.1007/s11229-019-02390-8>

Rehurek, R. and Sojka, P. (2010). [genism]. "Software framework for topic modelling with large corpora", in: *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, pp. 45–50. <https://radimrehurek.com/gensim/>

Smith, R. (2019). *tesseract 4.1.1*. <https://tesseract-ocr.github.io/>

Who CAREs, really?

Vom schwierigen Umgang mit digitalisierten Kulturgütern aus kolonialen Kontexten

Lange, Felix

felix.lange@zib.de
Zuse-Institut Berlin, Germany

Kuper, Heinz-Günter

kuper@zib.de
Zuse-Institut Berlin, Germany

Müller, Anja

anja.mueller@zib.de
Zuse-Institut Berlin, Germany

Amrhein, Kilian

amrhein@zib.de
Zuse-Institut Berlin, Germany

Klindt, Marco

klindt@zib.de
Zuse-Institut Berlin, Germany

Nowicki, Anna-Lena

nowicki@zib.de
Zuse-Institut Berlin, Germany

Spätestens seit dem Sarr/Savoy-Report (Sarr/Savoy 2018) ist die Diskussion über den Umgang europäischer Gedächtnisinstitutionen mit ihren Sammlungen aus kolonialen Kontexten neu [vgl. Savoy 2021] entbrannt. Welchen Ansprüchen muss in diesem Zusammenhang die Digitalisierung von Objekten des kulturellen Erbes genügen? Das Poster veranschaulicht diese Frage am Beispiel der "Ethnografica"-Sammlung im Nachlass von Karl Schmidt-Rottluff aus dem Bestand des Brücke-Museums Berlin. Die Digitalisierung¹ der Sammlung wird durch das Forschungs- und Kompetenzzentrum Digitalisierung Berlin (digiS)² in beratender Funktion begleitet. Schmidt-Rottluff, Gründungsmitglied der expressionistischen Künstlergruppe "Brücke" (1905–1913), sammelte in großem Umfang Skulpturen und Objekte aus Gebieten, die zum Teil unter deutscher Kolonialherrschaft standen. Provenienz und Erwerbsumstände sind in den meisten Fällen nicht endgültig geklärt. Die Art und Weise der digitalen Verfügbarmachung birgt daher im Vergleich zu europäischen Kulturgütern eine große Brisanz. digiS untersucht in diesem Zusammenhang bestehende Technologien hinsichtlich des Urheberrechts, der Dateninfrastruktur zur Veröffentlichung der Digitalisate und des Metadatenmodells.

Urheberrecht

Der Forderung von Sarr und Savoy nach einer systematischen Digitalisierung und Open-Access-Publikation von Kulturgut aus kolonialen Kontexten durch europäische Institutionen (Sarr/Savoy 2018: 58) halten Pavis und Wallace (2019) entgegen, dass dies einer Übertragung kolonialer Machtverhältnisse in den digitalen Raum gleichkäme. Dieses Argument spiegelt sich in den Forderungen indigener Gruppen in Australien, Neuseeland und Nordamerika nach "Indigenous Data Governance" (Carroll u.a. 2020) im Gegensatz zur unkontrollierten weltweiten Verfügbarmachung wider. Offensichtlich müssen also die FAIR-Prinzipien auch in Bezug auf digitale Objekte aus kolonialen Kontexten um ein Äquivalent des aus dem indigenen Bereich stammenden CARE-Regelwerks (dies. 2020) ergänzt werden, um gemeinsamen Nutzen und Mitspracherechte der betroffenen Gruppen zu gewährleisten. Digitalisierte Objekte aus außereuropäischen kolonialen Kontexten können also nicht pauschal als gemeinfreies Kulturgut betrachtet und also solches veröffentlicht werden.

Metadaten und Dateninfrastruktur

Eine wissenschaftliche Erschließung auf der Grundlage europäischer Erfassungskonventionen und für europäische Portale läuft Gefahr, außereuropäische Perspektiven nicht hinreichend abzubilden (vgl. Pavis/Wallace 2019). Damit ist der vieldiskutierte "Bias" in digitalen (Meta-)Daten angesprochen (van Erp/de Boer 2021). Eine objektive, operationalisierbare Definition dieses Begriffs wäre aus wissenschaftlicher Sicht wünschenswert, wird aber zu Recht bspw. von Blodgett u.a. (2020) zu Gunsten normativer Ansprüche an digitale Ressourcen verworfen, bei denen die Interessen von betroffenen marginalisierten Gruppen im Vordergrund stehen. Eine Vision zur Minimierung von Bias in diesem Sinne ist van Erps und de Boers (2021) "polyvoka-

les" Semantic Web. Demnach sollten Metadaten ihre Autorschaft in Daten, Abfragesprachen und User Interfaces explizit ausweisen. Aufbauend auf dieser Arbeit unternimmt digiS eine Analyse der im Museumsbereich gängigen Metadatenformate. Dabei müssen drei Ebenen differenziert werden: Das Austauschformat (bspw. LIDO³), Kontrollierte Vokabulare und die in dieses Gerüst geschriebenen Erschließungsinformationen. Wir untersuchen im Sinne der geforderten Polyvokalität problematische Begriffe in LIDO und erarbeiten Modellierungstechniken, die verschiedene Erschließungsperspektiven sichtbar machen. Bei den Kontrollierten Vokabularen folgen wir dem Ratschlag von Hardesty (2020), nach Möglichkeit Ressourcen zu verwenden, die von den darin genannten Gruppen selbst veröffentlicht werden.

Aus unserer Sicht ist der metadatenzentrierte Ansatz sinnvoll, muss aber durch die infrastrukturelle Ebene der Datenportale ergänzt werden. Denn die Veröffentlichung europäischer Metadaten auf einer großen europäischen Plattform wie Europeana wäre in den meisten Fällen die dominierende Erzählung zu den beschriebenen Objekten gegenüber kleineren lokalen Projekten. Auf der Ebene der Dateninfrastruktur sollen vielmehr (auch) die folgenden drei verschiedene Veröffentlichungsorte unterstützt werden:

1) *Das digitale Repositorium der sammelnden Institution* – in diesem Fall die Online-Sammlung des Brücke-Museums⁴. Hier könnten die außereuropäischen Sammlungsobjekte in den kunsthistorischen Kontext der Brücke-Gruppe eingeordnet und mit dem entsprechenden begrifflichen Instrumentarium erschlossen werden.

2) Durch die Präsentation der Sammlungsobjekte auf einer *global vernetzten, thematisch nicht spezialisierten Plattform wie Wikimedia Commons* wird der engere fachspezifische Kontext verlassen und eine angemessene Reichweite im Web gewährleistet. Das genannte Projekt setzt diesen Ansatz um⁵. Dabei sind eine sorgfältige, ethnografisch qualifizierte Auswahl des Materials und ein zunächst minimales Metadaten-set wesentlich, dass durch diverse externe Akteur:innen und Expert:innen nach dem Prinzip des Web 2.0 ergänzt werden kann.

3) Die dritte Veröffentlichungsoption betrifft direkt die Communities aus und in den Herkunftsregionen der Sammlungsobjekte. Ihnen soll durch *Exportschnittstellen* die Möglichkeit gegeben werden, Objekte in ihren eigenen digitalen Sammlungen ohne fremde Vorgaben zu präsentieren.

Unserer Ansicht nach müssen die drei Ebenen des Urheberrechtes, des Metadatenmodells und der Dateninfrastruktur gemeinsam in den Blick genommen werden, um eine angemessene digitale Repräsentation von Kulturgut aus kolonialen Kontexten zu ermöglichen.

Fußnoten

1. <https://www.bruecke-museum.de/de/museum/64/forschung>
2. <https://www.digis-berlin.de/>
3. <http://www.lido-schema.org/>
4. <https://www.bruecke-museum.de/de/sammlung/>
5. https://commons.wikimedia.org/wiki/Category:Karl_Schmidt-Rottluff%27s_Collection_of_Objects_from_Colonial_Contexts_in_the_Br%C3%BCke-Museum_Berlin

Bibliographie

Blodgett, Sue Lin / Barocas, Solon / Daumé III, Hal / Wallach, Hanna (2020): "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP". <https://arxiv.org/abs/2005.14050v2>.

Carroll, Stephanie Russo / Garba, Ibrahim / Figueroa-Rodríguez, Oscar L. / Holbrook, Jarita / Lovett, Raymond / Materechera, Simeon / Parsons, Mark / Raseroka, Kay / Rodriguez-Lonebear, Desi / Rowe, Robyn / Sara, Rodrigo / Walker, Jennifer D. / Anderson, Jane / Hudson, Maui (2020): "The CARE Principles for Indigenous Data Governance", in: *Data Science Journal*, 19(1): 43. DOI: <http://doi.org/10.5334/dsj-2020-043>.

Erp, Marieke van / de Boer, Victor (2021): "A Polyvocal and Contextualised Semantic Web", in: *The Semantic Web*, 506–12. Lecture Notes in Computer Science. Virtual Event: Springer International Publishing, 2021. <https://www.springerprofessional.de/the-semantic-web/19211334>.

Hardesty, Juliet (2020): "Mitigating Bias Through Controlled Vocabularies". Gehalten auf der *2020 DLF Forum*, ONLINE, 9. November 2020. <https://youtu.be/X0PVYgwHhVo>.

Pavis, Mathilde / Wallace, Andrea (2019): "Response to the 2018 Sarr-Savoy Report: Statement on Intellectual Property Rights and Open Access relevant to the digitization and restitution of African Cultural Heritage and associated materials", in: *JIPITEC* 10, Nr. 2 (10. Juli 2019). <https://www.jipitec.eu/issues/jipitec-10-2-2019/4910>.

Sarr, Felwine / Savoy, Bénédicte (2018): *Rapport sur la restitution du patrimoine culturel africain. Vers une nouvelle éthique relationnelle*. Paris, November 2018. http://restitutionreport2018.com/sarr_savoy_fr.pdf.

Savoy, Bénédicte (2021): *Afrikas Kampf um seine Kunst: Geschichte einer postkolonialen Niederlage*. München: C.H.Beck, 2021.

Zeitgeschichte untersuchen Topic Modeling von #blackouttuesday-Inhalten auf Instagram

Knierim, Aenne

aenne.knierim@stud.uni-regensburg.de
Universität Regensburg, Germany

Achmann, Michael

michael.achmann@ur.de
Universität Regensburg, Germany

Wolff, Christian

christian.wolff@sprachlit.uni-regensburg.de
Universität Regensburg, Germany

Einleitung/ Forschungsfrage

Die Black Lives Matter-Bewegung zur „Bekämpfung von schwarzem Rassismus auf der ganzen Welt“ entstand 2013 als Reaktion auf den Freispruch des weißen Polizisten George Zimmermann, der den 17-jährigen afroamerikanischen Schüler Trayvon Martin erschossen hatte (Black Lives Matter, n.d.). Um eine Bewegung über nationale Grenzen hinaus zu verbreiten, starteten drei afroamerikanische Aktivistinnen das Hashtag #blacklivesmatter (Black Lives Matter, n.d.). Dieser Prozess wird von Caliandro und Grahams als Grammatization bezeichnet, da das digitale Objekt Hashtag die Bewegung ermöglicht und strukturiert hat (2020). Nach Mundt et al. waren Soziale Medien für das Wachstum der Bewegung signifikant, da sie Beziehungen und Koalitionen mit anderen Gruppierungen der Bewegung ermöglichten und strategischen Aktionismus vereinfachten (2018).

Des Weiteren kann die Black Lives Matter-Bewegung auf Grund ihres transnationalen Charakters nach Vincent Millers Begriff des New Social Movement klassifiziert werden (2020). Miller zufolge sorgt "die ungehinderte Erstellung und Verbreitung von Informationen [...] für mehr Bewusstsein und Perspektiven zu Themen und Informationen" (2020). Genau das kann man an #blacklivesmatter beobachten: Nach dem Todeskampf des Afroamerikaners George Floyd, der gefilmt wurde und noch am selben Tag viral ging, erreichte die Popularität des New Social Movements einen neuen Höhepunkt. Das Ereignis führte zur Schöpfung eines neuen Hashtags: #blackouttuesday. Unter dem Hashtag posteten am 02.06.2020 mehr als 20 Millionen Menschen in Solidarität mit George Floyd ein schwarzes Quadrat auf Instagram. In den Beiträgen findet ein Diskurs über racial justice in Millionen von Beiträgen statt (Gallagher, 2017).

Caliandro und Graham argumentieren, dass Instagram die Mainstream-Medien als Raum für die Bekanntmachung und Diskussion relevanter gesellschaftlicher Themen ersetzt und nennt als Beispiel das Hashtag #blacklivesmatter (2020). Nach dem #blackouttuesday fand Diskurs um Rassismus in den herkömmlichen Medien statt, sodass einzelne Stimmen schwarzer Influencer*Innen, Autor*Innen und Aktivist*Innen Beachtung erlangten. Die Erforschung eines größeren Korpus würde Aufklärung über die Wahrnehmung der Thematik der Hashtag-Nutzer*innen eröffnen.

Nach Keightley und Daphi wird das kulturelle Gedächtnis unter anderem durch moderne Kommunikationstechnologien vermittelt (2020). Demnach dient das schwarze Quadrat des #blackouttuesday, das noch immer auf vielen Accounts sichtbar ist, als visuelle Erinnerung an die Ermordung Floyds und an systemischen Rassismus. Mithilfe der Methoden der Digital Humanities und durch die Untersuchung großer Korpora ist es nun möglich, die Stimmen der breiteren Öffentlichkeit am #blackouttuesday zu untersuchen. Das digitale Objekt Hashtag ist eine strukturelle Eigenschaft von Instagram, die als Marker für die wichtigsten Themen, Ideen, Ereignisse, Orte oder Emotionen eines Posts dient (Highfield und Leaver, 2015). Bisherige Forschung um #blacklivesmatter ist überwiegend qualitativ, quantitative Forschung konzentriert sich meist auf den #blacklivesmatter Diskurs auf Twitter. Dies ist nicht zuletzt damit zu erklären, dass die Plattform Instagram den Zugang zu seinem Feed nach dem Cambridge Analytics Skandal 2016 erheblich erschwert hat (Puschmann 2019, Bruns, 2019, Caliandro and Graham 2019). Da der #blackouttuesday auf Instagram seinen Ursprung hat und hauptsächlich dort stattfand, konzentriert sich dieses Projekt auf dieses Soziale Medium.

Datenerhebung

Da die schwarzen, quadratischen Posts kein spannendes Datenkorpus darstellen, ist das Ziel, deren Bildunterschriften, sog. Captions, inklusive ihrer Hashtags zu untersuchen. Nach Gallagher entstehen aus dem Netzwerk der Interaktionen in den Sozialen Medien Sujets und Themen, die den Rahmen jeder individuellen Konversation verlassen (2017). Die Themen, die sich aus diesem Diskurs entwickeln, werden wiederum von Individuen geformt (Gallagher 2017). Um diese Sujets, hier Topics, zu greifen und deren quantitative Analyse zu ermöglichen, soll ein Korpus aus den Bildunterschriften von Posts mit dem #blackouttuesday vom 02.06.2020 erstellt werden, diese sollen einschließlich der Hashtags in den Captions mit Selenium gecrawlt werden. Dies ist besonders wichtig, da die Verwendung eines Hashtags andere Absichten als die des ursprünglichen Nutzers widerspiegeln und irrelevant zum Medieninhalt sein können, einschließlich Spam. Das ist besonders an der aktuellen Nutzung des Hashtags zu sehen. Aus datenschutzrechtlichen Gründen können nur öffentliche Accounts gecrawlt werden. Dies stellt allerdings kein Manko dar, da private Nutzer*Innen häufig öffentliche Accounts nutzen.

Datenanalyse

Die erhobenen Daten sollen mittels Topic Modeling untersucht werden. Es wird ein besonderer Fokus auf die sinnhafte linguistische Datenvorverarbeitung gelegt. Auf Grund des spezifischen Sprachgebrauchs in den Sozialen Medien sowie der Nutzung von Emoticons stellt insbesondere die Tokenisierung des Korpus eine besondere Herausforderung dar (Singh und Sachan, 2017). Mit Topic Modeling wird es möglich, aus den Bildunterschriften und Hashtags Topics zu extrahieren und diese anschließend zu analysieren. Dafür wird der Latent Dirichlet Allocation benutzt (Blei et al., 2018). Für das Projekt soll sich auf englischsprachige Posts bezogen werden. Da der #blackouttuesday transnational verwendet wurde, ist es allerdings schwierig, eine genau geographische Grenze zu ziehen. Da sich die Bewegung in den USA gegründet hat, wird angestrebt, den Language Identifier „en-US“ zu verwenden. Die Ergebnisse sollen mit Word Clouds, Diagrammen und weiteren visualisiert werden. So können bereichernde Einblicke in die Gedanken und Gefühle der Unterstützer*innen des Black Live Matter Movements mit besonderem Blick auf #blackouttuesday gegeben werden.

Diskussion

Unsere Forschung soll das virale Event #blackouttuesday beleuchten und kann so zum besseren Verständnis von New Social Movements beitragen. Außerdem soll es qualitative Analysen zu #blacklivesmatter um eine neue quantitative Perspektive bereichern. Es soll ersten Aufschluss darüber geben, welche Topics allgemein durch #blacklivesmatter und spezifisch durch #blackouttuesday im gesellschaftlichen Diskurs durch das Soziale Medium Instagram relevant und sichtbar geworden sind.

Bibliographie

Bainotti, L./ Caliandro, A./ Gandini, A. (2021): "From archive cultures to ephemeral content, and back: Studying Instagram

Stories with digital methods.” In: *New Media & Society*, 23(12), 3656–3676. <https://doi.org/10.1177/1461444820960071>.

Bruns, A. (2019): “After the ‘APIcalypse’: social media platforms and their fight against critical scholarly research.” In: *Information, Communication & Society*, 22(11), 1544–1566. <https://doi.org/10.1080/1369118X.2019.1637447>.

Caliandro, A./ Graham, J. (2020): “Studying Instagram Beyond Selfies.” In: *Social Media + Society*, 6(2), 205630512092477. <https://doi.org/10.1177/2056305120924779>.

Gallagher, Ryan (2017): “Disentangling Discourse: Networks, Entropy, and Social Movements.” University of Vermont.

Gilowsky, Julia/ Heinrich, Horst-Alfred (2018): “Wie wird kommunikatives zu kulturellem Gedächtnis? Aushandlungsprozesse auf den Wikipedia-Diskussionsseiten am Beispiel der Weißen Rose.” In: G. Sebald (Hrsg.), (*Digitale*) *Medien und soziale Gedächtnisse*. Springer Fachmedien Wiesbaden.

Highfield, T. / Leaver, T. (2015): “A methodology for mapping Instagram hashtags.” In: *First Monday. Vorab-Onlinepublikation*. <https://doi.org/10.5210/fm.v20i1.5563>.

Marres, N. (2015): “Why Map Issues? On Controversy Analysis as a Digital Method. Science, technology & human values”. <https://doi.org/10.1177/0162243915574602>.

Merrill, S./ Keightley, E./ Daphi, P. (2020): “Social Movements, Cultural Memory and Digital Media” In: *Springer International Publishing*. <https://doi.org/10.1007/978-3-030-32827-6>.

Mundt, M./ Ross, K./ Burnett, C. M. (2018): “Scaling Social Movements Through Social Media: The Case of Black Lives Matter.” In: *Social Media + Society*, 4(4), 205630511880791. <https://doi.org/10.1177/2056305118807911>.

Pfeiffer, Jasmin (2018): “Rahmungen von Erinnerungen: Zur Metapher des Paratexts.” In: G. Sebald (Hrsg.), (*Digitale*) *Medien und soziale Gedächtnisse* (S. 281–299). Springer Fachmedien Wiesbaden.

Sebald, G. (2018): “(Digitale) Medien und soziale Gedächtnisse.” In: *Springer Fachmedien Wiesbaden*.

Singh, S. K. (2017): “Importance and Challenges of Social Media Text.”

Sommer, Vivien. (2018): “Mediatisierte Erinnerungen. Medienwissenschaftliche Perspektiven für eine Theoretisierung digitaler Erinnerungsprozesse.” In: G. Sebald (Hrsg.), (*Digitale*) *Medien und soziale Gedächtnisse*. Springer Fachmedien Wiesbaden.

Zeitler, Anna (2018): “#MediatedMemories: Twitter und die Terroranschläge von Paris im kollektiven Gedächtnis.” In: G. Sebald (Hrsg.), (*Digitale*) *Medien und soziale Gedächtnisse* (S. 123–143). Springer Fachmedien Wiesbaden.

Workshops

Annotorious

Eine JavaScript-Bibliothek für die Entwicklung maßgeschneiderter Bildannotationstools

Rainer, Simon

Rainer.Simon@ait.ac.at

Austrian Institute of Technology in Wien, Österreich

Radisch, Erik

e.radisch@gmx.de

Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland

Dieser Workshop richtet sich an Software-Entwickler im Digital Humanities-Bereich, die für die Konzeption und Umsetzung von Software-Tools, Forschungsinfrastrukturen und User Interfaces verantwortlich sind, bzw. in deren Arbeitsbereich der Umgang mit visuellen Medien fällt. Der Workshop behandelt speziell das Thema der digitalen Bildannotation, und zeigt, wie mit Hilfe der open source JavaScript-Bibliothek *Annotorious* mit geringem Aufwand maßgeschneiderte, an Projektbedürfnisse angepasste Bildannotations-Interfaces entwickelt werden können. Der Workshop geht dabei auch auf typische Use Cases ein, wie z.B. die semantische Annotation mit kontrollierten Vokabularen oder die Annotation zum Zweck des "ground truth building" für machine learning-Anwendungen. In praktischen Übungen vermittelt der Workshop die Grundlagen der Verwendung der Bibliothek, sowie die Möglichkeiten zur Integration in bestehende Infrastrukturen und der Anpassung und Erweiterung über die umfangreiche Programmierschnittstelle und das Plugin-Ökosystem.

Bildannotation in den Geisteswissenschaften

Als einer der sieben "scholarly primitives" (Unsworth 2000) kommt der Annotation in der geisteswissenschaftlichen Forschungspraxis eine wichtige Bedeutung zu. Die Idee, Dokumente mit Notizen oder Marginalien zu versehen, reicht mindestens bis zu den Manuskripten des Mittelalters zurück. Aber vor allem im digitalen Kontext eröffnet die Praxis der Annotation Wissenschaftlern neue Möglichkeiten, um Wissen und Forschungsergebnisse zu organisieren, zu teilen, sich kollaborativ mit anderen Forschern auszutauschen, und gemeinsam an der Analyse und Interpretation von Quellmaterial zu arbeiten (Barker und Terras 2016). Annotationen können dabei vielfältige Formen annehmen. Sie erweitern den Kontext, indem sie ein Quelldokument mit Zusatzinformationen anreichern, z.B. über Provenienz, Aufbau oder Autorenschaft; sie tragen zur Verbesserung der Auffindbarkeit in digitalen Sammlungen bei, insbesondere für Laien, die nicht oder wenig mit domänenspezifischer Terminologie vertraut sind (Hunter et al. 2008); sie machen die Struktur eines Dokuments transparent und explizit, und unterstützen damit bei der Analyse; oder sie liefern zusätzliche Details über bestimmte Aspekte, die bei Interpretation und Verständnis helfen können (Haslhofer et al. 2009).

Besonders in geisteswissenschaftlichen Projekten variieren die Anforderungen an Annotationswerkzeuge für digitale Inhalte sehr stark. Das betrifft nicht nur Fragen der grundlegenden Benutzerinteraktion und Usability, sondern vor allem die fachlichen und technischen Aspekte: zum Beispiel die Einbindung projektspezifischer Taxonomien und Eingabe-Schemata; die Anpassung von User Interface-Elementen an bestehende Arbeitsabläufe oder Kuratierungspraktiken; die Integration in bestimmte existierende Systeme wie ein spezifisches Sammlungs- oder Content-Management-System; oder die Anbindung an verschiedene Datenbank-Backends. Diese Heterogenität, und die oft sehr spezifischen Probleme geisteswissenschaftlicher Projekte führen in vielen Fällen dazu, dass überlegt wird, maßgeschneiderte Annotationswerkzeuge von Grund auf selbst zu entwickeln - was allerdings wiederum zeit- und ressourcenintensiv ist.

Genau für dieses Problem bietet Annotorious eine Lösung an. Einerseits bietet es vorgefertigte Basis-Komponenten, mit deren Hilfe mit geringem Aufwand ein umfangreiches Bild-Annotationstool in die eigene Forschungsumgebung eingebunden werden kann. Andererseits bietet es dabei aber viele Möglichkeiten, den Annotationsprozess individuell zu gestalten, und maßgeschneidert an die Bedürfnisse des jeweiligen Projektes anzupassen.

Über Annotorious

Annotorious ist eine JavaScript-Bibliothek mit deren Hilfe Bildannotationsfunktion einfach in eine bestehende Webseite oder eine Browser-basierte Webapplikationen eingebunden werden kann. Annotorious ist dabei bewusst keine off-the-shelf"-Lösung. Es bedarf der Erstellung einer eigenen Lösung, um Annotorious zu nutzen. Das mag im ersten Augenblick abschrecken. Allerdings lässt sich eine Basisvariante schon mit wenigen Codezeilen ein Standardszenario umsetzen, in welchem eine Annotationsebene an bestehende Bilder "angeheftet" wird. Je nachdem, welche Bedürfnisse die jeweilige Fragestellung mit sich bringt, stehen Benutzern verschiedene Zeichenwerkzeuge zur Verfügung (Rechteck, Polygon, Freihandzeichnen etc.). In einem weiteren Schritt mit den gezeichneten Markierungen verschiedene Zusatzinformationen verknüpft werden, z.B. Kommentare oder Tags (siehe Abbildung 1, 2 und 3). Hier liegt die wirkliche Stärke des Annotationstools. Sowohl die Zeichenwerkzeuge als auch der integrierte Annotations-Editor sind dabei erweiterbar. Sie lassen sich nach Bedarf um projektspezifische Elemente, wie zum Beispiel Eingabefelder für strukturierte Daten, ergänzen. Das Tool ist dadurch mit sehr wenigen Codezeilen hochgradig individualisierbar und je nach Nutzungsszenario erweiterbar, was es perfekt für geisteswissenschaftliche Problemstellungen macht.

Annotorious läuft ausschließlich im Browser, und hat keine server-seitigen Abhängigkeiten. Über die umfangreiche Programmierschnittstelle (API) lässt es sich nahtlos in bestehende Umgebungen integrieren, und an verschiedene graphische Anforderungen (visuelles Erscheinungsbild, Branding etc.) anpassen. Als Bildformate unterstützt Annotorious nicht nur die gängigen browser-kompatiblen Dateitypen (JPEG, PNG), sondern auch eine Reihe von Formaten für hochauflösende zoombare Bilder, wie insbesondere IIIF, Zoomify oder DeepZoom. Diese werden über den in den Geisteswissenschaften weit verbreiteten Image Viewer OpenSeadragon¹ unterstützt, der ebenfalls sehr einfach in bestehende Systeme integriert werden kann, und für den Annotorious als Plugin-Variante zur Verfügung steht. Um Annotationsdaten in einem möglichst interoperablen und zukunftssicheren Format zu halten, setzt Annotorious als Speicherformat das vom

World Wide Web Consortium standardisierte *W3C Web Annotation Data Model* ein.

Die Entwicklung von Annotorious wurde unter anderem durch *Pelagios* ermöglicht, einer internationalen Digital Humanities Initiative die sich insbesondere mit dem Thema der semantischen Annotation von räumlich-/geographischen Bezügen in Texten und Bildern beschäftigt. Die weitere Entwicklung läuft inzwischen aber unabhängig als selbstständiges Community-Projekt weiter, das die Bibliothek unter der BSD-3-Clause open source Lizenz zur uneingeschränkten Verwendung zur Verfügung stellt. Die Entwickler-Gemeinde wird dabei durch eine *Projektwebseite mit umfassender Dokumentation*, sowie einen aktiven *Support Chat-Kanal* tatkräftig unterstützt. Aktuelle Nutzer von Annotorious im geisteswissenschaftlichen Kontext, bzw. im Bibliotheksbereich sind z.B. das Metropolitan New York Library Council², die britische Crowdsourcing-Plattform MicroPasts³, oder das Projekt “Wissenschaftliche Bearbeitung der buddhistischen Höhlenmalereien in der Kuča-Region der nördlichen Seidenstraße” der Sächsischen Akademie der Wissenschaften.⁴

Lernziel

Ziel des Workshops ist es, den Teilnehmern umfassenden Einblick in den Funktionsumfang und in die Einsatzmöglichkeiten von Annotorious zu geben, und sie in die Lage zu versetzen, selbstständig Annotationsumgebungen zu erstellen, sie gemäß den eigenen Forschungsbedürfnissen individuell anzupassen, und über Plugins zu erweitern. Im Rahmen von praktischen Übungen werden die Teilnehmer dazu Basisszenarios auf dem eigenen Notebook entwickeln und Grundkomponenten (visuelles Erscheinungsbild, Annotations-Editor) anpassen. Weiters werden fortgeschrittene Themen angesprochen und die nötigen Konzepte vermittelt, die die Teilnehmer benötigen um auch komplexere Erweiterungen und eigene Plugins für Annotorious selbst zu schreiben.

Voraussetzungen

Die Teilnehmer benötigen ein eigenes Notebook um an den praktischen Übungen teilzunehmen, sowie grundlegende Kenntnisse der Web-Entwicklung (HTML, CSS, JavaScript). Fortgeschrittene Erfahrung mit JavaScript, bzw. die Installation erweiterter JavaScript Entwicklerwerkzeuge (node, npm) ist nicht zwingend erforderlich aber, vor allem hinsichtlich der fortgeschrittenen Themen (Entwicklung eigener Erweiterungen) von Vorteil. Die Teilnehmerzahl ist auf 15 Personen beschränkt. Ein separater Call for Papers ist nicht nötig.

Workshop Ablauf

Der Workshop wird einen halben Tag in Anspruch nehmen, und gliedert sich in 3 Teile (jeweils ca. 45min-1h30min), gefolgt von einem allgemeinen Frage & Antwort-Teil.

Im ersten Teil wird gemeinsam ein Basis-Setup von Annotorious in einer eigenen Webseite erstellt. Hierbei wird auf grundlegende Fragen wie Annotationsmodi (Zeichenwerkzeuge, headless mode, Annotorious standard vs. OpenSeadragon), Backend-Anbindung und Integration von bestehendem Benutzermanagement eingegangen. Die wichtigsten Funktionen und Konzepte der Pro-

grammierschnittstelle werden besprochen; und es wird anhand von zwei unterschiedlichen Beispielen gezeigt, wie Datenspeicher für Annotorious realisiert und angebunden werden können.

Im zweiten Teil werden bereits existierende Erweiterungen vorgestellt. Es wird in das Plugin-Ökosystem eingeführt, und gezeigt wie typische Use Cases (z.B. die Integration einer visuell an die eigenen Bedürfnisse angepassten Zeichenwerkzeugleiste, oder der Umgang mit mehrseitigen Bilddokumenten) mit existierenden Plugins stark vereinfacht werden können. Es wird näher auf das von Annotorious verwendete Datenformat, das *W3C Web Annotation Data Model* eingegangen, und der Zusammenhang zwischen Datenmodell und Editor-Komponenten besprochen. Es wird gezeigt Annotationsdaten sinnvoll in andere gängige Formate konvertiert werden können. Der Fokus liegt dabei insbesondere auf Formaten die relevant für die Annotation von Landkarten sind (GeoJSON oder WKT), bzw. Auf Formaten für die Verwendung von Annotationen als Trainingsdaten für Neuronale Netze (Coco-Dataset-Format).

Im letzten Teil wird eine kurze Einführung in die internen Plugin Schnittstellen von Annotorious gegeben, die die Entwicklung eigener Erweiterungen - z.B. neuer Zeichenwerkzeuge oder von Zusatzkomponenten für den Annotations-Editor - ermöglichen. Der Workshop schließt mit einem Überblick zu fortführendem Material, das den Teilnehmern das Selbststudium von fortgeschrittenen Themen ermöglicht, sowie einer allgemeinen abschließenden Frage-und-Antwort- bzw. Diskussionsmöglichkeit.

Benötigte technische Ausstattung

Es wird ein Beamer und ein Flipchart benötigt.

Beitragende

Dr. Rainer Simon ist Senior Research Software Engineer in der Forschungsgruppe Data Science & Artificial Intelligence am Austrian Institute of Technology in Wien, wo er sich vorwiegend mit der Entwicklung von digitalen Tools und Plattformen zur Unterstützung geisteswissenschaftlicher Forschung beschäftigt. Er arbeitet seit mehr als 15 Jahren im Bereich der Digital Humanities, und hat während dieser Zeit mit einer Vielzahl von Partner aus dem akademischen und dem GLAM (Galleries, Libraries, Archives, Museums) Bereich zusammengearbeitet. Er ist der Urheber und Haupt-Betreuer des Annotorious Open Source Projektes.

Dr. Erik Radisch ist wissenschaftlicher Mitarbeiter der Sächsischen Akademie der Wissenschaften zu Leipzig, wo er augenblicklich vor allem im Projekt “Wissenschaftliche Bearbeitung der buddhistischen Höhlenmalereien in der Kuča-Region der nördlichen Seidenstraße” involviert ist. Hier integrierte er ebenfalls Annotorious in die Projektseite und erweiterte die Annotationsmöglichkeiten um die Erfassung von Multipolygonen. Bereits vor seiner Arbeit an der sächsischen Akademie der Wissenschaften beschäftigte er sich intensiv mit wissenschaftlichen Methoden der Bildanalyse im Bereich der Digital Humanities.

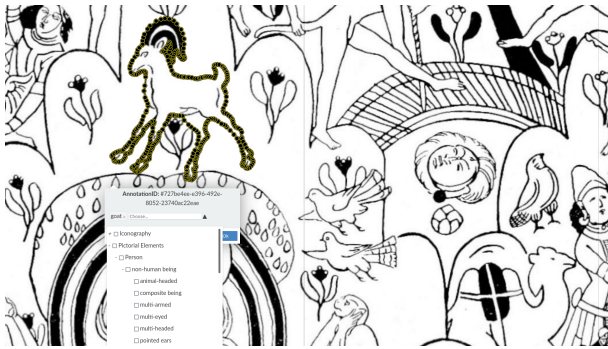


Abb. 1: Beispiel für eine Polygonannotation mit einem Annotationseditor, der eine Taxonomie in Baumstruktur integriert hat.

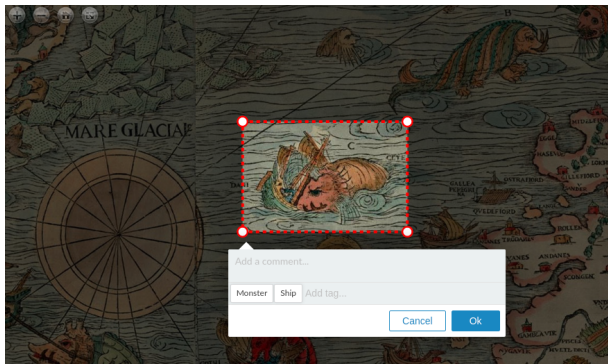


Abb. 2: Der Annotationseditor ermöglicht es, verschiedene Informationen mit der Bildannotation zu verknüpfen.

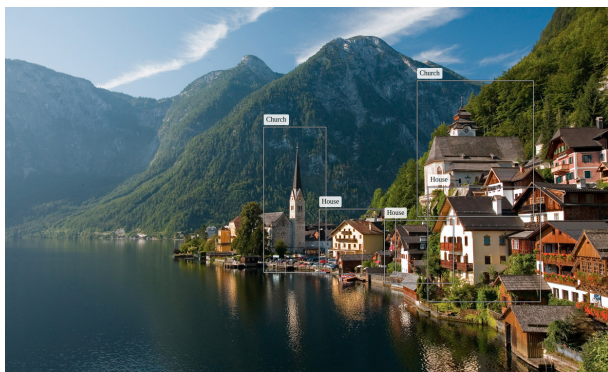


Abb. 3: Beispiel der Standardannotation mit Rechtecken.

Fußnoten

1. <https://openseadragon.github.io/examples/in-the-wild/> (zuletzt geprüft: 30.11.2021)
2. <https://github.com/esmero/archipelago-deployment>; <https://metro.org/> (zuletzt geprüft: 30.11.2021)
3. <https://crowdsourced.micropasts.org/> (zuletzt geprüft: 30.11.2021)
4. <https://kucha.saw-leipzig.de/> (zuletzt geprüft: 30.11.2021)

Bibliographie

Barker, E. / Terras, M. (2016) Greek literature, the digital humanities, and shifting technologies of reading. Oxford Handbooks Online, Oxford. DOI: 10.1093/oxfordhb/9780199935390.013.45.

Haslhofer, B./ Jochum, W./ King, R./ Sadilek, C./ Schellner, K. (2009) The LEMO Annotation Framework: Weaving Multimedia Annotations with the Web. In: *International Journal on Digital Libraries*, 10(1) : 15-32.

Hunter, J./ Khan, I./ Gerber, A. (2008) HarvANA – Harvesting Community Tags to Enrich Collection Metadata. In: *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*, Pittsburgh, Pennsylvania, United States, June 16–20, 2008: 147–156.

Unsworth, J. (2000) Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? In: *Humanities Computing: formal methods, experimental practice*. King's College, London, May 2000. Available at: <https://www.johnunsworth.name/Kings.5-00/primitives.html> (zuletzt geprüft: 30.11.2021).

Barcamp "Headlines & Highlights" der AG Zeitungen & Zeitschriften

Rißler-Pipka, Nanette

rissler-pipka@sub.uni-goettingen.de

Niedersächsische Staats- und Universitätsbibliothek Göttingen, Germany

Roeder, Torsten

dh@torstenroeder.de

Bergische Universität Wuppertal, Germany

Einleitung

Als Arbeitsgruppe "Zeitungen & Zeitschriften" im Verband Digital Humanities im deutschsprachigen Raum befassen wir uns in wissenschaftlichen und infrastrukturellen Kontexten mit historischen Zeitungen und Zeitschriften. Mit deren Digitalisierung, Digitalisaten, digitalen Präsentationsformen und -formaten sowie mit digitalen Analyseverfahren setzen wir uns kritisch auseinander und setzen uns für deren Weiterentwicklung ein. Die Bandbreite der möglichen Themen ist dementsprechend groß und orientiert sich am Interesse und Engagement der Mitglieder, die erfreulicherweise aus den unterschiedlichsten Kontexten stammen und damit einen offenen Diskurs zwischen Infrastruktur und Wissenschaft befördern, der aktuell vor allem im Kontext der NFDI Grundvoraussetzung für eine nachhaltige Weiterentwicklung ist. Um die ganze Bandbreite der laufenden und möglichen AG-Aktivitäten abbilden zu können und mit interessierten Menschen ins Gespräch zu kommen, die (noch) nicht aktives Mitglied der AG sind, haben wir uns für das Format eines Barcamps entschieden, bei dem sich verschiedene Kleingruppen konkreten Arbeitsfeldern

widmen können und deren Ergebnisse abschließend im Plenum präsentiert und diskutiert werden.

In einem halbtägigen Preconference-Barcamp (4 Stunden mit einer Pause) zu neun vorgeschlagenen Themen rund um die Arbeit der AG zu digitalen historischen Periodika wollen wir gemeinsam mit den Teilnehmer*innen aus und außerhalb der AG weitere Handlungsfelder erschließen und begonnene Aktivitäten voran bringen. Es handelt sich um ein offenes Angebot, das sich an alle richtet, die an dem Thema "Historische Zeitungen und Zeitschriften" interessiert sind und dazu ein spezifisches Arbeitsfeld in einer Kleingruppe vertiefen möchten. Die AG Zeitungen & Zeitschriften stellt vorhandene Kompetenzen bereit, möchte aber bewusst die offene Gestaltung durch die Teilnehmenden anregen. Zusammenhänge mit den AG-Aktivitäten sind natürlich willkommen, aber nicht notwendig. Es ist vorgesehen, dass die Ergebnisse der jeweiligen Kleingruppen am Ende des Barcamps vorgestellt werden und möglichst zeitnah in eine verwertbare und weiternutzbare Form überführt werden.

Themen

Die folgenden Themenvorschläge sind Angebote von AG-Mitgliedern. Die Liste kann im Vorfeld des Barcamps noch erweitert werden. Auf dem Barcamp selbst wird entschieden, welche davon in den Kleingruppen behandelt werden. Die Moderation wird jeweils von AG-Mitgliedern übernommen.

Agenda der DHd-AG Zeitungen & Zeitschriften: Visionen, Aufgaben und Ziele

Die AG möchte eine eigene "Agenda" erstellen, in der Interessen und Arbeitsfelder definiert und Visionen niedergelegt sind. In der Diskussion dazu wird zum einen die Abgrenzung des Forschungsgegenstands Zeitungen und Zeitschriften diskutiert sowie die Veränderung desselben im digitalen Zeitalter reflektiert. Mittel- und langfristige Ziele der AG in der internationalen Vernetzung von Forschenden und Anbietenden, in der Funktion eines Expertenforums, in der Nachwuchsförderung sowie in Consulting für Fördergeber bestehen. Ein erster Entwurf der Agenda liegt bereits vor und wird zur Diskussion gestellt, anschließend überarbeitet und zur weiteren Abstimmung an die gesamte AG übergeben.

AG-Workshopreihe: Von Metadaten bis zur Volltextanalyse

Hier trifft sich die Vorbereitungsgruppe des nächsten anstehenden Events aus der Reihe der "Methoden-Workshops", die im Jahr 2020 erfolgreich gestartet wurde und über einen längeren Zeitraum fortgesetzt wird. Mit den Teilnehmenden wird gemeinsam die Planung verfeinert, Bedarfe außerhalb der AG abgefragt und das bestehende Workshop-Konzept in Hinblick auf Lehrmethode, Zeitplanung und Vermittlungsziele diskutiert. Dabei fließen die Erfahrungen aus den drei vergangenen Workshops zu den Themen Metadaten und Korpuserstellung ein. Insbesondere werden thematische und konzeptionelle Ideen von Teilnehmenden außerhalb der AG begrüßt. Hier können auch konkrete Wünsche für die Vermittlung bestimmter Kompetenzen an die AG herangetragen werden.

Layout-Labor: OCR, OLR, Kodierungsfragen und Analysemethoden

In dieser Gruppe steht der fachliche Austausch über ein relativ neues Arbeitsfeld im Vordergrund. Aktuelle technische Entwicklungen im Bereich der Layout-Analyse und OCR/OLR werden hier vorgestellt, ausprobiert und diskutiert. Dazu sind gegebenenfalls Vorkenntnisse sowie technische Voraussetzungen erforderlich, die im Vorfeld bekannt gegeben werden. Der aktuelle Stand der Entwicklung und der noch bestehenden Bedarfe wird ausgetestet und z.B. in einem Blogartikel zusammengefasst. Ferner kann eruiert werden, inwieweit die existierenden Analysewerkzeuge für die Vermittlung an ein breiteres, technisch nicht in gleichem Maße versiertes Publikum auch in Form eines Methoden-Workshops bereits geeignet sind. Hier ergibt sich möglicherweise eine Schnittstelle zur DHd-AG „OCR“, mit der schon beim ersten AG-Workshop zum Thema OCR zusammen gearbeitet wurde.

FAIR Review: Full Text Corpora

Der letzte Workshop der AG befasste sich mit der Korpuserstellung vom Retrieval zu Balancing. In dieser Gruppe werden bestehende Ressourcen und Datenangebote einem Review hinsichtlich FAIR Data unterzogen. Ein Ziel kann darin bestehen, die aktuellen Datenangebote mit einer Beschreibung und einem Expert-Review zu versehen. Dies dient nicht nur den AG-Mitgliedern, sondern kann darüber hinaus auch auf der AG-Homepage oder in einem noch zu bestimmenden Review Journal als kollaborativer Beitrag veröffentlicht werden.

FAIR Review: Metadata

Ähnlich wie in der vorhergehenden Gruppe knüpft diese an bereits stattgefundene Workshops an. Hier steht das FAIR-Review von aktuellen Metadaten-Angeboten im Zentrum, ebenso mit dem Ziel, dies im Nachgang zu veröffentlichen. Analog zur Korpus-Idee bezieht sich das Review hier auf die angebotenen Metadaten, zu denen in vielen Fällen eine hinreichende Beschreibung fehlt. Daneben soll eine Einschätzung zur weiteren Verwendbarkeit dieser Daten gegeben werden.

Intermedialität - Multimodalität - Materialität

In dieser Gruppe geht es vornehmlich um die theoretische Debatte, die sich auch auf den Forschungsgegenstand der historischen Zeitungen und Zeitschriften bezieht: Wie kann das Medium in seiner Komplexität adäquat als digitales Objekt abgebildet werden? Welche Elemente müssen im Rahmen der Digitalisierung bereits bedacht und erfasst werden? Wie können ggf. flexibel später Merkmale von Intermedialität, Multimodalität und Materialität digital annotiert werden? Hier ergibt sich möglicherweise eine Schnittstelle zur DHd-AG "Theorie", die sich schon in Vorgesprächen angedeutet hat und bei dieser Gelegenheit vertieft werden kann.

TEI für Periodicals

Diese Gruppe schließt thematisch direkt an das Thema der Multimodalität an, diskutiert diese indessen aber aus der Sicht der kon-

kreten und aktuellen Möglichkeiten, welche in den TEI-Proposals gegeben sind. Die TEI bietet momentan noch keine ausreichenden oder praxisnahen Umsetzungs-Vorschläge für die Kodierung von Zeitungen und Zeitschriften. An Beispielen kann dies experimentell ausgelotet werden, woraus möglicherweise Erweiterungsvorschläge entwickelt werden können. Zusammenhänge mit OCR-Formaten als häufige Vorstufe einer TEI-Kodierung spielen hier ebenfalls mit hinein. Aber ist TEI überhaupt das universelle "Wunschformat" - oder benötigen bestimmte Use-cases andere Lösungen?

Internationale Community

Welche internationalen Arbeitsgruppen zu historischen Zeitungen und Zeitschriften bestehen derzeit? Welche Themen werden dort bearbeitet, welche Diskurse sind dort aktuell, wie ist dort die Balance zwischen Forschungs- und Angebotsseite? Lässt sich von deren Arbeit etwas lernen für die AG? Wie lassen sich Kontakte und Verbindungen, wie beispielsweise zur Special Interest Group "Periodicals" der Text Encoding Initiative, aufbauen oder intensivieren? Zudem liegt eine aktuelle Anfrage der Kooperation oder gemeinsamen Bildung einer entsprechenden Arbeitsgruppe im ADHO vor: Wie diese strategisch zu konkretisieren wäre, kann ebenfalls in dieser Gruppe diskutiert werden.

Öffentlichkeitsarbeit

Die AG Zeitungen und Zeitschriften pflegt eine Homepage, auf der eine umfangreiche Ressourcensammlung zum Thema enthalten ist. Diese Sammlung ist redaktionell zu überarbeiten, so dass sie auch einen Mehrwert für Personen außerhalb der AG bietet. Daneben sind die vorhandenen Informationen auf der Homepage (z.B. vergangene Veranstaltungen, Workshop-Material, etc.) insgesamt für die Nachnutzung aufzubereiten. Außerdem können auch Konzepte für weitere Informationsangebote wie z.B. Bibliographie oder Veranstaltungshinweise entwickelt werden.

Organisation

Die Vorbereitung der Gruppenarbeit erfolgt im Vorfeld durch die AG unter Berücksichtigung des offenen Inputs von Barcamp-Teilnehmer*innen und dem anvisierten Output. Für die Arbeit in den Kleingruppen wird eine schlichte Struktur vorgeschlagen: 30 Minuten für die Erläuterung des Themas durch die jeweiligen Moderator*innen sowie die Festlegung eines Ziels und einer Herangehensweise, gefolgt von 90 Minuten Arbeits- oder Diskussionszeit, und abschließend nochmals 30 Minuten Zeit, um gemeinsam die Ergebnisse zusammenzufassen, zu sichern und für eine kurze Präsentation vorzubereiten. Dabei können sowohl digitale als auch analoge Hilfsmittel verwendet werden - je nach Präferenz und Ausstattung: digitale Folien oder Pinnwand und Karten aus dem Moderatorenkoffer. Abschließend präsentiert jeder Tisch seine Ergebnisse. Diese werden unmittelbar im Speedblogging- oder Tweet-Verfahren verbreitet.

Wir benötigen einen großen Raum (s. Teilnehmerzahl) mit verteilten Tischen und einem Beamer. Begrüßenswert (aber nicht zwingend notwendig) wären klassische Kommunikationsmittel wie Whiteboard, Pinnwand und Moderationskoffer.

Wir erwarten maximal 50 Personen und erbitten eine möglichst verbindliche Anmeldung im Vorfeld, damit ggf. die Zahl der Kleingruppen vorausschauend angepasst werden kann. Eine

spontane Anmeldung vor Ort ist jedoch ebenso möglich. Bei der Anmeldung wird unverbindlich das Interesse an bevorzugten Themen erfragt, um das Themenangebot ggf. entsprechend anzugleichen.

Ablauf

30 Min - Begrüßung, kurze Sammlung der Themenvorschläge und Wahl der jeweiligen Moderatorinnen/Moderatoren, Aufteilung der Teilnehmenden in Kleingruppen

150 Min - Arbeit in Kleingruppen: 30 Min Einführung und gemeinsame Zielsetzung, 90 Min Diskussions- oder Arbeitszeit, 30 Min Zusammenfassung und Ergebnispräsentation

(währenddessen 15 Min Pausenzeit in Abstimmung mit der Organisation vor Ort)

45 Min - Ergebnispräsentationen im Plenum (jeweils max. 5 Min.) und Schlussdiskussion

Die Sprache der Erinnerung – analysieren und verstehen Korpuslinguistische Zugänge zu Oral-History-Daten

Gerstenberg, Annette

gerstenberg@uni-potsdam.de
Universität Potsdam, Germany

Leh, Almut

almut.leh@fernuni-hagen.de
Fernuniversität Hagen, Germany

Möbus, Dennis

dennis.moebus@fernuni-hagen.de
Fernuniversität Hagen, Germany

Pagenstecher, Cord

cord.pagenstecher@fu-berlin.de
Freie Universität Berlin, Germany

Die Motivation des Workshops ist es, die interdisziplinären Potentiale der Anwendung korpuslinguistischer Tools auszuloten, ohne die Spezifik und Sensibilität von Oral-History-Interviews aus dem Blick zu verlieren.

Beschreibung des Themas

Oral-History-Interviews sind narrative, meist lebensgeschichtliche Erinnerungsinterviews, die in der zeithistorischen Forschung, aber auch in den Sozial- und Kulturwissenschaften als Quellen bzw. Datengrundlage genutzt werden. Als Teil des kulturellen Erbes werden sie an verschiedenen Forschungs- und Gedächtniseinrichtungen gesammelt und für Sekundäranalysen aufbereitet. Interview-Archive sind Gedächtnisinstitution in zweifacher Hin-

sicht: einmal in der Bewahrung und Vermittlung von Wissensbeständen, die in ihrer Gesamtheit das kulturelle Gedächtnis einer Gemeinschaft bilden. Zusätzlich aber auch in einem unmittelbaren Sinn, insofern die hier archivierten Wissensbestände selbst Erinnerungen, also Gedächtnisinhalte, sind, und gehören als solche zum digitalen Gedächtnis. Und da die Audio- oder Videoaufzeichnung sowie die Transkripte vielfach in elektronischer Form vorliegen, gehören Interview-Archive zum digitalen Gedächtnis. Als maschinenlesbare Daten sind Oral-History-Interviews nicht nur für die Humanities, sondern auch für Linguistik und Informatik interessante multimodale, freilich wenig strukturierte Daten.

In den historischen Wissenschaften werden Oral-History-Interviews in intensiver hermeneutischer Arbeit analysiert und interpretiert, wobei sich die Fragestellung im Spannungsfeld persönlicher und kollektiver Relevanz situiert. Die digitale Erschließung dieser Quellen eröffnet die Möglichkeit, gerade auch überindividuelle Muster der Erinnerung und ihrer sprachlichen Verfasstheit mit Hilfe von korpuslinguistischen Tools zu erschließen.

Auf Basis von Praxisbeispielen wird im Workshop die Frage nach dem Mehrwert dieser technikgetriebenen Analysen diskutiert. Können solche Tools das hermeneutische Verstehen unterstützen oder gar bereichern, oder besteht die Gefahr, die Subjektivität der Erzählung und die individuelle Entstehungssituation der Quellen aus dem Blick zu verlieren? Sind die digital erkannten Muster am Ende Artefakte oder können sie Schlüssel zum tieferen Verstehen sein?

Die Motivation des Workshops ist es, die interdisziplinären Potentiale der Anwendung korpuslinguistischer Tools auszuloten, ohne die Spezifik und Sensibilität von Oral-History-Interviews aus dem Blick zu verlieren.

Im ersten Block des Workshops widmen sich drei Impulsreferate diesen Koordinaten des Themas. Zuerst wird thesenartig entwickelt, welche Implikationen die Digitalisierung für die Auswertung des Quellentyps Oral-History-Interviews hat (Almut Leh). Daran anschließend wird das Projekt Oral-History.Digital vorgestellt, das verschiedene Interviewbestände in einer webbasierten Erschließungs- und Forschungsumgebung verbindet und damit sammlungsübergreifende und vergleichende Zugänge ermöglicht (Cord Pagenstecher). Mit dem Interesse, die sprachliche Konstruktion der Erinnerung auszuloten, werden darauf aufbauend linguistische Fragestellungen der Pragmatik und Semantik entwickelt, die sich auf Oral-History-Interviews anwenden lassen (Annette Gerstenberg).

Im zweiten Block werden anhand eines vorbereiteten Arbeitskorpus ausgewählter Oral-History-Interviews drei Anwendungsszenarien vorgestellt. Sie zeigen, wie die in den Impulsreferaten entwickelten Fragestellungen konkret bearbeitet werden können.

Zur Einordnung wird zunächst das ausgewählte Arbeitskorpus vorgestellt: im Hinblick auf thematischen Schwerpunkt, Entstehungskontext und sprachliche Spezifika der enthaltenen Teiltex-te. Mit sprachstatistischen Basisdaten und automatisch ermittelten „Schlüsselwörtern“ werden Unterschiede der verwendeten Teiltex-te des Arbeitskorpus erläutert und visualisiert. Geplant sind darauf aufbauend drei Arbeitseinheiten, in denen jeweils eine in der Oral-History bisher wenig genutzte Analysemöglichkeiten im Mittelpunkt steht. In jeder Arbeitseinheit wird die Analyse demonstriert und zugleich die Möglichkeit gegeben, jeden Schritt selbst nachzuvollziehen.

(1) Auf Basis des lemmatisierten und nach Wortarten ausgezeichneten Datensatzes werden Häufigkeiten, von Wortarten und Grundwörtern untersucht. Als aufschlussreich hat sich zum Beispiel der Vergleich der Vorkommen von Pronomina (*ich* vs. *wir*) erwiesen (Knowles et al. 2021). Weiterhin werden die häufigsten

Verben semantisch kategorisiert, wobei Verben des Erinnerns und des Sagens besonders berücksichtigt werden.

(2) Ausgehend von der Analyse häufig vorkommender Wortfolgen (n-grams) werden Instanzen formelhaften Sprechens ermittelt. Gerade im Vergleich verschiedener Interviews erweisen sich solche als „Floskeln“ unterschätzte Redewendungen als charakteristisch und aussagekräftig für eine distanzierende oder aktualisierende Rahmung der erzählten Erinnerungen. Häufig verwendete und in der Analyse meist übersehene Wendungen wie *das ist alles lange her*, *das werde ich nie vergessen* oder *das hat uns geprägt* helfen dabei zu beobachten, wie das Erinnerte eingeordnet wird.

(3) Die Themen des Arbeitskorpus werden mit einer vorbereiteten Topic Modeling-Analyse, bei der lexikalische Cluster herausgearbeitet und visualisiert werden, vorgestellt. Dieser statistische Zugriff auf den Wortschatz wird in Kollokations-Analysen fortgesetzt und durch Kontextanalysen ergänzt. Dabei werden statistisch relevante gemeinsame Vorkommen ermittelt, aus denen deutlich wird, wie typische Erlebnisse sprachlich kodiert werden – wenn zum Beispiel das Kollokat von *Krieg* der *Schützengraben* ist oder zu *Gefangenen* häufig die Herkunft (*französische*, *russische*) angegeben wird.

Jedes Szenario wird mit einer technischen Anleitung zum Mitmachen verbunden. Auf diese Weise können wir zielgerichtet und konkret mit den Sprachdaten arbeiten und davon ausgehend weiterdenken. In der Diskussion wird es darum gehen, welche Aussagekraft wir den Ergebnissen zumessen und ob die unterschiedlichen Sichtweisen auf den gleichen Datenbestand neue Fragestellungen aufwerfen oder bekannte Interessen neu akzentuieren.

Helfen uns die digitalen Werkzeuge, den Interviews besser zuzuhören und die Sprache der Erinnerung besser zu verstehen? Als Ergebnis des Workshops versprechen wir uns Antworten auf diese Fragen der Erkenntnismöglichkeiten von Oral-History mit dem Werkzeugkasten der Korpuslinguistik.

Format

Halbtägiger Workshop, 7.3.2022

Arbeitsform: 1. Phase: Impulsreferate mit Diskussion, 2. Phase: Anwendungsszenarien von Korpus-Tools mit vorbereiteten Datensamples und Anleitung zur eigenen Datenexploration, 3. Phase: Diskussion der Ergebnisse und ihrer Forschungsrelevanz.

Der Workshop gibt Gelegenheit zur eigenen Datenexploration, wofür ein eigener Laptop notwendig ist. Die verwendeten Tools sind einfach zu installieren. Vor dem Workshop werden Datensätze zur Verfügung gestellt und Hinweise zu Tools gegeben, die installiert werden sollten, die verwendeten Werkzeuge sind sicher und gut handhabbar. Interessensbekundungen können mit spezifischen Fragen an Oral-History-Daten eingereicht werden, ggf. auch in Verbindung mit der Beschreibung eigener Datenbestände, die dann in die Workshoparbeit einbezogen werden können. Am Workshop kann auch ohne eigene Datenarbeit teilgenommen werden.

Die Vorbereitung der Texte wird thematisiert und mit Hilfe von Screenshots nachvollziehbar gemacht (Bereinigung aus der Datenbank extrahierter Daten; Erstellen der Nur-Text-Version; Tokenisierung, POS-Tagging, Lemmatisierung mit WebLight und TreeTagger).

Die gemeinsame Arbeit im Workshop nutzt diese vorbereiteten Dateien. Für die Analyse von Verbformen, Pronomina (1) und n-grams (2) werden robuste Werkzeuge (AntConc, TextPad) verwendet. Es werden Frequenzen ermittelt und Vorkommen im Kontext überprüft und für unterschiedliche Oral History-Quellen

verglichen; in der Diskussion werden die in den Eröffnungsreferaten aufgemachten Perspektiven auf die Ergebnisse bezogen. Das Topic Modeling (3) wird mit Python auf Jupyter-Konsolen in einem Google Colab durchgeführt, Datengrundlage werden die Transkriptionen lebensgeschichtlicher Interviews sein, als Topic Modeling-Engines kommen Gensim und dessen Mallet-Implementation zum Einsatz

Angaben zum Zielpublikum, insbesondere zu notwendigem Vorwissen

Wir laden Interessierte aus allen Disziplinen ein, die mit großen Textkorpora arbeiten und daran interessiert sind, die Anwendbarkeit computerlinguistischer Verfahren auf hermeneutische Probleme zu diskutieren. Es gibt die Möglichkeit, die Analyse mitzumachen oder nur zu verfolgen und die Relevanz der Ergebnisse auszuleuchten. Für den Workshop ist ein interdisziplinär zusammengesetztes Plenum aus den Digital Humanities sowie aus relevanten Themenfeldern wie Geschichtswissenschaften, Kulturwissenschaften, Literaturwissenschaft, Linguistik, Soziologie willkommen.

Interesse an Oral-History bzw. interviewbasierter Forschung in anderen Disziplinen und / oder an Anwendungsszenarien korpuslinguistischer Tools wird vorausgesetzt.

Zahl

10–30 Teilnehmerinnen und Teilnehmer

Technische Voraussetzung

Der Raum sollte mit Projektor und W-LAN ausgestattet sein. Eigene Laptops werden nach Verfügbarkeit mitgebracht.

Bibliographie

Apel, Linde / Almut Leh / Cord Pagenstecher (in print), “Oral History im digitalen Wandel. Interviews als Forschungsdaten“, in: Linde Apel (Hrsg.): *Erinnern, erzählen, Geschichte schreiben. Oral History im 21. Jahrhundert*.

Fechner, Martin / Andreas Weiß (2017): “Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts“, in: *Zeitschrift für digitale Geisteswissenschaften* 2. http://doi.org/10.17175/2017_005.

Gerstenberg, Annette (2017): “A Difficult Term in Context: The Case of French STO“, in: Erich Kasten / Katja Roller / Joshua Wilbur (Hrsg.): *Oral History Meets Linguistics*. Fürstenberg: Kulturstiftung Sibirien: 159–184.

Graham, Shawn / Ian Milligan / Scott Weingart (2014): *Exploring Big Historical Data: The Historian's Macroscope*. London.

Knowles, Anne Kelly / Paul B. Jaskot / Tim Cole / Alberto Giordano (2021): “Mind the Gap: Reading across the Holocaust Testimonial Archive“, in: Tim Cole / Simone Gigliotti (eds.): *The Holocaust in the 21st Century: Relevance and Challenges in the Digital Age*. United States: Northwestern University Press: 216–241.

Leh, Almut / Joachim Köhler / Michael Gref / Nikolaus P. Himmelmann (2019): “Audiovisual Data in Digital Humanities“, in: *VIEW Journal of European Television History and Culture* 7/14: 138.

Pagenstecher, Cord (2019): “Digital Humanities und biographische Forschung“, in: *BIOS* 30/1-2/2017: 76–91.

Pagenstecher, Cord / Stefan Pfänder (2017): “Hidden dialogues. Towards an Interactional Understanding of Oral History Interviews“, in: Erich Kasten / Katja Roller / Joshua Wilbur (Hrsg.): *Oral History Meets Linguistics*. Fürstenberg: Kulturstiftung Sibirien: 185–207.

Pagenstecher, Cord / Doris Tausendfreund (2013): “Das Online-Archiv ‘Zwangsarbeit 1939–1945‘“, in: Nicolas Apostolopoulos / Cord Pagenstecher (Hrsg.): *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt*. Berlin: Metropol: 71–96.

Philipps, Axel (2018): “Text Mining-Verfahren als Herausforderung für die rekonstruktive Sozialforschung“, in: *Sozialer Sinn* 19/2, S. 367–388.

Salman, Munir / Felix Engel / Almut Leh / Matthias Hemmje (2019): “Digital Humanities und biographische Forschung“, in: *BIOS* 30/1-2/2017: 92–100.

Einführung in DraCor Programmable Corpora für die digitale Dramenanalyse

Börner, Ingo

ingo.boerner@uni-potsdam.de
Universität Potsdam

Fischer, Frank

frank.fischer@dariah.eu
National Research University Higher School of Economics
Moscow; DARIAH-EU

Milling, Carsten

cmil@hashtable.de
Universität Potsdam

Sluyter-Gäthje, Henny

sluytergaeth@uni-potsdam.de
Universität Potsdam

Zielstellung des Workshops

In dem ganztägigen Workshop wird DraCor (<https://dracor.org>), eine offene Plattform zur Erforschung von Dramen in verschiedenen Sprachen, vorgestellt und anhand von praktischen Beispielen aus der digitalen Dramenanalyse erprobt. Im Zentrum von DraCor stehen so genannte ‘Programmable Corpora’. Hierunter verstehen wir infrastrukturell-forschungsorientierte, offene, erweiterbare, Linked-Open-Data-freundliche Volltextkorpora, die es ermöglichen sollen, auf niederschwellige Weise diverse Forschungsfragen aus dem Bereich der digitalen Literaturwissen-

schaft anhand von Korpora datenbasiert, nachvollziehbar und reproduzierbar zu bearbeiten (Fischer u. a. 2019).

Der Workshop richtet sich an Personen, die

- mit literarischen Texten und insbesondere mit Dramen arbeiten oder arbeiten möchten und dazu eigene Korpora erstellen oder bereits vorhandene Korpora nachnutzen möchten;
- Methoden der digitalen Dramenanalyse (Netzwerkanalyse, Stilometrie) erlernen oder auf Basis des Programmable Corpora-Ansatzes erproben wollen;
- Interesse an den Möglichkeiten zur Erforschung von literarischen Texten mithilfe von Linked Open Data (LOD) haben.

Es erfolgt eine Vorstellung des Konzepts der ‚Programmable Corpora‘ sowie einer Demonstration der exemplarischen Umsetzung in der Plattform DraCor inklusive einer Vorstellung aller Komponenten. In Form von Hands-on-Tutorials wird den Teilnehmer*innen eine praktische Einführung in das Erstellen und Kuratieren eigener Dramenkorpora zur Analyse mit DraCor gegeben. Ein weiterer Teil führt anhand praktischer Beispiele zu den Methoden Stilometrie und Netzwerkanalyse in die Verwendung der DraCor-API sowie der Python-Bibliothek PyDraCor ein. Die API-Schnittstelle (Application Programming Interface) ermöglicht den maßgeschneiderten direkten Zugriff auf bestimmte Teile der Korpora. Die Möglichkeiten zu korpusübergreifenden Abfragen und Einbeziehung von Informationen aus der Linked-Open-Data-Cloud mit SPARQL werden ebenso erprobt.

Das Konzept der ‚Programmable Corpora‘

Den Kern von DraCor bilden Korpora von Dramen in elf Sprachen (Deutsch, Russisch, Französisch, Italienisch, Schwedisch, Spanisch, Altgriechisch, Elsässisch, Lateinisch, Baschkirisch und Tatarisch) sowie zwei weitere Autoren-Korpora (Shakespeare, Calderón), zu denen die Plattform eine Vielzahl an möglichen Forschungszugängen bietet: Die Dramen sind als XML-Dateien entsprechend der TEI-Guidelines kodiert und unter einer offenen Lizenz frei über GitHub unter <https://github.com/dracor-org> verfügbar. Sie können von dort geladen, gegebenenfalls selbst transformiert oder angereichert und zur weiteren Beforschung in beliebigen Tools weiterverwendet werden.

Neben diesem „klassischen“ modus operandi der korpusbasierten Forschung bietet DraCor als offenes digitales Ökosystem jedoch noch weitere Schnittstellen und angeschlossene Tools (Netzwerkvisualisierungen, Shiny App, Easy Linavis). Grundlegend hierfür ist die DraCor REST API (<https://dracor.org/doc/api>), die sowohl Funktionen zum Abrufen der Daten in unterschiedlichen Formaten (TEI, JSON, Plaintext, RDF, GEXF, GraphML) als auch einige eingebaute Analysefunktionalitäten (bspw. zu Netzwerkmetriken) bereitstellt. Über die API können neben Struktur- und Metadaten auch die Volltexte ohne weiteres Markup abgerufen werden, um so ohne weiteren Zwischenschritt zur Entfernung von Markup Methoden wie stilometrische Analysen oder Topic Modeling anzuwenden. Die DraCor API ist im OpenAPI-Standard dokumentiert und kann in einer mittels Swagger UI implementierten interaktiven Dokumentation (<https://dracor.org/documentation/api>) direkt aus dem Webbrowser heraus verwendet werden.

Für die Programmiersprachen Python (PyDraCor: <https://github.com/dracor-org/pydracor>) und R (rdracor: <https://github.com/dracor-org/rdracor>) sind API-Bibliotheken verfügbar, die eine schnelle und auf die jeweilige Programmiersprache angepasste Einbindung der API-Funktionalitäten ermöglichen. Für

komplexe Abfragen steht auf der Plattform ein SPARQL-Endpoint (<https://dracor.org/sparql>) zur Verfügung. Hierüber sind sowohl korpusübergreifende als auch kombinierte Abfragen (federated queries) möglich, bei denen DraCor gleichzeitig mit anderen als LOD verfügbaren Ressourcen, wie beispielsweise Wikidata, abgefragt werden kann.

Digitale Dramenanalyse mit DraCor

Korpusbasierte, in der Regel quantitative Methoden verwendende Analysen von Dramen haben sich in den vergangenen Jahren zu einem eigenen Subfeld der Computational Literary Studies (CLS) entwickelt (vgl. Willand et al. 2017; Reiter 2021). Dabei hat sich die Bereitstellung gemeinsam kuratierter und offener Ressourcen wie DraCor als produktiv auch für angrenzende Disziplinen wie die Computerlinguistik erwiesen (vgl. beispielsweise Pagel, Reiter 2020).

Auf Wortebene operierende Verfahren haben sich dabei etwa auf die Autorschaftsattributions (Schöch 2014) oder Genreklassifikation mit Topic Modeling (Schöch 2017) fokussiert. Aktuell werden vielversprechende Neukonzeptualisierungen stilometrischer Maße wie das Kontrastmaß Zeta entwickelt und angewendet (Schöch 2018). Auf der Grundlage von strukturell ausgezeichneten Korpora lassen sich darüber hinaus gezielte Analysen etwa von Bühnenanweisungen durchführen, die mit POS-Informationen oder semantischen Feldern operieren (Trilcke et al. 2020).

Im Bereich der strukturellen Analyse wurden Dramenkorpora früh schon, beginnend mit den Arbeiten von Stiller, Nettle, Dunbar (2003) und fortgesetzt etwa bei Moretti (2011), mit netzwerkanalytischen Ansätzen untersucht. Typologische Arbeiten beispielsweise zum Konzept der Small Worlds (Trilcke et al. 2016) stehen hier u.a. neben Ansätzen zur quantitativen Klassifizierung von Figurentypen (Fischer et al. 2018).

Wenngleich semantische Technologien mittlerweile zum festen Bestandteil des Methodenspektrums der Digitalen Geisteswissenschaften zählen, gelangen sie in den korpusbasierten CLS bisher selten Anwendung (zu Prosa bspw. Frank und Ivanovic 2018; Dittrich 2017). Die Erfassung von Metadaten als Linked Data und die Anbindung an externe Referenzressourcen, insbesondere Wikidata, ermöglichen jedoch weitreichende Abfragemöglichkeiten und lassen sich zur Analyse von literarischen Korpora gewinnbringend nutzen. Beispielsweise sind in den DraCor-Korpusdaten keine detaillierten Informationen zu Autor*innen und Aufführungsorten enthalten. Da aber zu den einzelnen Stücken die eindeutigen Wikidata-Identifikatoren hinterlegt sind, können diese Informationen per federated queries in SPARQL abgerufen und in unterschiedlichen Visualisierungsformen, wie zum Beispiel als Karte, dargestellt werden.

Lernziele und Ablauf des Workshops

Im ersten Teil des Workshops wird zunächst das Konzept der ‚Programmable Corpora‘ eingeführt und diskutiert. Daran anschließend werden die Plattform DraCor und die einzelnen Komponenten vorgestellt, wobei auch immer wieder kürzere Übungsphasen vorgesehen sind, in denen die Teilnehmer*innen die vorgestellten Komponenten und Tools unmittelbar ausprobieren können. Insbesondere werden die unterschiedlichen Möglichkeiten zum Bezug und zur Analyse der Korpusdaten erprobt. Ein Fo-

kus liegt dabei auf der Verwendung der API. Anhand der interaktiven Dokumentation werden die API-Funktionalitäten erläutert und können von den Teilnehmer*innen ausgiebig getestet werden. Im Anschluss daran wird ein kurzer Überblick zur Korpuserstellung und zu den Besonderheiten der TEI-Kodierung geben, wie sie in DraCor zum Einsatz kommen.

Den zweiten Teil des Workshops bilden Gruppenarbeitsphasen, in denen drei Themenbereiche vertieft werden können:

(1) Korpuserstellung und -kuratierung mit DraCor: Die Teilnehmenden vertiefen die TEI-Kodierung von Dramen anhand von praktischen Übungen und lernen, wie eine lokale Instanz der Plattform mittels Docker aufgesetzt, gegebenenfalls angepasst und mit eigenen Korpora bestückt werden kann.

(2) Dramenanalyse mit DraCor-API und Python: Mittels Jupyter Notebooks mit ausführlich dokumentiertem Python-Programmcode werden die Teilnehmer*innen an Methoden der digitalen Dramenanalyse unter Verwendung der DraCor-API herangeführt. Die Notebooks sollen es auch Teilnehmer*innen, die bisher noch keine Erfahrungen im Programmieren mit Python gemacht haben, im Sinne eines Literate-Programming-Ansatzes ermöglichen, die einzelnen Analyseschritte nachzuvollziehen und auch selbst adaptieren zu können. Die Notebooks setzen konkrete Forschungsfragen zur Dramenanalyse um, etwa zur literaturhistorischen Entwicklung netzwerkanalytischer Maße oder zur quantitativen Dominanz von Figuren.

(3) Dramenanalyse mit Linked Data: Den Schwerpunkt bilden praktische Analysen, die aus der Anbindung von DraCor an die Linked Open Data Cloud möglich werden. Im Workshop wird ein kurzer Crashkurs in die Abfragesprache SPARQL gegeben, um dann im Anschluss gemeinsame Abfragen von DraCor und Wikidata vorzunehmen und die Ergebnisse zu visualisieren.

Die Ergebnisse der Arbeitsgruppen werden anschließend im Plenum präsentiert und diskutiert.

Organisatorisches

Anzahl der möglichen Teilnehmer*innen: 25

Teilnehmer*innen benötigen einen eigenen Laptop mit Internetzugang; Hinweise zu vorab zu installierender Software (Oxygen XML-Editor, Docker, ...) werden im Vorfeld bekanntgegeben. Die Materialien werden auf GitHub bereitgestellt; die Jupyter Notebooks werden unter (<https://github.com/dracor-org/dracor-notebooks>) veröffentlicht.

Weitere benötigte technische Ausstattung am Veranstaltungsort: Beamer, WLAN

Beitragende / Kontaktdaten

Ingo Börner (ingo.boerner@uni-potsdam.de) arbeitet als wissenschaftlicher Mitarbeiter im Projekt „CLSIInfra“ an der Universität Potsdam an der Weiterentwicklung von DraCor. Seine Arbeitsschwerpunkte umfassen Datenmodellierung und Linked Open Data.

Frank Fischer (frank.fischer@darjah.eu) ist Associate Professor an der Higher School of Economics in Moskau und einer der Direktoren von DARIAH. Seine Beschäftigung mit digitaler Dramenanalyse geht zurück auf das Projekt zur Digitalen Literaturwissenschaftlichen Netzwerkanalyse DLINA (<https://dlina.github.io>), aus dem DraCor hervorgegangen ist.

Carsten Milling (cmil@hashtable.de) ist Webdeveloper und ist im Projekt „CLSIInfra“ an der Universität Potsdam für die Entwicklung der DraCor-Plattform zuständig.

Henny Sluyter-Gäthje (sluytergaeth@uni-potsdam.de) ist wissenschaftliche Mitarbeiterin an der Professur für deutsche Literatur des 19. Jahrhunderts an der Universität Potsdam. Sie hat ein Masterstudium of Science in Cognitive Systems mit Schwerpunkt Computerlinguistik abgeschlossen und arbeitet an der algorithmischen Verarbeitung literarischer Texte.

Fördernachweis

DraCor wird gegenwärtig im Rahmen des von der EU Horizon 2020 geförderten Projekts „CLSIInfra“ (Fördernummer: 101004984, <https://cordis.europa.eu/project/id/101004984>) weiterentwickelt.

Bibliographie

Dittrich, Andreas (2017): "Intra-Connecting an Exemplary Literary Corpus with Semantic Web Technologies for Exploratory Literary Studies" in: Bański, Piotr et al. (Hg.): *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017*. Mannheim: Institut für Deutsche Sprache. <https://nbn-resolving.org/urn:nbn:de:bsz:mh39-62441>.

Fischer, Frank / Trilcke, Peer / Kittel, Christopher / Milling, Carsten / Skorinkin, Daniil (2018): "To catch a protagonist: Quantitative dominance relations in german-language drama (1730–1930)" in: Digital Humanities 2018. Conference Abstracts. Mexico City: El Colegio de México / Universidad Nacional Autónoma de México / Red de Humanidades Digitales 193–201.

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtel, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): "Programmable Corpora: Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor" in: *DHd2019: »Digital Humanities: multimedial & multimodal«. Book of Abstracts*. Mainz/Frankfurt a. M.: Johannes Gutenberg Universität Mainz / Goethe Universität Frankfurt, 194–197.

Frank, Andrew / Ivanovic, Christine (2018): "Building Literary Corpora for Computational Literary Analysis – A Prototype to Bridge the Gap between CL and DH" in: Calzolari, Nicoletta et al. (Hg.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association.

Moretti, Franco (2011): "Network Theory, Plot Analysis" in: *Stanford Literary Lab Pamphlets 2*. <http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf> [letzter Zugriff 13.7.2021].

Pagel, Janis / Reiter, Nils (2020): "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German" in: *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Marseille 55–64 <http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf> [Letzter Zugriff: 15.7.2021].

Reiter, Nils (2021): "Möglichkeiten Quantitativer Dramenanalyse" in: *Comparatio. Zeitschrift für Vergleichende Literaturwissenschaft* 12(2): 39–52.

Schöch, Christof (2017): "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama" in: *Digital Humanities Quarterly* 11, Nr. 2 <http://www.digitalhu>

manities.org/dhq/vol/11/2/000291/000291.html [Letzter Zugriff: 15.7.2021].

Schöch, Christof (2018): "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie" in: Bernhart, Toni et al. (Hg.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin: de Gruyter 77–94 doi: 10.1515/9783110523300-004.

Schöch, Christof (2014): "Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik" in: Schneider, Lars / Schöch, Christof (Hg.): *Literaturwissenschaft im digitalen Medienwandel*. Beihefte zu Phin 7 <http://web.fu-berlin.de/phin/beihefte7/b7t08.pdf> [Letzter Zugriff: 15.07.2021].

Stiller, James / Nettle, Daniel / Dunbar, Robin I. M. (2003): "The Small World of Shakespeare's Plays" in: *Human Nature* 14: 397–408.

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkasper, Dario / Kittel, Christopher (2016): "Theatre Plays as ›Small Worlds‹? Network Data on the History and Typology of German Drama, 1730-1930" in: *Digital Humanities 2016. Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków 385-387 https://dh2016.adho.org/abstracts/static/dh2016_abstracts.pdf [Letzter Zugriff: 15.07.2021].

Trilcke, Peer / Kittel, Christopher / Reiter, Nils / Maximova, Daria / Fischer, Frank (2020): "Opening the Stage. A Quantitative Look at Stage Directions in German Drama" in: *Digital Humanities 2020. Conference Abstracts*. Ottawa: University of Ottawa https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html [Letzter Zugriff: 15.07.2021].

Willand, Marcus / Trilcke, Peer / Schöch, Christof / Reißler-Pipka, Nanette / Reiter, Nils / Fischer, Frank (2017): "Aktuelle Herausforderungen der Digitalen Dramenanalyse" in: *DHD 2017. Digitale Nachhaltigkeit. Konferenzabstracts*. Bern: Universität Bern 175–180 doi: 10.5281/zenodo.3684825.

Ethisch - transparent - offen

Die CARE-Prinzipien und ihre Implikationen für geisteswissenschaftliche FDM-Services

Moeller, Katrin

katrin.moeller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Germany

Söring, Sibylle

sibylle.soering@fu-berlin.de
Universitätsbibliothek der Freien Universität Berlin, Germany

Imeri, Sabine

sabine.imeri.1@ub.hu-berlin.de
Universitätsbibliothek und Fachinformationsdienst Sozial- und Kulturanthropologie an der Humboldt-Universität zu Berlin, Germany

Lemaire, Marina

marina.lemaire@uni-trier.de
Servicezentrum eSciences der Universität Trier

Reichert, Nils

Nils.Reichert@hla.hessen.de
Hessisches Landesarchiv Marburg

Einführung

Format: Workshop, 4 h (halbtags)

Zielgruppe: FDM-Beratungspersonal und -Infrastrukturentwickler:innen; Fachwissenschaftler:innen

Gruppengröße: max. 30 Teilnehmende

Techn. Ausstattung: Beamer, Whiteboards/Pinnwände, Medienkoffer

Bei den Einreichenden handelt es sich vor allem um Vertreter:innen von Datenzentren und Infrastruktureinrichtungen (AG Datenzentren des DHd-Verbandes), deren Aufgabe es u.a. ist, Forschende bei der Entwicklung und Umsetzung des Forschungsdatenmanagements (FDM) in den Geistes-, Sozial- und Kulturwissenschaften zu unterstützen und Forschungsinfrastrukturen sowie Daten für diese Disziplinen bereitzustellen. Dabei fallen häufig schon Beratungs- und Kompetenzvermittlungsaufgaben an, die weit in rechtliche und ethische Thematiken ausgreifen, für die Mitarbeiter:innen von Datenzentren und institutioneller Infrastruktureinrichtungen häufig aber nicht umfassend ausgebildet oder geschult sind, und für die aus geisteswissenschaftlicher Perspektive auch noch kaum Handreichungen existieren.

Während die Einreichenden im Rahmen des Workshops ihre disziplinäre und infrastrukturelle Expertise und Erfahrung in der FDM-Beratung und dem Aufbau von FDM-Services einbringen, werden fachliche Beiträge zu Anwendungskontexten, zur Relevanz und disziplinären Ausweitung der CARE-Prinzipien durch die Ethnologin Sabine Imeri, den Juristen Thomas Henne, das Netzwerk "Koloniale Kontexte" und den Historiker Cord Pagensteher eingebracht. Die Verschränkung von FDM-Expertise und Problemszenarien der interdisziplinären Forschungspraxis soll beide Handlungsebenen des FDMs besser miteinander in Bezug setzen und die gegenseitige Wahrnehmung von Bedarfen, Herausforderungen und Lösungskonzepten fördern.

Workshopidee und Zielgruppe

Während die FAIR-Prinzipien (Wilkinson 2016; Kraft 2017) die Aufmerksamkeit vor allem auf Eigenschaften von Forschungsdaten als Voraussetzung für gelingenden und nachhaltigen Datenaustausch richten, werden ethische Fragestellungen, Machtdynamiken und historische Kontexte des Umgangs mit Forschungsdaten bisher nicht systematisch im FDM berücksichtigt. Ein wichtiger Impuls hierzu wurde mit den CARE-Prinzipien (GIDA 2019) für die Handhabung von Forschungsdaten geschaffen, die sich auf Indigene Gruppen und Gemeinschaften beziehen: Die Initiative fordert dazu auf, komplementär zu FAIR mit den vier Dimensionen Collective Benefit, Authority to Control, Responsibility und Ethics auch die Zwecke von Datentransparenz und -austausch sowie deren Auswirkungen auf Indigene Gruppen und Gemeinschaften regelmäßig zu reflektieren. Die CARE-

Prinzipien ergänzen die FAIR-Prinzipien in der Absicht, in der Open Science- und Open Data-Bewegung "Indigene Datensouveränität" (Carroll et.al 2020) und damit die Selbstbestimmungsrechte Indigener Gruppen (UNDRIP 2007, insb. Art. 31) mit Blick auf deren Wissen und kulturelles Erbe zu stärken (vgl. z.B. das digitale Archiv der Passamaquoddy People). In einer breiteren Perspektive scheinen diese Überlegungen geeignet, generell unterschiedliche Aspekte von Verantwortung in Prozessen der Erzeugung und Öffnung von Forschungsdaten zu thematisieren. Darunter fallen etwa die zeitlichen Dimensionen der Archivierung von Forschungsdaten oder die Historizität von Standards und Regelwerken sowie von Forschungs- und Erhebungsmethoden. Gleichfalls sind ethische Belange stärker zu berücksichtigen, etwa mit Blick auf Material aus kolonialen Kontexten, der Darstellung bestimmter jeweils stigmatisierter Menschen (z.B. uneheleiche oder behinderte Menschen in bestimmten Gesellschaften) oder Quellen, die Übergriffe im Kontext von Diktaturen thematisieren. Fragestellungen reichen dabei vom Umgang mit Beutegut und kolonialen Darstellungen in Museen/Forschungsdaten/Vokabularen bis hin zur Verwendung von Ortsnamen der Vergangenheit oder den Umgang mit Methoden der verstehenden Deutung.

In vielen geistes- und kulturwissenschaftlichen Forschungsprojekten sind solche Fragen virulent, da diese Sachverhalte nicht nur personengebundene Daten betreffen, die aufgrund gesetzlicher Regelungen noch einen besonderen Schutz genießen. Auch Datenzentren, die historische Daten anbieten, müssen ihre gesellschaftliche Verantwortung reflektieren und die vertretenen ethischen Prinzipien transparent machen. So wie momentan um Ausstellungskonzeptionen mit kolonialem Erbe oder die Benennung von Straßen und Einrichtungen gerungen wird, sind ähnliche Debatten auch hinsichtlich von öffentlichen Datenbeständen zu erwarten. Es bleibt fraglich, ob solche Probleme allein durch eine kritische Positionierung (Quellenkritik) bearbeitbar sind. Gleichzeitig müssen forschungsspezifische Prinzipien (Erhalt der Kontextualisierbarkeit) beachtet werden, welche die Einbettung von Wissen und ihre Rezeption rekonstruieren können und müssen, um sie zeitspezifisch zu analysieren (Schwerhoff 1992). Die geisteswissenschaftlichen Forschungsdatenzentren, die sowohl Forschungsprojekte bei der Entwicklung ihrer FDM-Strategien beraten als auch Infrastrukturen für die Datenerhebung, -aufbereitung, -analyse, -publikation und -archivierung bereitstellen, verfügen in dieser Hinsicht bisher kaum über professionelle Strukturen. Dies betrifft sowohl Handlungskonzepte in der Beratungspraxis, der Infrastrukturentwicklung und -bereitstellung im Hinblick auf CARE, als auch die Frage, wie ggf. widersprechende Anforderungen der FAIR-Prinzipien oder fachspezifischen Methoden hier in Einklang zu bringen sind. Eine Umfrage im Kreis der Vertreter:innen der DHd AG Datenzentren zeigte, dass hier weder umfangreiche Kompetenzen verfügbar, noch solche Fragestellungen bisher überhaupt Teil einer größeren Debatte sind. Weder können Forschungsprojekte hinsichtlich ethischer oder rechtlicher Belange adäquat beraten werden, noch lässt sich bisher einschätzen, welche Maßnahmen in der Entwicklung und für den Betrieb von Forschungsinfrastrukturen ergriffen werden müssen, um unter Berücksichtigung der CARE-Prinzipien Daten zu verarbeiten, bereitzustellen und zu kontextualisieren. Ebenso bestehen meist keine Ressourcen im weiteren Feld der Einrichtungen, um reguläre Services und Dienste hierfür anbieten oder vermitteln zu können.

Vor diesem Hintergrund soll der Workshop die Möglichkeit geben, die vielfältigen interdisziplinären Herausforderungen und Lösungsansätze in Theorie und Praxis kennenzulernen. So existieren bereits verbindliche Regularien und Prinzipien im Bereich der ethischen Herausforderungen für medizinische Versuche oder

Beratungen (Frewer/Bruns/May 2012) bzw. längere Debatten in der qualitativen Sozialforschung und der Ethnologie (von Unger/Dilger/Schönhuth 2016), die hier Impulse zur Diskussion bieten. Ebenso gibt es in einem größeren Kontext Überlegungen zur Treuhänderschaft von Daten, die als neue Grundlage die Berechtigung von Institutionen zur Datenhaltung und -weitergabe thematisieren (RfII 2021). Denn nicht nur die Kompetenzen der einzelnen Mitarbeiter:innen müssen sich hinsichtlich dieser Rahmenbedingungen weiterentwickeln, auch die gesetzlichen Grundlagen der Institutionen bleiben bisher auf einzelne Einrichtungstypen orientiert. So gelten bspw. besondere gesetzliche Grundlagen im Umgang mit personenbezogenen Daten spezifisch für Archive, nicht aber für Datenzentren.

Programm des Workshops

Die Diskussion und Erarbeitung von Richtlinien für einen ethisch angemessenen Umgang mit Forschungsdaten in geisteswissenschaftlichen Disziplinen ist daher insgesamt noch wenig konsequent verfolgt worden. So bleibt bisher unklar, was CARE im hier skizzierten breiteren Verständnis für den Datenaustausch, den Aufbau infrastruktureller Lösungen bzw. für die Nutzung bereits bestehender Infrastrukturen bedeuten kann, welche Themen adressiert und welche Richtlinienkompetenz die Datenzentren hier entwickeln wollen oder müssen.

Daher soll der Workshop zunächst einen vertiefenden Einblick und Verständnis für die CARE-Prinzipien schaffen und eine offene Diskussionsplattform bieten, um Ideen zu entwickeln, wie die CARE-Prinzipien in das Forschungsdatenmanagement, die FDM-Beratung und die Forschungsinfrastrukturentwicklung integriert und wie eine Rückwirkung auf die forschende Community ausgestaltet werden können. Anhand von spezifischen Praxisbeispielen werden konkrete Anwendungsszenarien vorgestellt, die für die systematisierende Diskussion als Impulse dienen.

Der Workshop sieht ein dreistufiges Format vor:

1. Der erste Teil macht die Vermittlung von Wissen zu den CARE-Prinzipien und der Systematisierung / Strukturierung damit verbundener Themen und Fragestellungen zum Gegenstand, die sich aus rechtlichen, ethischen und gesellschaftlichen Prinzipien ableiten lassen.

Einführend wird die Ethnologin Dr. Sabine Imeri die CARE-Prinzipien erläutern und grundlegende damit verbundene Wissenskonzepte und Anwendungskontexte illustrieren.

2. Daran anschließend stellen sich Projekte und Initiativen vor, die mit den von CARE adressierten Problematiken umgehen müssen, die aber in der tatsächlichen Umsetzung von CARE derzeit nur punktuell auf gesicherten Prinzipien jenseits gesetzlicher Regularien aufbauen können.

Diese Praxisbeispiele speisen sich nach den momentanen Planungen aus folgenden drei Use Cases:

- 2.1. Dr. Cord Pagenstecher vom Center für Digitale Systeme (CeDiS) der Freien Universität Berlin) berichtet aus dem Bereich der Oral History über ethische und datenschutzrechtliche Herausforderungen im Kontext des Aufbaus übergreifender audiovisueller Zeitzeugenarchive aus der NS-Zeit und im adäquaten Umgang mit historischen Unrechtssystemen. Der Projektverbund baut eine digitale Informationsinfrastruktur für wissenschaftliche Sammlungen von audiovisuell aufgezeichneten narrativen Interviews, v. a. für die zeitgeschichtliche Forschung, auf.

- 2.2. Prof. Dr. jur. Thomas Henne, LL.M. (Berkeley) eröffnet eine archivrechtliche Perspektive, die im Kontext der CARE-Prinzipien aufkommende Herausforderungen für Datenzentren und Produzent*innen von Forschungsdaten Orientierung und Verläss-

lichkeit bieten kann. Der im Archivrecht festgeschriebene öffentliche Auftrag von Archiven geht mit Pflichten und Möglichkeiten einher, Daten zu schützen, zu veröffentlichen und – von ihrer Entstehung an – zu erhalten.

2.3. Prof. Dr. Henning Schreiber und Dr. Katrin Pfeiffer vom Asien-Afrika-Institut der Universität Hamburg thematisieren insb. den Aspekt 'Collective Benefit' anhand aktueller Überlegungen zu Fragen von Digital Data Literacy aus afrikanistischer Perspektive. Grundlage ist ein Kooperationsprojekt, in dem die Sammlungen des National Centre Of Arts And Culture (NCAC) in Gambia digitalisiert und katalogisiert wurden.

3. Die anschließende Diskussion der Teilnehmenden mit den Expert:innen kann auf dieser Basis die Relevanz der CARE-Prinzipien in Theorie und Praxis für geisteswissenschaftliche Forschungsdaten und deren Verarbeitung in Datenzentren zum Gegenstand machen. Im Zentrum steht dabei die Identifizierung wesentlicher fachübergreifender Themen einer auf CARE-Prinzipien beruhenden, generalisierbaren Beratung zu ethischen Fragen im FDM. Darüber hinaus soll die Berücksichtigung wesentlich geisteswissenschaftlicher Ansätze und Problemlagen, etwa der Herausforderung zum Erhalt historischer Narrative und ihrer analytischen Kontextualisierbarkeit im Rahmen von historischen Studien sowie darauf aufbauender Methoden, in den Blick genommen werden.

Diese Systematisierung wird hinsichtlich bestehender Lösungsansätze anderer Fachdisziplinen, adaptierbarer Lösungspotentiale und perspektivischer Anknüpfungspunkte diskutiert. Damit können als Ergebnis des Workshops Ideen und Lösungsschritte dokumentiert werden, die Empfehlungen zum Umgang mit den CARE-Prinzipien in der Alltagspraxis der geisteswissenschaftlichen FDM-Beratung und der Datenzentren entwickeln. Für die langfristige Bearbeitung dieser Themenkomplexe werden Herausforderungen und Problemszenarien sowie noch offene Lösungsstrategien gebündelt. Die Art der Ergebnissicherung wird Bestandteil des Workshops sein.

Bibliographie

Carroll, SR, et al. (2020): "The CARE Principles for Indigenous Data Governance". *Data Science Journal*, 19: 43, S. 1-12. DOI: 10.5334/dsj-2020-043.

Frewer, Andreas / Bruns, Florian / May, Arnd T. (Hrsg.) (2019): *Ethikberatung in der Medizin*, Berlin, Heidelberg, DOI: 10.1007/978-3-642-25597-7.

GIDA, Global Indigenous Data Alliance (2019): *CARE Principles of Indigenous Data Governance*. [Online]. <<https://www.gida-global.org/care>> (Zugriff am 09.07.2021).

Kraft, Angelina (2017): *The FAIR Data Principles for Research Data* TIB-Blog. <<https://blogs.tib.eu/wp/tib/2017/09/12/the-fair-data-principles-for-research-data/>> (Zugriff am 19.11.2020).

Netzwerk Koloniale Kontexte (2021): "Website des Netzwerkes Koloniale Kontexte", in: *EVIFA Fachportal Ethnologie der Humboldt-Universität Berlin*. <<https://www.evifa.de/de/ueber-uns/fid-projekte/netzwerk-koloniale-kontexte>> (Zugriff am 12.07.2021).

Passamaquoddy People: Passamaquoddy People: At Home on the Ocean and Lakes, <https://passamaquoddypeople.com/> (Zugriff am 13.07.2021).

Rat für Informationsinfrastrukturen (Hrsg.) (2021): *Datentreuhänder: Potenziale, Erwartungen, Umsetzung*. Workshop-Bericht der AG Datentreuhänderschaft des RfII am 25. September 2020, Göttingen 2021, urn:nbn:de:101:1-2020052646.

Schwerhoff, Gerd (1992): "Devianz in der alteuropäischen Gesellschaft. Umrisse einer historischen Kriminalitätsgeschichte", in: *Zeitschrift für Historische Forschung* 19, Nr. 4, S. 385-414.

UNDRIP (2007): *United Nations Declaration on the Rights of Indigenous Peoples*, <https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-indigenous-peoples.html> (Zugriff am 13.07.2021).

von Unger, Hella / Dilger, Hansjörg / Schönhuth, Michael (2016): "Ethics Reviews in the Social and Cultural Sciences? A sociological and Anthropological Contribute to the Debate", in: *Forum Qualitative Sozialforschung*, 17 (3), Art. 20, <http://nbn-resolving.de/urn:nbn:de:0114-fqs1603203>.

Wilkinson, Mark D. u.a. (2016): "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3, Article 160018, DOI: 10.1038/sdata.2016.18.

FAIRRes Datenmanagement mit dem DARIAH-DE Repository

Jander, Melina

jander@sub.uni-goettingen.de
SUB Göttingen, Germany

Weimer, Lukas

weimer@sub.uni-goettingen.de
SUB Göttingen, Germany

Einleitung

Forschende, Lehrende und Studierende produzieren bei ihrer Arbeit kontinuierlich Daten. Im Zuge des kulturellen Wandels bleibt die Frage danach, wie mit diesen Daten den FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable; Wilkinson et al. 2016) entsprechend umgegangen werden kann, stets aktuell und nach wie vor ungelöst. Im direkten Austausch mit der Community hat CLARIAH-DE (durch qualitative Interviews mit verschiedenen Zielgruppen)¹ unterschiedliche Bedarfe ermittelt, u.a. die Möglichkeit, Forschungsdaten auffindbar, zugänglich und nachnutzbar zu speichern. Das DARIAH-DE Repository wird diesen Anforderungen gerecht, indem sein Publikationswerkzeug, der DARIAH-DE Publikator, mit grafischer Nutzeroberfläche und detaillierten (mehrsprachigen) Beschreibungs- und Hilfetexten auch jene Nutzende abholt, die keine Vorerfahrung im Forschungsdatenmanagement mitbringen. Um diese Qualitätsmerkmale zertifiziert zu belegen, befindet sich das DARIAH-DE Repository derzeit im Beantragungsprozess des Core-TrustSeals (<https://www.coretrustseal.org/>). Das Repository ist Teil des von der SUB Göttingen für das NFDI-Konsortium Text + (<https://www.text-plus.org/>) angebotenen Dienstes der DARIAH-DE Data Federation Architecture (DFA), die insgesamt fünf Dienste in sich vereint. Diese adressieren eine nachhaltige Publikation, aussagekräftige Beschreibung, Mapping und Auffindbarkeit der Forschungsdaten. Dabei stellt die DFA schrittweise jene

Dienste bereit, die wichtige Aspekte des Research Data Lifecycles (Puhl et al. 2015) abdecken.

Im Workshop soll das DARIAH-DE Repository mit seinen Funktionalitäten sowohl im übergeordneten Kontext der FAIR-Prinzipien als auch im technischen Kontext der DFA vorgestellt werden. Mit den Teilnehmenden werden die einzelnen Schritte des Publikationsprozesses durchlaufen, um mit diesem vertraut zu werden und mögliche Hemmschwellen bei der Datenpublikation zu senken. Gleichzeitig bietet der Workshop Raum für konstruktives Feedback. Perspektivisch kann dies dazu beitragen, den Wandel innerhalb der akademischen Publikationskultur hin zu einem grundlegenden Bewusstsein von Open Access voranzutreiben und somit die Kartierung der Forschungsdatenlandschaft positiv zu beeinflussen.

Der Workshop richtet sich an Geisteswissenschaftler*innen in allen Stufen der akademischen Laufbahn und unabhängig von ihrer zugehörigen Institution, ihrer Arbeit oder ihren Forschungsinteressen, da der Bedarf an nachhaltiger Datenpublikation in all diesen Bereichen gleichermaßen hoch ist. Das Repository wird sowohl von Einzelforscher*innen als auch kollaborativ in Forschungsprojekten genutzt. Es beinhaltet aktuell 267 Kollektionen mit mehr als 1.700 Dokumenten.

Repositorien in den Geistes- und Kulturwissenschaften

Das Angebot an Forschungsdatenrepositorien ist vielfältig. Manche Repositorien beschränken sich auf einzelne Fachdisziplinen (z.B. AMAD für Mittelalterstudien), andere auf bestimmte Formate und Communities (bspw. das Bildarchiv Foto Marburg oder das Deutsche Textarchiv) oder Publikationstypen (Forschungs(roh)daten vs. Publikationen). Ferner bieten manche nur Archivfunktionen ohne Veröffentlichung an und viele können ihre Dienste nur für eine kurze Zeit aufrechterhalten. Andere wiederum sind kostenpflichtig (z.B. RADAR).

Das DARIAH-DE Repository kombiniert die oben genannten Elemente und geht somit über reine Archivfunktionen hinaus. Dennoch kann es als generisches Repository nicht alle Bedarfe einzelner Fachcommunities adressieren (z.B. Helling et al. 2020). Seit 2017 ist es durch die DFA (vgl. Abb. 1) und seine Anbindung an DARIAH- und CLARIAH-DE Teil größerer Infrastrukturen, die Nachhaltigkeit garantieren und an geisteswissenschaftliche NFDI-Konsortien angeschlossen sind (Brünger-Weilandt et al. 2020). Nach Ablauf des Förderzeitraums von DARIAH-DE wurde es im Rahmen der DARIAH-DE Betriebskooperationsvereinbarung weitergeführt und war außerdem Teil des Angebots von CLARIAH-DE, das nun als Angebot der SUB Göttingen in Text + übergegangen ist. Via CLARIAH-DE existiert ein Helpdesk², über den Fragen gestellt werden können. Technisch wird das DARIAH-DE Repository von der GWDG und SUB Göttingen betrieben. Um die Fachwissenschaft gezielt anzusprechen und einen thematischen Rahmen für die Forschungsdaten zu bieten, ist es zwar geistes- und kulturwissenschaftlich ausgerichtet, hierbei aber nicht an Einzelwissenschaften gebunden. Es ist ferner durchsuchbar, nicht rein institutionell und bietet neben zitierfähigen Links bspw. auch die Vergabe von Persistenten Identifikatoren (DataCite DOI und ePIC Handle). Darüber hinaus sind die Verwendung des Repositoriums sowie der gesamten DFA, die Speicherung von Daten und alle zusätzlichen Services nicht mit Folgekosten für die Nutzenden verbunden.³

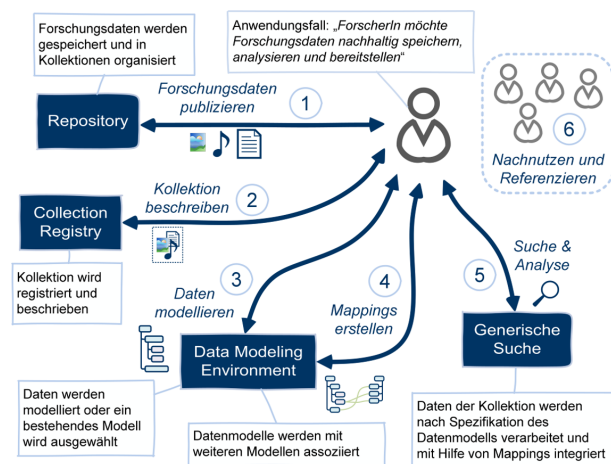


Abb. 1: Schematische Darstellung der DARIAH-DE DFA.

Das DARIAH-DE Repository in der Forschungsdaten-Föderationsarchitektur

Im Rahmen digitalen Forschens gelten die FAIR-Prinzipien zunehmend als Leitlinien. Für Forschende ist es daher wichtig, dass schon der Publikationsprozess auf eine FAIRe Publikation hin ausgerichtet ist (Ivanović et al. 2019). Diesem Anspruch wird das DARIAH-DE Repository auf nutzerfreundliche Weise gerecht: Mit einem DARIAH- oder Föderationsaccount greifen Nutzende auf den Publikator in einem grafischen Interface zu und können mit wenigen Klicks mit dem Einspielen der Daten und der Auszeichnung der Metadaten beginnen. Das Dublin Core Metadatenchema liegt hier als Standard zugrunde und ermöglicht es auch jenen Forschenden, die nur wenige Metadaten eingeben möchten, ohne großen Aufwand ihre Daten zu beschreiben. Das Design von Eingabemaske und Fileupload wird dabei als eines wahrgenommen, mit dem intuitiv gearbeitet werden kann (Cremer 2018).

Die im Publikator erstellten Kollektionen können in einem nächsten Schritt in der Collection Registry eingetragen und mit weiteren deskriptiven Metadaten ausgezeichnet werden. Mit der Generischen Suche beinhaltet die DFA außerdem ein Front-End für die in der Collection Registry verfügbar gemachten Daten und deren Metadaten. Durch den modularen Aufbau der DFA können die in ihr vereinten Werkzeuge und Dienste – unter Nutzung u.a. der DARIAH-DE Authentifizierungs- und Autorisierungsinfrastruktur (AAI) sowie der DARIAH-DE Storage API zur Speicherung von Forschungsdaten (Schmunk / Funk 2018) – somit sowohl kombiniert in einem Workflow als auch individuell genutzt werden. Das technische Workaround für einen Datenimport ins Repository lässt sich dabei folgendermaßen beschreiben (vgl. Abb. 2):

1. Publikator: Die Nutzenden laden ihre zur Publikation vorbereiteten Daten über die Eingabemaske ins Repository und versehen sie mit Metadaten entsprechend des Dublin Core Metadatenschemas. Sowohl die Objekte als auch ihre Metadaten werden als Kollektion per API an den DARIAH-publish Service übermittelt.
2. DARIAH-publish Service: Die Metadaten werden validiert, Referenzen auf Objekte innerhalb der einzuspielenden Kol-

dahingehend zu optimieren (Friedrichs / Jander / Reißler-Pipka, in Veröffentlichung).

2. "Support," CLARIAH-DE, Zugriff am 18. November 2021, <https://www.clariah.de/support>.

3. Für eine ausführliche Dokumentation des DARIAH-DE Repositoriums siehe DARIAH-DE, *DARIAH-DE Repository Documentation: Release 2020-06-25* (2020), https://repository.de.dariah.eu/doc/services/dhrep_doc.pdf.

Bibliographie

Brünger-Weilandt, Sabine / Bruhn, Kai-Christian / Busch, Alexandra W. / Hinrichs, Erhard / Maier, Gerald / Paulmann, Johannes / Rapp, Andrea / von Rummel, Philipp / Schlottheuber, Eva / Schmidt, Dörte / Schrade, Torsten / Simon, Holger / Stein, Regine / Teich, Elke (2020): "Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies." Zugriff am 08. Juli 2021. <https://zenodo.org/record/4045000#.YORQuOgZPY>

CLARIAH-DE (2021): "Support". Zugriff am 18. November 2021. URL: <https://www.clariah.de/support>.

CoreTrustSeal (2021): Zugriff am 12. Juli 2021. URL: <https://www.coretrustseal.org/>.

DARIAH-DE (2020): "DARIAH-DE Repository Documentation": Release 2020-06-25 (2020). Zugriff am 08. Juli 2021. URL: https://repository.de.dariah.eu/doc/services/dhrep_doc.pdf.

Cremer, Fabian (2018): "DARIAH-DE Repository: Notizen zum Nutzen jenseits der Nutzung." DHd-Blog. Zugriff am 08. Juli 2021. URL: <https://dhd-blog.org/?p=10368>.

Ivanović, Dragan / Schmidt, Birgit / Grim, Rob / Dunning, Alastair (2019): "FAIRness of Repositories & Their Data: A Report from LIBER's Research Data Management Working Group." Zugriff am 05. Juli 2021. <http://doi.org/10.5281/zenodo.3251593>.

Friedrichs, Sonja / Jander, Melina / Reißler-Pipka, Nanette (in Veröffentlichung): "User Studies zur digitalen Forschungsinfrastruktur von CLARIAH-DE: Konzept, Umsetzung, Erkenntnisse", in: *DARIAH-DE Working Papers*.

Helling, Patrick / Jung, Kerstin / Pielström, Steffen (2020): "Standards and harmonized components of technical/structural infrastructures for long-term archiving and publishing of complex and heterogeneous data packages". Text+ User Story. Zugriff am 08. November 2021. URL: <https://www.text-plus.org/en/research-data/user-story-337/>.

Puhl, Johanna / Andorfer, Peter / Höckendorff, Mareike / Schmunk, Stefan / Stiller, Juliane / Thoden, Klaus (2015): "Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften", in: *DARIAH-DE Working Papers* 11.

Schmunk, Stefan / Funk, Stefan E. (2016): "Das DARIAH-DE- und das TextGrid-Repositorium: Geistes- und kulturwissenschaftliche Forschungsdaten persistent und referenzierbar langzeitspeichern." *Bibliothek – Forschung und Praxis* 40, no. 2: 213–221.

Wilkinson, Marc D. / Dumontier, Michel / Aalbersberg, I. Jan et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship", in: *Scientific Data* 3.

Flexibles Arbeiten mit OCR4all

Massenvolltextdigitalisierung von Drucken mithilfe von OCR-D und hochqualitative Transkription von Handschriften

Langhanki, Florian

florian.langhanki@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Wehner, Maximilian

maximilian.wehner@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Baierer, Konstantin

konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preussischer Kulturbesitz

Hinrichsen, Lena

hinrichsen@hab.de
Herzog August Bibliothek Wolfenbüttel

Reul, Christian

christian.reul@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Einleitung

Die automatisierte Texterkennung von historischen Drucken und Handschriften gilt aus geisteswissenschaftlicher wie aus informatischer Perspektive in ganz unterschiedlichen Forschungs- und Anwendungskontexten auch weiterhin als anspruchsvolle und problembehaftete Aufgabe. Während die OCR (Optical Character Recognition) moderner Texte mit ihren zeilenbasierten OCR-Ansätzen (Breuel et al. 2013) weithin als informatisch quasi gelöstes Problem angesehen wird, stellen v. a. die höchst komplexen Layoutstrukturen vormoderner Werke (speziell der vor 1700) und ihr teils schlechter Druck- bzw. Erhaltungszustand immer noch ein großes Problem bei der Herstellung maschinenles- und -verarbeitbarer Texte dar. Verglichen mit der Vielfalt und Varianz der in Drucken verwendeten Typen und Schriftarten, gestaltet sich die Erkennung von Handschriften durch die vielfältigen Ausprägungen einzelner Schriftarten in Kombination mit unterschiedlichen Schreiberhänden noch einmal komplizierter. Selbst der kommerzielle State of the Art der Texterkennungssoftware wie bspw. ABBYY Finereader¹ wird in der Produktion wissenschaftlich nutzbarer Daten hier vor erhebliche Probleme gestellt. Die bereits bekannten Schwierigkeiten einer OCR auf historischen Daten müssen demnach um jene einer HTR (Handwritten Text

Recognition) mittelalterlicher und frühneuzeitlicher Werke erweitert werden.

Besonders die neueren Forschungsfelder innerhalb der Geisteswissenschaften und Digital Humanities (Text Mining, Sentiment Analysis etc.) haben diese Schwierigkeiten bei gleichzeitigem Bedarf großer Textmengen zur Anwendung quantitativer Analyseverfahren erkannt. Hier stellt sich zunehmend die Frage nach Möglichkeiten einer Texterkennung historischer Drucke und Handschriften, die sowohl hohen Qualitätsansprüchen als auch einem ebensolchen Automatisierungsgrad genügen.

Es ist unstrittig, dass entsprechende Werkzeuge frei verfügbar sein, kohärente OCR- bzw. HTR-Workflows zur Verfügung stellen müssen und außerdem einfach und selbstständig durch nicht-informatische, geisteswissenschaftliche Nutzer:innen bspw. über eine grafische Benutzeroberfläche nutzbar sein sollten. Hinzu kommen jene spezifischen Anforderungen, die mit der Massenverarbeitung von Texten einhergehen, sowie der Wunsch nach größtmöglicher Flexibilität und nach Vielfalt von Werkzeugen. Den besonderen Anforderungen einer massenhaften Textdigitalisierung wendet sich besonders das DFG-geförderte Projekt OCR-D (Engl et al. 2020) mit dem Ziel zu, die Werke in den Verzeichnissen der deutschsprachigen Drucke (VD 16–18) durch vollautomatische Texterkennung als Forschungsdaten zur Verfügung zu stellen. Während in OCR-D also der Fokus auf Massenverarbeitung, Skalierbarkeit und Flexibilität sowie vielfältigen Anwendungsmöglichkeiten liegt, vereint die an der Universität Würzburg entwickelte Software OCR4all² (Reul et al. 2019) die erstgenannten Notwendigkeiten einer einfachen Nutzbarkeit entsprechender Technologien mithilfe einer grafischen Benutzeroberfläche und richtet sich dabei dezidiert an Geisteswissenschaftler:innen.

Mit dem im Juli 2021, im Rahmen der dritten Projektphase von OCR-D³, gestarteten Würzburger Forschungsprojekt OCR4all-libraries⁴ rückt mit der geplanten Integration der OCR-D-Lösungen in die dort entwickelte Software nun noch einmal verstärkt eine notwendige Vereinfachung und Individualisierung komplexer und projektspezifisch flexibel anwendbarer OCR- und HTR-Workflows in den Fokus. Die Anwendung der Software im Spannungsfeld einer Massenvolltextdigitalisierung wie jener der VD16–18⁵ und einer hochqualitativen Erfassung mittelalterlicher Handschriften erfährt hier einen neuen wie nachhaltigen Rückenwind.

OCR4all

Die im Workshop verwendete Software orientiert sich in seinem Aufbau an den Hauptkomponenten eines OCR-Workflows (s. u.), gliedert diesen jedoch noch einmal in unterschiedliche Teilmodule. Dieser modulare Aufbau erlaubt eine Einbindung und Verwendung bereits bestehender Softwarelösungen, die gemäß ihren Stärken zu einem kohärenten OCR-Workflow kombiniert werden. Im Allgemeinen umfasst der typische Ablauf einer OCR bzw. HTR die **Vorverarbeitung** (Preprocessing), die **Regionen- und Zeilensegmentierung** (Region-, Line-Segmentation), die **Texterkennung** (Recognition) und die **Nachkorrektur** (Post Correction) (s. Abb. 1).



Abb. 1: Hauptkomponenten eines typischen OCR-Workflows. Von links nach rechts: Originalbild, Vorverarbeitung, Segmentierung, Texterkennung, Nachkorrektur.

Im Preprocessing werden die Einzelbilder gerade gestellt und binarisiert oder in Graustufen umgewandelt (s. Abb. 1). Dabei werden alle gängigen Eingabeformate für Bilddateien unterstützt. Dem schließt sich die Layouttypisierung mithilfe des Segmentierungstools LAREX⁶ (s. Abb. 2) an. Hier können werkspezifische Parameter zur Text- und Bildtypisierung festgelegt sowie zu erkennende Layoutregionen (Haupttext, Überschriften, Marginalien, Seitenzahlen, Anstreichungen, Randnotizen etc.) definiert werden. Je nach Komplexität des vorliegenden Seitenlayouts ist nach einer automatischen Layouterkennung ein Eingriff in das vorliegende Ergebnis mittels unterschiedlicher Korrekturwerkzeuge möglich. Weiterhin kann in LAREX die Lesereihenfolge der Layoutbestandteile markiert werden, um den Lesefluss des Originals später vorlagengetreu nachbilden zu können. Vor allem für die Verwendung des maschinenverarbeitbaren Textes in digitalen Editionen sind viele der beschriebenen Funktionen unverzichtbar.

Der Layouttypisierung folgt die Zeilensegmentierung. In dieser werden die Text beinhaltenden Layoutbestandteile in einzelne Zeilenbilder zerteilt (OCRopus⁷), um damit die eigentliche OCR vorzubereiten.

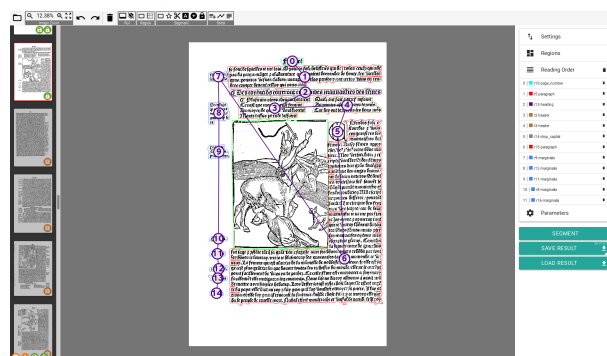


Abb. 2: Im Teilmodul der Segmentierung erfolgen die Typisierung der Layoutelemente sowie die Festlegung der Lesereihenfolge entweder von Grund auf oder in Form der Korrektur eines automatisch generierten Ergebnisses.

Anschließend wird bei der Recognition aus den vorliegenden Einzelzeilenbildern (mittels der OCR-Engine Calamari⁸) maschinenverarbeitbarer Text generiert. Dazu können in OCR4all bereits standardmäßig integrierte gemischte Modelle für Schriftarten unterschiedlicher Epochen genutzt werden. Als 'gemischt' werden Modelle bezeichnet, deren Trainingsgrundlage aus einer Vielzahl verschiedener Drucktypen und Schriftarten besteht. Nach der Recognition können die entstandenen Texte in einem Editor komfortabel korrigiert werden (s. Abb. 3).



Abb. 3: Im Editor kann generierter Text mithilfe einer virtuellen Tastatur (rechts) zeichengetreu korrigiert werden.

Für die Berechnung der Fehlerrate einer Zeichenerkennung kann im Evaluationsmodul der ursprünglich erkannte Text mit der durch die Nutzer:innen vorgenommenen Korrektur verglichen werden.

Darüber hinaus bietet OCR4all die Möglichkeit, unter Verwendung vorgenommener Textkorrekturen, selbstständig werkspezifische Modelle zu trainieren, anzuwenden und iterativ zu verbessern. Besonders für Werke mit großer Typenvielfalt und -varianz, auf denen bereits bestehende gemischte Modelle keine hinreichende Erkennungsergebnisse erzielen, werden auf diese Weise dennoch sehr hohe Erkennungsraten erreicht.

Im abschließenden Modul zur Nachkorrektur können die während des Workflows generierten Texte editionsreif korrigiert und anschließend als Plain Text und im Kontext weiterer Strukturdaten als PAGE-XML⁹ ausgegeben werden. Letzteres Format beinhaltet neben dem erkannten und ggf. nachkorrigierten Text so auch die Koordinaten aller ausgezeichneten Layoutelemente der Scanseite sowie deren semantische Funktion innerhalb des originalen Seitenlayouts.

Derzeit ist der Workflow auf die hier erläuterten Methoden beschränkt. Im Verlauf des OCR4all-libraries Projekts werden bis zum Workshop jedoch auch die im Rahmen des OCR-D-Projekts erarbeiteten Lösungen verfügbar gemacht werden, wodurch die Nutzer:innen den Workflow eigenständig um Einiges flexibler gestalten und präziser auf den eigenen Anwendungsfall abstimmen können.

In Abhängigkeit des Ausgangsmaterials variiert der zum Erreichen einer sehr hohen Genauigkeit benötigte Arbeitsaufwand zwischen wenigen Minuten bei Werken mit einfachen Layoutstrukturen und einigen Stunden bei sehr komplexen Werken, für die spezifische Erkennungsmodelle erst noch trainiert werden müssen (Reul et al. 2019).

Workshopkonzeption

Der ganztägige Workshop soll einem informatisch wie technisch nicht speziell geschulten Nutzer:innenkreis einen einfachen und verständlichen Einstieg in das Themen- und Problemfeld der OCR und HTR historischer Materialien bieten. Er wird dazu befähigen, mithilfe der vorgestellten Software eigenständig und innerhalb kurzer Zeit qualitativ hochwertige Texte aus ganz unterschiedlichen Ausgangsdaten zu generieren. Die Workshopkonzeption erfolgt deshalb besonders praxisbezogen. Dies bedeutet einen angeleiteten und stets individuell anpassbaren Durchgang des oben vorgestellten OCR- bzw. HTR-Workflows anhand verschiedener, nach Layoutkomplexität, Typographie und Schriftart, Erhaltungszustand und Entstehungszeitraum gruppierter Drucke und Handschriften. Dabei sollen anwendungsbezogen u. a. die folgenden Grundfragen der OCR und HTR beantwortet werden:

- Auf welchen Daten ist OCR4all anwendbar? Was ist OCR-D und welchen Mehrwert bringt die Integration von OCR-D-Lösungen?
- Wie verändert sich entsprechend des Ausgangsmaterials die Anwendung des in OCR4all integrierten OCR- bzw. HTR-Workflows und der in ihm enthaltenen Submodule?
- Mit welchem (manuellen) Aufwand ist in unterschiedlichen Bearbeitungsphasen des Materials zu rechnen?
- Wie stark lässt sich der Workflow in Abhängigkeit des vorliegenden Materials automatisieren?
- Wie und nach welchen Maßgaben können (im Rahmen eines iterativen Ansatzes) projekt- und werkspezifische Texterkennungsmodelle trainiert werden? Welche Erkennungsgenauigkeiten sind zu erwarten?
- Welcher Aufwand ist mit Blick auf die spätere Verwendung der produzierten Texte überhaupt sinnvoll?

Da sich neben den Spezifika des Ausgangsmaterials auch eine grundlegende technische Expertise der Anwender:innen im Bereich der OCR und HTR als Grundbedingung der Produktion hochwertiger maschinenlesbarer Texte herausgestellt hat, strebt der Workshop neben einer praktischen Handlungsanleitung auch die Vermittlung der wichtigsten Funktionskonzepte der in OCR4all integrierten Submodule an.

Darüber hinaus umfasst die Veranstaltung auch Fragen der Einrichtung und Installation der Software, um den Teilnehmer:innen eine stabile und nachhaltige Anwendung von OCR4all über den Workshopkontext hinaus zu ermöglichen. Um einen reibungslosen Ablauf des Workshops selbst zu garantieren, wird durch die Antragsteller:innen eine Serverversion der Software zur Verfügung gestellt. Die max. 25 Teilnehmer:innen benötigen für die Teilnahme deshalb lediglich einen internetfähigen Laptop. Die Verwendung einer Maus wird empfohlen. Digitalisate werden zur Verfügung gestellt, gerne darf aber auch eigenes Material mitgebracht und im Workshop bearbeitet werden.

Forschungsinteressen der Beitragenden

Florian Langhanki ist Wissenschaftlicher Mitarbeiter am 'Zentrum für Philologie und Digitalität' der Universität Würzburg. Seine Forschungsinteressen sind Übersetzungsliteratur und Zweisprachigkeit in Mittelalter und Früher Neuzeit sowie die OCR und HTR frühneuzeitlicher Werke und Sammelhandschriften.

Maximilian Wehner ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für ältere deutsche Philologie der Universität Würzburg. Seine Forschungsinteressen sind die wissensvermittelnde Literatur der Frühen Neuzeit, die OCR bzw. HTR mittelalterlicher und frühneuzeitlicher Drucke und Handschriften sowie deren Nutzung in universitärer und schulischer Lehre.

Konstantin Baierer ist Wissenschaftlicher Mitarbeiter an der Staatsbibliothek zu Berlin und betreut dort seit 2018 das OCR-D-Projekt.

Lena Hinrichsen ist Wissenschaftliche Mitarbeiterin an der Herzog August Bibliothek Wolfenbüttel und Projektkoordinatorin von OCR-D. Ihre Forschungsinteressen sind OCR und Objekterkennung sowie Bild-Text-Verhältnisse.

Dr. Christian Reul leitet die Digitalisierungseinheit des 'Zentrum für Philologie und Digitalität' der Universität Würzburg. Seine Forschungsschwerpunkte sind die OCR/HTR auf histori-

schem Material sowie die Neu- und Weiterentwicklung der entsprechenden Software.

Fußnoten

1. <https://www.abbyy.com/de-de/finereader/>
2. <http://ocr4all.de/>
3. <https://ocr-d.de/de/phase3/>
4. <https://www.uni-wuerzburg.de/zpd/news/single/news/ocr4all-libraries-genehmigt/>
5. <https://ocr-d.de/de/about/>
6. <https://github.com/OCR4all/LAREX>
7. <https://github.com/tmbdev/ocropy>
8. <https://github.com/Calamari-OCR/calamari>
9. <https://www.primaresearch.org/tools/PAGELibraries>

Bibliographie

Breuel, Thomas M. / Ul-Hasan, Adnan / Al-Azawi, Mayce Ali / Shafait, Faisal (2013): High-Performance OCR for Printed English and Fraktur Using LSTM Networks, in: *12th International Conference on Document Analysis and Recognition*: 683-687.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings, in: *Applied Sciences* 2019. (9) 22. URL: <https://www.mdpi.com/2076-3417/9/22/4853>

Engl, Elisabeth / Boenig, Matthias / Baierer, Konstantin / Neudecker, Clemens / Hartmann, Volker (2020): Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke, in: *Zeitschrift für Historische Forschung* 47 (2), 2020, S. 223–250. URL: <https://elibrary.duncker-humblot.com/journals/id/28/vol/47/iss/5737/art/58179>

GitMA oder CATMA für Fortgeschrittene Projektdaten via Git abrufen und mittels Python-Bibliothek weiterverarbeiten

Schumacher, Mareike

schumacher@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Vauth, Michael

vauth@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Gerstorfer, Dominik

gerstorfer@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Meister, Malte

meister@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Dieser CATMA-6-Workshop richtet sich an fortgeschrittene CATMA User*innen mit Vorkenntnissen in digitaler Annotation, die im Rahmen der eigenen Arbeit oder von Forschungsprojekten mit größeren Mengen von Annotationsdaten operieren (wollen). Im Zentrum steht die Weiterverarbeitung und Analyse von Annotationsdaten. Wie greife ich über Git auf meine CATMA-Annotationsdaten zu? Wie erstelle ich individuelle, interaktive Visualisierungen meiner Annotationsdaten? Wie berechne ich die Übereinstimmung zwischen mehreren Annotator*innen? Diese und ähnliche Fragen werden während des Workshops beantwortet.

CATMA (Gius et al. 2021) ist eine webbasierte, kollaborative Textannotations- und Analyse-Plattform, die seit 2008 an der Universität Hamburg und im Rahmen des DFG-geförderten Projektes forTEXT seit 2020 an der Technischen Universität Darmstadt entwickelt wird.¹ Hauptzielgruppe sind traditionell-analog arbeitende Geisteswissenschaftler*innen, die über eine intuitiv bedienbare GUI Texte annotieren und analysieren können. Mit dem Release von CATMA 6 im Jahr 2019 wurde für die Plattform ein auf Git basierendes Backend eingeführt. Für zahlreiche Projekte, die bereits auf sehr fortgeschrittenem Niveau CATMA nutzen, und Interessierte aus der Digital-Humanities-Community mit Erfahrung in der Nutzung von Git und Grundkenntnissen in Python eröffnet sich dadurch eine Reihe neuer Funktionen, die es in bisherigen CATMA-Versionen nicht gab. Einige dieser Funktionen werden im Laufe dieses Ganztagesworkshops vorgestellt und vermittelt.

Der Workshop bietet:

- kurze Einführung in die Nutzung von CATMA über das graphische Userinterface
- Kennenlernen der Datenstrukturen des Backends
- Zugriff auf das Backend mit Git
- Weiterverarbeitung der Daten mit Hilfe eines zur Verfügung gestellten Python-Packages

Annotation in CATMA 6 – projektorientiert, gemeinsam, vielfältig

Eine der wichtigsten Neuerungen von CATMA 6 gegenüber früheren Versionen ist die Umstellung auf eine projektzentrierte Nutzungsarchitektur. Am Beginn der Arbeit mit CATMA steht das Anlegen eines Projektes mit beliebig vielen Dokumenten, die analysiert werden sollen, und beliebig vielen Team-Mitgliedern, die daran arbeiten wollen. Zur Annotation können eigene Taxonomien entworfen oder auf der Plattform fortext.net bereitgestellte Ressourcen genutzt werden. Die Annotationskategorien können frei gestaltet werden und jede Passage im Text kann frei damit annotiert werden. Einzelne und Mehrfachannotationen, einander überlagernde oder überlappende Annotationen oder sogar widersprüchliche Annotationen – in CATMA ist durch die Speicherung der Daten als Standoff-Markup vieles möglich. Eine weitere

Neuerung im Funktionsumfang ist die Möglichkeit, Textstellen und Annotationen zu kommentieren. Offene Fragen, nicht zu Ende gedachte Interpretationsansätze oder auch der Austausch mit den anderen Team-Mitgliedern können über die Kommentarfunktion in den Annotationsprozess integriert werden. Sowohl Annotationen als auch Kommentare können über die Analyse-Funktionen von CATMA durchsucht, in tabellarische Form gebracht oder visualisiert werden. Der Umfang dessen, was über die CATMA-GUI umgesetzt werden kann, ist also recht groß. Und doch macht die Einführung des auf Git basierenden Backends das Tool für die Digital-Humanities-Community erst richtig interessant. Der undogmatische Zugang, der bisher nur zu Annotationen und Annotationstaxonomien ermöglicht wurde, erstreckt sich nun bis zu den Annotationsdaten und der Weiterverarbeitung derselben (siehe Abbildung 1). Dieser neue Teil des CATMA-Workflows wird in diesem Workshop vermittelt werden.

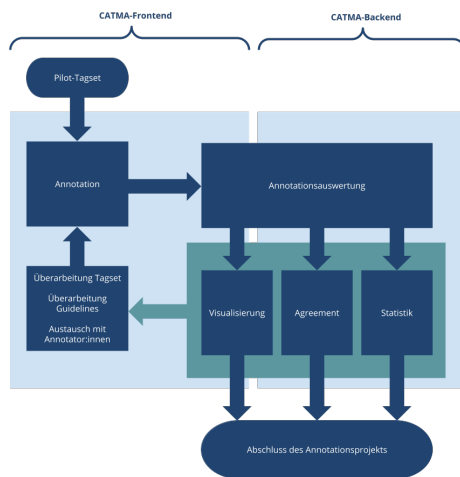


Abb. 1: Im Workshop vermittelter Workflow zur Annotationsauswertung und -überarbeitung mit dem CATMA-Backend

Standards und Best Practices nicht aus den Augen verlieren mit GitMA

Niedrigschwelligkeit und Nähe zu traditionell-analogen Methoden der Geisteswissenschaften sind nach wie vor wichtige Grundsätze, die in CATMA implementiert sind. Doch mit zunehmender Verbreitung des Tools in den digitalen Geisteswissenschaften sind neben der Möglichkeit zu hermeneutisch-vielfältiger Textanalyse auch die Einhaltung von Best Practices und Standards, die innerhalb der Digital-Humanities-Community entwickelt wurden, von Bedeutung. Eine Verschmelzung von CATMA und Git zu "GitMA" ermöglicht beides. Dabei bleibt der Annotationsprozess selbst völlig frei gestaltbar. Die resultierenden Daten aber können zum Beispiel nach der Übereinstimmung der Annotierenden untereinander ausgewertet werden. Es ist möglich eine der Annotationen als 'Silver Annotation' festzulegen und die anderen

daran zu messen. Das festgestellte Disagreement kann zur Grundlage eines Disagreement-Tagsets werden, das über das Backend auch wieder ins Frontend der CATMA-GUI zurückgespielt werden kann (siehe Abb. 1). Dasselbe gilt für die nicht übereinstimmend annotierten Passagen, welche wiederum selbst durch Annotationen dargestellt/hervorgehoben werden können. So ergibt sich ein harmonischer Workflow vom Frontend zum Backend und zurück, der in Zukunft auch die Erstellung von Goldannotationen unterstützen wird.

Die GitMA-Funktionalitäten werden im Rahmen dieses Workshops erstmals einem Fachpublikum vorgestellt. Neben der Vermittlung von Nutzungskompetenzen möchten wir darum auch eine kritische Diskussion anregen. Feedback zu Idee und Umsetzung der CATMA-Backend-Nutzung sind uns überaus willkommen!

Format und Ablauf des Workshops

Der Workshop wird als ganztägiges hands-on Tutorial angeboten, das an einem oder an zwei aufeinander folgenden (halben) Tagen stattfinden kann.

Ablauf:

Teil 1

1. CATMA Backend (45 Minuten)
2. kurze Einführung in das CATMA-Frontend
3. Struktur: Tagsets, Documents, Annotation Collections
4. Annotationsrepräsentation: JSON-Files
5. Zugriff auf Annotationsdaten über Git (45 Minuten)
6. wie clone ich ein CATMA Project?
7. wie update ich ein CATMA Project, um neue Annotationen zu laden?

Pause

1. Zugriff auf ein CATMA Project mit Python (45 Minuten)
2. Installation des Packages
3. Laden eines Projects
4. Zugriff auf Annotation Collections, Dokumente und Tagsets

Teil 2

1. Annotationsauswertungen (90 Minuten)
2. Visualisierungen zum Annotationsfortschritt und zur Exploration von Annotationen (Plotly)
3. IAA Auswertung von zwei Annotation Collections des gleichen Dokuments (15 Minuten)
4. weiterführende Auswertungen mit Pandas

Pause

1. Unterstützung der Goldannotation (75 Minuten)
2. Festlegung der Silver Annotations
3. Umgang mit Annotationsspannen
4. Automatische Erstellung eines Disagreement Tagsets
5. Darstellung von Disagreement als Annotationen
6. Manuelle Überarbeitung von automatischen Goldannotationen
7. Diskussion und Feedback (60 Minuten)

Zielgruppe:

Nutzer*innen, die Annotationen mit CATMA in Forschungsprojekten oder Lehrsituationen managen, sowie alle, die einen

schnellen Workflow zwischen Annotation bzw. Annotationsbearbeitung und Annotationsauswertung benötigen.

Zahl der möglichen Teilnehmer*innen:

30

Technische Voraussetzungen:

Die benötigten Vorinstallationen von Git, Anaconda und Plotly können durch die Bereitstellung eines Docker-Image vermieden werden. Die Teilnehmer*innen sollten die Installation von Docker selbst auf einem eigenen Laptop (Touch Devices werden nicht unterstützt), den sie zum Workshop mitbringen, möglichst schon erledigt haben. Für die Durchführung des Workshops benötigen wir außerdem einen Beamer.

Zur Vorbereitung sollten Teilnehmer*innen außerdem schon einen CATMA-Account erstellt (unter <https://app.catma.de/catma/>) und sich mit der CATMA-Nutzung bekannt gemacht haben (z.B. mithilfe von der forTEXT-Lerneinheit zu CATMA 6: *Manuelle Annotation mit CATMA*). Wenn eigene CATMA-Annotationsdaten vorhanden sind, können diese während des Workshops analysiert werden. Für Teilnehmende, die nicht an eigenen Daten arbeiten möchten, stellen wir ein Demo-Projekt zusammen, dem man während des Workshops beitreten kann.

Benötigte Vorkenntnisse:

Die Teilnehmer*innen sollten über grundlegende Kenntnisse der Kommandozeile, Git und Python sowie Jupyter verfügen.

Beitragende

Michael Vauth, M.Ed.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Michael Vauth promoviert über "Zur Annotation intradiegetischen Erzählens. Binnenerzählungen im literarischen Werk Heinrich von Kleists" an der Technischen Universität Darmstadt. Er ist wissenschaftlicher Mitarbeiter im Forschungsprojekt EvENT (Evaluating Events in Narrative Theory) an der Technischen Universität Darmstadt. Zuvor hat er an der Technischen Universität Hamburg im Projekt herMA (Automatisierte Modellierung hermeneutischer Prozesse - Der Einsatz von Annotationen für sozial- und geisteswissenschaftliche Analysen im Gesundheitsbereich) gearbeitet. Er beschäftigt sich insbesondere mit der digitalen Narratologie und der Methodik der Netzwerkanalyse.

Dominik Gerstorfer, M.A.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Dominik Gerstorfer promoviert über "Philosophische Fragen der Digital Humanities" an der Universität Stuttgart. Derzeit ist er im DFG-Projekt forTEXT tätig, zuvor war er im Digital-Humanities-Projekt CETA in Stuttgart beschäftigt. Dominik hat an der Universität Tübingen Philosophie, Politikwissenschaften und Soziologie (M.A.) studiert. Seine Forschungsschwerpunkte liegen in den Bereichen Wissenschaftstheorie, formale Methoden und Ar-

gumentationsanalyse. Im Rahmen von forTEXT beschäftigt sich Dominik u.a. mit Intertextualität, Ontologien und der Entwicklung von Kategoriensystemen.

Malte Meister, B.Sc.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Malte Meister hat 2009 sein Informatik-Diplom (B.Sc.) in Kapstadt erworben. Im Rahmen des Abschlussprojekts für sein Diplom wurde er beauftragt, das Text-Annotations und -Analysetool CATMA, für die Universität Hamburg zu erstellen. Bis Anfang 2010 wirkte er im Team an CATMA mit, bevor er sich auf seine Karriere in der freien Wirtschaft konzentrierte. Nach mehr als zehn Jahren Berufserfahrung als Softwareentwickler und Teamleiter entschied er sich, wieder in die CATMA-Entwicklung einzusteigen. Er ist seit 2021 technischer Mitarbeiter an der TU Darmstadt und beschäftigt sich dort im Rahmen von forTEXT hauptsächlich mit dem Betrieb und der Weiterentwicklung von CATMA und den damit verbundenen Systemen.

Mareike Schumacher, M.A.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Landwehrstraße 50A, 64293 Darmstadt

Mareike Schumacher koordiniert das DFG-Projekt forTEXT (<https://fortext.net>), in dem neben der Dissemination von digitalen Routinen, Ressourcen und Tools in die traditionelleren Fachwissenschaften auch die Weiterentwicklung von CATMA eine wesentliche Rolle spielt. Sie promoviert als digitale Literaturwissenschaftlerin über Orte und Räume im Roman, beschäftigt sich besonders mit den Methoden des *distant reading* (u. a. *Named Entity Recognition* oder Stilometrie), der Digital-Humanities-Theorie und der Verbindung von digitalen Methoden und theoriebasierter Literatur- und kulturwissenschaftlicher Forschung.

Fußnoten

1. CATMA (Computer Assisted Text Markup and Analysis) erscheint zum Beispiel im *TAPoR Toolverzeichnis*, sowie in „Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften“ (Frey-Endres & Simon 2021).

Bibliographie

Frey-Endres, Marcel / Simon, Tobias (2021): „Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und3 Netzwerkanalyse in den Geisteswissenschaften“. In: *Digital Philology / Working Papers in Digital Philology* 02/2021. Darmstadt: TUPrints. URL: https://tuprints.ulb.tu-darmstadt.de/17850/1/Digital_Philology__Working_Papers_in_Digital_Philology_vol002.pdf [letzter Zugriff 24. November 2021]

Gius, Evelyn / Meister, Jan Christoph / Meister, Malte / Petris, Marco / Bruck, Christian / Jacke, Janina / Schumacher, Mareike / Gerstorfer, Dominik / Flüh, Marie / Horstmann, Jan (2021): CATMA 6 (Version 6.3). Zenodo. DOI: 10.5281/zenodo.1470118. URL: <https://catma.de/> [letzter Zugriff 24. November 2021]

HISB vorgestellt: Eine virtuelle Arbeitsumgebung für die akademische Forschung wie auch die Digitalisierung von strukturierten Informationen aus Archivalien

Eine Anwendung der virtuellen Forschungsplattform Geovistory

Knecht, David

david.knecht@kleiolab.ch
KleioLab GmbH, Schweiz

Beretta, Francesco

francesco.beretta@cnrs.fr
LARHRA–CNRS/Université de Lyon/ENS, Frankreich

Hotz, Gerhard

gerhard.hotz@unibas.ch
Integrative Prähistorische und Naturwissenschaftliche Archäologie (IPNA), Universität Basel

Kontext

Die Digitalisierung von Archivalien erlaubt einen neuartigen Zugang zum kulturellen und gesellschaftlichen Gedächtnis. Zum einen sollen sich so breitere Gesellschaftsgruppen einfach mit digitalisierten Zeitzeugnissen auseinandersetzen können (Kansy 2012: 3). Zum anderen soll dank den vielfältigen Methoden zur Bearbeitung digitaler Objekte auch die Beantwortung neuer Forschungsfragen ermöglicht werden. Dabei ist es besonders herausfordernd, Arbeitsumgebungen zu schaffen, die es erlauben digitalisierte Archivalien im Sinne von primären Forschungsdaten zu verbinden mit den sekundären Forschungsdaten, welche im Rahmen der Bearbeitung eines Forschungsgegenstandes zusätzlich erfasst werden (Maier 2020). Erst recht, wenn das digitalisierte Archivgut in einem ersten Schritt prozessiert, strukturiert und in neuer Form zugänglich gemacht werden muss, damit es den nötigen Grad an Granularität der strukturierten Informationen aufweist, welcher zur Bearbeitung der Forschungsfragen von Nöten ist. Dieser Aufgabe hat sich das Projekt HISB (Historisch-genealogisches Informationssystem Basel) angenommen, welches sich auf die Bevölkerung der Stadt Basel im 19. Jahrhundert konzentriert. Das HISB hat sich zum Ziel gesetzt, Informationen zu den Bewohnern der Stadt Basel aus den Quellen des Staatsarchivs Basel-Stadt für die Forschung, wie auch für die Nutzung durch die interessierte Öffentlichkeit aufzubereiten.

Rund um das Projekt HISB

Das HISB baut dafür auf dem akkumulierten Datensatz des Bürgerforschungsprojekts Basel-Spitalfriedhof (BBS) auf, welches der Universität Basel (IPNA) angegliedert ist. Seit gut zwölf Jahren erschließen die rund 70 Bürgerforscher/innen des BBS historische Akten (1840-1870) aus dem Staatsarchiv Basel-Stadt. So entstanden Datensätze mit mehr als 300'000 historischen Personendaten. Die Daten helfen, Forschungslücken in der Basler Sozial- und Stadtgeschichte zu schließen, das Interesse an der Geschichte Basels zu wecken und die anthropologischen Forschungen an den identifizierten Skeletten des Basler Spitalfriedhofs (1845-1868) zu kontextualisieren. Sie dienen als Basis universitärer Forschung (Hotz et al. 2018).

Viele der im BBS erfassten Personen sind aber in mehreren Quellen erwähnt. Als Beispiel sei «Babette Saxer» exemplarisch angeführt, zu welcher es zwölf unterschiedliche Akteneinträge gibt. Das HISB aggregiert die vorhandenen Informationen zu jeder Person. Diese sollen über ein benutzerfreundliches Interface abgefragt, ergänzt und visualisiert werden. Dieser durch das BBS-Team erschlossene Aktenkomplex mit ausgewählten seriellen Daten soll solcherart einen detaillierten Einblick in die Lebensbedingungen beliebiger Personen und Personengruppen Basels im 19. Jahrhundert für den Zeitraum 1840-1870 erlauben. Die Auswahl der thematisch breitgefächerten Aktendossiers soll ermöglichen, eine große Anzahl unterschiedlicher Fragestellungen zu den Lebensbedingungen und dem Lebensalltag Basels im 19. Jahrhundert, auch unabhängig von Geschlecht, geografischer Herkunft und sozialem Stand der anvisierten Personen zu untersuchen. Die Erschließungen der Volkszählungen 1850, 1860 und 1870 lassen Migrationsbewegungen nach Basel und künftig sogar innerhalb der Stadt nachverfolgen.

Die Zielsetzung des HISB ist dreifach:

1. Historische Personendaten zur Stadt Basel aus dem 19. Jahrhundert sollen in einem webbasierten Informationssystem aufbereitet werden und für die akademische Forschung bereitstehen.
2. Eine webbasierte Arbeitsumgebung soll für die freiwilligen Mitarbeiter/innen des BBS erstellt werden, die es erlaubt, neue Digitalisierungen direkt im Informationssystem HISB vorzunehmen.
3. Es soll mittelfristig eine Online-Auftritt erstellt werden, über den die interessierte Öffentlichkeit im read-only Zugang direkt und auf einfach zugängliche Weise das Informationssystem HISB befragen, sowie eigene Analysen wie auch Visualisierungen erstellen kann.

Geovistory – die virtuelle Forschungsumgebung

Um die Projektziele zu erreichen, sollen die Daten des BBS in Geovistory, einer Webplattform zur Bearbeitung, Auswertung und Publikation von historischen Informationen, importiert und anschließend publiziert werden.

Geovistory ist eine vom DH-Startup KleioLab entwickelte Forschungsumgebung die sich in der Beta-Phase befindet und den Forschenden zum Testen frei zugänglich ist. Geovistory ist für geistes- und insbesondere geschichtswissenschaftliche Forschungsprojekte nach der partizipativen Methode des «User-

Experience-Design» entwickelt. Geovistory soll Forscher/innen auf innovative und einfach zugängliche Art als digitales Werkzeug unterstützen und deren Forschung auf attraktive Weise Geschichtsinteressierten zugänglich machen. Dafür bildet Geovistory den gesamten Forschungsprozess digital ab: von der Erfassung der Quellen, über die Informationsextraktion, die Verwaltung von projektspezifischen kontrollierten Vokabularen, die Verlinkung mit externen Ressourcen, den Aufbau eines Informationsnetzes und die (räumliche) Analyse der Forschungsdaten bis hin zur Web-Publikation der Ergebnisse.

Number	Name/Vorname	HISB ID	State/Canton	Description
1	Weber Gottlieb	3	Aargau	Aarau CH
2	Weber Marie	3	Aargau	No entity matched
3	Weber Gottlieb	3	Aargau	No entity matched
4	Weber Rudolf	3	Aargau	No entity matched
5	Weber Elisabeth	3	Aargau	No entity matched
6	Glaser Rudolf	9	Basel-Stadt	No entity matched
7	Glaser Catharina	9	Basel-Stadt	No entity matched
8	Glaser Maria	9	Basel-Stadt	No entity matched
9	Glaser ungetauft	9	Basel-Stadt	No entity matched

Abb. 1: Digitalisierte & verlinkte Volkszählung Basel 1860.

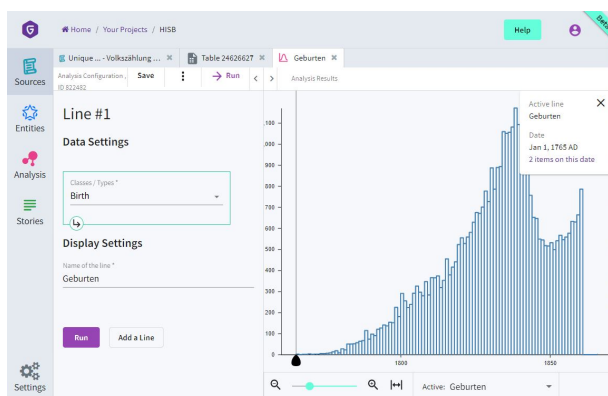


Abb. 2: Visualisierung der im HISB erfassten Geburten. Quellenangabe: Visualisierung aus der Projektumgebung des HISB. Eigene Erstellung.

Um offene und wiederverwertbare Daten nach den FAIR-Kriterien zu produzieren, baut das Datenmodell auf dem *de facto* Standard CIDOC-CRM (ISO 21127:2014) mit einigen, den Forschungsbedürfnissen entsprechenden Erweiterungen auf. Dies geschieht in Anknüpfung an die kollaborative Plattform Ontology Management Environment OntoME, betrieben vom Laboratoire de recherche historique Rhône-Alpes (CNRS / Université de Lyon) im Rahmen des Data for History Consortium. Ein Poster dieses Workflows wurde auf der jährlichen Time-Machine Konferenz 2019 präsentiert (Beretta et al. 2019).

Phase I des Projekts HISB

In der Umsetzung der ersten Projektphase (2019-2021) wurde zum einen die Plattform Geovistory maßgeblich weiterentwickelt, um den Anforderungen des HISB an eine Arbeitsumgebung für

seine Mitarbeitenden gerecht zu werden (Importer-Funktionalitäten). Zum anderen wurden erste Datensätze (Volkszählung 1850 und 1860 der Stadt Basel) des BBS ins HISB importiert und integriert (gematched). Bei beiden Aufgaben stellten sich verschiedene Herausforderungen typisch für ein Digitalisierungs- und Datentransformations-Projekt, das darauf abzielt, fein-granulare, sauber strukturierte und semantifizierte Daten zu produzieren. Zwei Herausforderungen waren besonders gross:

Erstens hat die Entwicklung des Importer-Funktionalitäten ein vielfaches der ursprünglich budgetierten Zeit gebraucht: Zum einen wegen der Komplexität der funktionalen Anforderungen, zum anderen wegen technischen Herausforderungen, um die neue Komponente in das Gesamtsystem zu integrieren.

Zweitens stellte sich das Matching der Volkszählung 1850 und 1860 als wissenschaftlich anspruchsvolle Aufgabe heraus. Die Herausforderungen waren dabei auf mehreren Ebenen: Innerhalb der einzelnen Tabellen von 1850, respektive 1860 gab es Inkonsistenzen wie beispielsweise verschiedene Schreibarten für die Konfession „katholisch“, die vereinheitlicht werden mussten. Anspruchsvoller aber war es Unterschiede in der Datenerhebung 1850 und 1860 zu identifizieren und zu bereinigen, um ein Matching dieser Quellen zu ermöglichen. Beispielsweise wurde 1850 bei einer in Basel wohnhaften Person die Staatszugehörigkeit „Königreich Sardinien“ vermerkt. Im Jahr 1860 aber dann „Königreich Italien“, welches dazumal gerade gegründet wurde. Solche Unterschiede beruhen auf historische Gegebenheiten, zu deren Interpretation es eine/n kundige/n Historiker/in braucht. Nur so kann anschließend ein sinnvolles Matching durchgeführt werden. Das heißt, dass eine enge Zusammenarbeit zwischen den Domänen- & Quellenexperten sowie den Data Engineers notwendig ist. Diese Herausforderungen führten dazu, dass das Matching ebenfalls deutlich mehr Zeit beanspruchte als ursprünglich budgetiert.

Insgesamt aber ist die HISB Phase I klar als Erfolg zu werten. Denn die gestellte Aufgabe konnte gemeistert werden und mit dem HISB wurde eine webbasierte Plattform bereitgestellt, die

1. den Mitarbeitenden des BBS dienen kann, künftig direkt in einem Informationssystem strukturierte Daten aus den Archivalien zu erfassen. Das vereinfacht zu großen Teilen den Prozess der Datenaggregation (Matching).
2. der Forschung dienen kann, um darauf aufbauend Forschungsfragen zu bearbeiten. Dies dank den bestehend ausgeklügelten Daten-Kurationen, sowie Visualisierungs- und Analyse-Funktionalitäten.
3. den Sockel bietet, um weitere schon digitalisierte Datenquellen zu importieren und zu integrieren (matchen), um so ein stetig reicheres Informationssystem für Basel einer breiten Öffentlichkeit bereitzustellen.

Hands-on Workshop an der DHd-Konferenz 2022

Der angebotene Workshop ist als Tutorial geplant. Er soll den Teilnehmenden ermöglichen, einen Einblick in ein spannendes und ambitioniertes Digitalisierungsprojekt zu gewinnen, inklusive dessen zu meisternde Herausforderungen wie auch großes Potential. Dabei ist die Zielsetzung des Workshops zweifach:

- Zum einen möchten wir gemachte Erfahrungen – positive wie negative – in der Konzeptualisierung und Umsetzung einer virtuellen Arbeitsumgebung zur fortlaufenden Digitalisierung

von Archivgütern (im Sinne der Bereitstellung fein-granularer, strukturierter Daten) teilen.

- Zum anderen möchten wir die Teilnehmenden einladen, die Funktionalitäten der virtuellen Arbeitsumgebung von HISB hands-on zu testen. Im Sinne der Datenerfassung, aber auch der Datenanalyse und –visualisierung.

Ablauf Workshop

- Teil 1: Einführung in das Projekt HISB: Kontext & Grundlage des HISB und der Plattform Geovistory, Erfahrungen & Resultate der HISB Phase I.
- Pause
- Teil 2: Hands-On-Einblick ins HISB & die darunterliegende Plattform Geovistory. Anhand von Beispielesdaten folgen wir dem Arbeitsprozess vom HISB und testen direkt in der Webplattform Geovistory die verschiedenen Funktionalitäten (Arbeit mit digitalisierten Quellen, Annotation der Digitalisate, Erfassen fein-granularer Informationen, Abfrage & Visualisierung dieser Informationen etc.).
- Teil 3: Offene Diskussion, um den Workshop anhand der gewonnen Eindrücke und der eigenen Erfahrungen der Teilnehmenden zu reflektieren.

Zielpublikum und Anforderungen

- Für diesen Workshop sind keine technischen Vorkenntnisse erforderlich. Der Workshop ist gedacht für Personen/Institutionen, die sich mit Digitalisierungsprojekten beschäftigen, wie auch für Forschende, die eine neue virtuelle Arbeitsumgebung kennenlernen möchten.
- Es können 20 bis 30 Personen am Workshop teilnehmen. Jede/r Teilnehmende benötigt einen Laptop mit Wifi-Zugang.

Vortragende

Dr. habil. Francesco Beretta, LARHRA–CNRS/Université de Lyon/ENS, Frankreich. francesco.beretta@cnrs.fr : Francesco Beretta ist habilitierter Kirchenhistoriker. Seit mehr als zehn Jahren forscht er am LARHRA in Lyon zu Fragen semantischer Modellierung historischer Informationen und der Entwicklung geohistorischer Informationssysteme.

M.A. David Knecht, KleioLab GmbH, Basel. david.knecht@kleiolab.ch : David Knecht ist Ökonom. Bei KleioLab ist er verantwortlich für die Begleitung von Forschungsprojekten und interessiert sich insbesondere für Fragen der Datenaggregation aus verschiedenen Quellen.

Co-Autor

Dr. Gerhard Hotz, Integrative Prähistorische und Naturwissenschaftliche Archäologie (IPNA), Universität Basel. gerhard.hotz@unibas.ch : Gerhard Hotz ist Anthropologe. Er leitet seit über zehn Jahren das Bürgerforschungs-projekt Basel-Spitalfriedhof (BBS) und verantwortet das HISB. Seine Forschungsinteressen drehen sich um interdisziplinäre Fragestellungen zwischen Anthropologie, Archäologie, Natur- und Geisteswissenschaften – insbesondere im Bereich Gesundheit, Krankheitsbelastung, Gesellschaft und Umwelt.

Bibliographie

Beretta, Francesco / Alamertery, Vincent / Derks, Sebastian / Petram, Lodewijk / Schneider, Jonas (2019): “Geohistorical FAIR data: data integration and Interoperability using the OntoME platform.” in *Time Machine Conference 2019*, Dresden. <https://halshs.archives-ouvertes.fr/halshs-02314003> [letzter Zugriff am 14. Juni 2021].

Hotz, Gerhard / Schneider, Jonas/ Beretta, Francesco / Knecht, David (2018): „Produktevision Historisch-Genealogisches Informationssystem Basel (HISB)“ KleioLab, Basel.

Kansy, Lambert (2012): „Digitalisierungsstrategie. Strategie für Digitalisierung von Archivgut (2013-2018).“ Präsidialamt des Kantons Basel-Stadt, Staatsarchiv. Basel.

Maier, Gerald (2020): „Die Bedeutung der Archive für Forschungsdaten in der Geschichtswissenschaft . Verband der Historiker und Historikerinnen Deutschlands“ in *Blog-Post*. <https://blog.historikerverband.de/2020/11/11/die-bedeutung-der-archive-fuer-forschungsdaten-in-der-geschichtswissenschaft/> [letzter Zugriff am 17. November 2021].

Introduction to Docker

Lampert, Marcus

lampert@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Short Description

Overview

First introduced in 2013, Docker is one of those tools that, not unlike the version control software Git, has become nearly ubiquitous in software development. As a lightweight virtualization software, it elegantly solves common problems in software engineering related to developing and deploying applications. Docker creates independent operating environments, so-called ‘containers’, with which one can deploy applications in a flexible and reproducible manner. A single Docker command on the command line can launch an entire application, either locally or on a server, with any number and kind of desired frontends, backends, services, databases, etc. This containerized approach to application development has many advantages for software engineering in general, and for the Digital Humanities specifically, including:

- Ease of setting up an existing project on a new server or personal computer. This is why, for example, open source projects often offer a Docker image, or describe how to set up their application using a standard Dockerfile.
- Continuity between local development and production. Building a project locally on a personal computer with Docker, versus building for production with Docker are often similar, if not identical processes.
- Broad language and version support. A server that runs Docker only needs to have Docker installed in order to support basically every major language and their various versions. If

a Docker image exists for a given language and version, then that application can be run. Specifically for the Digital Humanities, where projects have a long duration, this feature could potentially prolong the time for which an application can be supported.

In this workshop, we offer participants a practical introduction to Docker and show by example how it can be integrated into existing and new Digital Humanities projects. The two examples we use are Digital Humanities projects developed at the Berlin-Brandenburg Academy of Sciences and Humanities and are exemplars of two standard application types used in the Digital Humanities: a web application based on an SQL database, and a web application based on an XML NoSQL database. Participants will follow along with the presenter, deploying these applications on their own laptops using Docker. We will also look together at a live Docker application running on a server at the Berlin-Brandenburg Academy of Sciences and Humanities. By the end of the workshop, participants will have a general idea of what Docker is and how to use it, and can begin thinking about how they can use Docker to facilitate their Digital Humanities projects.

Background

The workshop is based on a workshop developed internally at the Berlin-Brandenburg Academy of Sciences and Humanities. At the Academy, we have recognized the advantages that Docker would bring for a number of our projects. This workshop is designed to be practical and accessible to a broad audience. It covers basic concepts and commands that arise in one's regular work with Docker.

Workshop Structure and Content

The workshop takes four hours, including breaks. It consists of four modules, each module lasting about 45 minutes followed by a 15 minute break:

- Introduction
- Docker
- Docker Compose
- Deployment

In each of these modules, the presenter will explain and demonstrate core concepts on his computer, and then participants will be asked to execute a number of Docker commands on their computers. At least one assistant will assist the presenter in going around the room helping people.

Module 1: Introduction

Topics covered:

- Installing Docker and Docker Compose
- Description of Docker: a software platform for enclosing applications and application components in containers that are encapsulated but share an operating system, i.e. lightweight virtualization
- Advantages of Docker:
 - Reproducible application builds in a reproducible environment.

- Can run multiple applications on a host, local or a server, without colliding dependencies.
- Easy to deploy, locally or on a server, an application with multiple parts (frontend, backend, database, etc).
- Disadvantages of Docker:
 - Learning curve
 - Additional work of dockerizing a project
 - Managing Docker on a server, potential security problems.
- The building blocks of Docker: images, containers, volumes, networks, etc.

Module 2: Docker

Objective: learn basic Docker commands.

Topics Covered:

- Anatomy of a docker project
 - 'Dockerfile'
 - '.dockerignore' (optional)
- Starting the example application without docker
 - 'cp .env.example .env'
 - 'php composer install or compose install'
 - 'php artisan serve'
- Starting applications with docker
 - 'docker build .'
 - 'docker run [IMAGE ID]'
 - From a new terminal window: 'docker ps'
 - 'docker exec -ti [CONTAINER ID/NAME] bash'
- A few problems:
 - Cannot access the port (see 'docker network ls')
 - Image ID is not very descriptive
- Stopping the container:
 - 'docker ps'
 - 'docker stop [CONTAINER ID/NAME]'
- Starting the applications with docker, improved
 - 'docker build -t [NAME] .'
 - 'docker run -p 8000:8181 [NAME]'
- Starting applications in the background
 - 'docker run -d -p 8000:8181 [NAME]'
 - 'docker logs -f [CONTAINER ID/NAME]'
- Starting verses running a container
 - 'docker container ls'
 - 'docker start [CONTAINER ID/NAME]'

Module 3: Docker Compose

Objective: learn how a docker-compose.yml file is composed and learn basic Docker Compose commands.

Topics Covered

- Problems with just using Docker
 - Stopping containers is annoying
 - Important things like ports, tags, network, volumes, etc. are only giving at runtime
 - How would you handle a project with multiple containers?
- Docker commands one might often use
 - 'docker exec -ti [CONTAINER ID/NAME] bash' - If you have a database in a docker container, for example
 - 'docker ps', 'docker image ls', 'docker volume ls'
- Basic Docker Compose commands
 - 'docker-compose build [services]'
 - 'docker-compose up [services]'
 - 'docker-compose up --build [services]'

- ‘docker-compose down’, ‘docker-compose down -v’
- Common things we do in a docker-compose.yml file
 - expose ports
 - talk with other containers in the same network
 - set environment variables
 - define volumes
- Use cases of Docker Compose
 - Production: deploy with one command
 - Development: start those services you aren't actively working on.

Module 4: Deployment

Objective: Gain a general idea of how to use Docker to deploy an application on a server

Topics Covered:

- One docker-compose.yml for each environment.
 - local.docker-compose.yml
 - production.docker-compose.yml
- Merging Multiple docker-compose.yml files: <https://docs.docker.com/compose/extends/>
- Vendor docker-compose.yml with a shell script.
 - Define a template.yml with variables:


```
environment:
  - BACKEND_URL: https://{BACKEND_URL}:{BACKEND_PORT}
```
 - For each env define values for the variables. So, env.local looks like:


```
BACKEND_URL=localhost
BACKEND_PORT=1234
```
 - Have shellscript execute ‘envsubst’ to generate the docker-compose.yml:


```
Input: ‘devops/vendor.sh local’
Output: ‘docker-compose.yml’
```

Presenter

Marcus Lampert

Wissenschaftlicher Mitarbeiter an der Berlin-Brandenburgischen Akademie der Wissenschaften
lampert@bbaw.de

Marcus works as a software engineer on various Digital Humanities projects at the Berlin-Brandenburg Academy of Sciences and Humanities. He received his PhD 2015 in German Literature from the University of Chicago writing on the content and influence of Johann Gottlieb Fichte's 1794 *Science of Knowledge*. Prior to joining the Academy, he worked for three years as a software engineer at PricewaterhouseCoopers Deutschland. He enjoys working as a fullstack developer, fiddling on both the frontend and backend, and solving devops challenges. Marcus is driven by the quest to use modern technology to facilitate the sharing of knowledge.

Format of the Workshop and Target Audience

- Language: German, or English depending on the needs of the participants.

- Up to 25 participants. No prior knowledge of Docker required.
- Participants should have an interest in writing and/or deploying websites or other related software applications.
- Basic use of the command line is helpful but not required.
- Experience with Linux systems is helpful but not required.

Participants should have a personal laptop with the following specs:

- admin rights so that they can install software such as Docker and Docker Compose.
- Preferably Mac or Linux as Operating System.
- Ability to connect to the internet.
- Some remaining storage space (50 GB) for using Docker.
- Very old machines might have performance or compatibility issues with Docker.

Technical Equipment

- Large display monitor/screen to connect to presenter's laptop.
- Power outlets for the participants' computers.
- Wireless internet connection for all the participants.

Bibliography

Docker Documentation. <https://docs.docker.com/> (accessed 1 December 2021).

En.wikipedia.org. Docker (software) - Wikipedia. [https://en.wikipedia.org/wiki/Docker_\(software\)](https://en.wikipedia.org/wiki/Docker_(software)) (accessed 1 December 2021).

Manifest für digitale Editionen

Fritze, Christiane

christiane.fritze@wienbibliothek.at
Wienbibliothek im Rathaus, Austria

Einleitung

Das Institut für Dokumentologie und Editorik sieht aus langjähriger Erfahrung mit vielen Editionsprojekten heraus einen drängenden Bedarf, die besonderen Rahmenbedingungen für digitale Editionen und einige gegenwärtig unbefriedigende Aspekte der wissenschaftlichen Arbeit bei allen Stakeholdern deutlicher zu machen um eine Verbesserung der Situation zu bewirken. Es sieht dafür die Form eines Manifests vor, das im Rahmen eines DHD-Workshops erarbeitet werden soll.

Digitale wissenschaftliche Editionen werden in der heutigen Zeit gewöhnlich in Form von Projekten von interdisziplinären Teams realisiert, die mit Drittmitteln gefördert werden. Den Anforderungen der Fördergeber hinsichtlich Nachhaltigkeit, Langzeitarchivierung und Berücksichtigung der FAIR-Prinzipien muss ebenso entsprochen werden wie den gestiegenen Bedürfnissen

der Nutzer*innen. Die gegenwärtige Praxis des Edierens impliziert mehrere Probleme, die in den Projekten regelmäßig adressiert werden müssen und Gegenstand des Manifests sind.

Zielgruppe des Manifests

Das Manifest richtet sich sowohl an Fördergeber, an Forschungseinrichtungen und das kulturelle Erbe bewahrende Infrastruktureinrichtungen als auch an Editionswissenschaftler*innen.

Zeitlichkeit und Nachhaltigkeit

Projekte sind per definitionem zeitlich begrenzte Unternehmungen. Eine digitale Edition ist mit ihrer Publikation aber nicht abgeschlossen, sondern hat ein vielfältiges Nachleben. Grundsätzlich sind Editionen darauf angelegt, ihre Ergebnisse langfristig nutzbar zu halten, im Idealfall sollen Editionen aber auch kontinuierlich weiterentwickelt werden: Daten müssen dauerhaft erreichbar sein, neue Erkenntnisse und Benutzer sollen einbezogen werden, digitale Systeme und Publikationen müssen kontinuierlich gewartet werden. Es bedarf also auch nach der Erstellung der digitalen Edition einer Infrastruktur, die die digitale Edition langfristig bewahrt und zur Nachnutzung verfügbar hält. Voraussetzung ist jedoch eine saubere Trennung zwischen Editionsdaten, die für sich genommen die Forschungsergebnisse der Editor*innen enthalten müssen, und der "Dissemination" (im Sinne des OAIS-Referenzmodells), die die Benutzerinteraktion mit der Edition realisiert. Konsequenterweise sollten sowohl die Editionsdaten als auch die genutzte Forschungssoftware nach den FAIR-Prinzipien aufbereitet, archiviert und zugänglich gemacht werden. Dies schließt auch die Erstellung der die Edition begleitenden Metadaten und eine persistente Identifikation mit ein. Die Zeitlichkeit der Edition schließt aber den verantwortungsbewussten Umgang mit Änderungen mit ein: Die digitale Edition braucht eine konsistente Versionierung (z.B. nach dem Modell des Semantic Versioning, <https://semver.org>).

Aus den FAIR-Prinzipien folgen Ansprüche an die Übernahme der digitalen Edition in Nachweissysteme ("Findable"), an die genaue Deklaration der Nutzungsrechte ("Accessible"), an die technische Verfügbarkeit z.B. über etablierte APIs ("Interoperable") und die Realisierung der Daten nach in der Community geläufigen Standards ("Reusable"). Standards müssen dabei so angewendet werden, dass sie konsistent innerhalb des Projektes und zur Community verwendet werden, sowie dass alle Abweichungen ausreichend dokumentiert sind.

Die FAIR-Prinzipien sind aber nur ein Mindeststandard. Erst eine Kapselung der Funktionalitäten in Containerformaten, eine kontinuierliche Betreuung oder eine Migration der Applikation erlaubt bei digitalen Editionen eine Benutzererfahrung, die den Erfahrungen mit der Zeitlichkeit des Mediums Buch gleichkommt.

Mindeststandards und Best Practices

Es sollte sich auch von selbst verstehen, dass bestimmte Mindeststandards eingehalten und Best Practices des digitalen Edierens angewendet werden. Dazu gehört zuvorderst, dass das digitale Paradigma von allen an der digitalen Edition Beteiligten verstanden und ernst genommen wird. Erst dann kann eine alle

Aspekte umfassende Diskussion zur Datenmodellierung mit allen Beteiligten erfolgen, die Grundlage für die Erstellung der Edition und möglichen sich anschließenden Nutzungsszenarien und Auswertungen ist.

Rollen und Zusammenarbeit

Treiber für digitale Editionen sind in der Regel Fachwissenschaftler*innen, die den Bedarf sehen, bestimmte wissenschaftlich relevante Quellen für die Nutzung durch die Fachcommunity aufzuschließen. Die Komplexität digitaler Editionen erfordert aber die Einbindung weiterer Akteure, vor allem aus dem methodischen, gestalterischen und technischen Bereich. Dabei entsteht häufig dadurch eine Schieflage, dass die zuletzt genannten Partner oft lapidar als "Techniker" bezeichnet, und nicht als Wissenschaftler*innen auf Augenhöhe akzeptiert werden. Um eine kritische Edition digital zu entwickeln, produzieren und zu präsentieren, bedarf es einer umfassenden Analyse des zu edierenden Materials hinsichtlich der Modellierung und Verdattung. Digitale Workflows müssen konzipiert werden, für die zu erschließenden Quellen muss ein passendes Datenmodell, das gängigen Standards zu entsprechen hat, erarbeitet werden, dieses Modell muss spätere Analysen, Präsentationen und noch nicht bekannte Nutzungsszenarien erlauben. Werkzeuge und ggf. "research software" müssen entwickelt und sachadäquate Präsentationsformen entworfen werden. Zudem ist der Produktlebenszyklus von Software und Online-Gegebenheiten im Allgemeinen kurz, so dass ein zusätzlicher Aufwand darin liegt, auf aktuelle Entwicklungen zu reagieren und immer wieder neue Lösungen zu finden.

Bei der Arbeit an komplexen digitalen Forschungsvorhaben treffen nicht nur distinkte Kompetenzbereiche, sondern auch unterschiedliche Arbeitsmethoden und Kommunikationskulturen aufeinander: Software-Entwickler*innen arbeiten derzeit häufig in agilen Settings mit streng getakteten Sprints, die keine Unterbrechungen dulden, Fachwissenschaftler*innen benötigen kurzfristig auf Zuruf Unterstützung. Es ist notwendig Kenntnis von und Verständnis für die jeweils andere Arbeitskultur zu haben und aufzubringen, um sie in das eigene Vorgehen ohne Reibungsverluste integrieren zu können. Nicht zuletzt müssen auch ausgeprägte Projektmanagementkompetenzen im Team vorhanden sein, um das Editionsprojekt erfolgreich durchführen zu können.

Schließlich müssen digitale Editionen sich auch frühzeitig mit den Personen in Verbindung setzen, die die Infrastrukturen für die Langzeitverfügbarkeit und den Nachweis betreiben (Repositorymanager*innen, Web-Hosts, Bibliothekar*innen etc.).

Workshop-Organisation

Die verschiedenen oben angerissenen Aspekte werden in einzelnen Abschnitten des Manifests in Kleingruppen erarbeitet.

Als zu diskutierender Vorschlag wären in einer Gliederung dabei zu unterscheiden: Präambel, Rahmenbedingungen digitaler Arbeit, sachliche Dimension (Mindestanforderungen an digitale Ressourcen), soziale Dimension (Rollen und Rollenverständnisse) und organisatorische Dimension (Editionen als Unternehmen).

Die Organisator*innen des Workshops werden zunächst in die Thematik einführen und für die aufgezeigten Probleme sensibilisieren. Anschließend wird das Konzept des Manifests vorgestellt werden. Vor der Einteilung in Kleingruppen wird Raum gegeben,

eigene Erfahrungsberichte beizusteuern. Jede Kleingruppe wird in einem Etherpad einen Abschnitt des Manifests bearbeiten. Um den Schreibprozess zu verkürzen sind die Etherpads mit Thesen und Statements vom IDE e.V. vorbereitet, die zur Diskussion und Weiterentwicklung angeboten werden. Parallel zur Formulierung der Forderungen im Manifest können entsprechende praktische Hinweise und Umsetzungsvorschläge entstehen, die am Ende in einem gesonderten Handreichungsdokument zusammengefasst werden. Nach der Gruppenarbeit werden die entstandenen Formulierungen im Plenum diskutiert, gegebenenfalls reformuliert, sodass sie von allen Teilnehmer*innen und dann Autor*innen des Manifests mitgetragen werden. Das Manifest soll inhaltlich einen Zustand erreichen, in dem es zu Konferenzende in einer Alpha-Version zur Veröffentlichung unter der Lizenz CC BY bereit steht. Nach dem Workshop wird es eine redaktionelle Überarbeitungsschleife geben, um den Gesamttext formal zu harmonisieren. Gegebenenfalls werden unfertige Abschnitte in einer virtuellen Session zu Ende geführt. Die Veröffentlichung wird in einem Forschungsdatenrepositorium erfolgen. Die Abschlussdiskussion wird vornehmlich darin bestehen, wie das Outreach zum Manifest aussehen kann und wie die virtuelle Zusammenarbeit für die Fertigstellung der zugehörigen praktischen Handreichungen organisiert wird.

Organisation/Format

- Umfang: halbtägiger Workshop
- Anzahl Teilnehmende: max. 30
- Zielgruppe: Editionswissenschaftler*innen, Digital Humanists. Kein Vorwissen notwendig.
- Ausstattung: Steckdosen, Internetzugang via WLAN
- Gruppenarbeit, jede Gruppe erarbeitet einen Aspekt
- ggf. Einbezug von anderen Akteuren via Social Media
- synchrones Schreiben / Gesamtreaktion

Output des Workshops / Ergebnisverwertung

- Manifesto als Open Access Dokument mit grafisch ansprechendem Layout, in Web- und Printformat
- ebenfalls als offene, versionierte (z.B. Git) Dokumente: Handreichungen für Digitale Editionen, welche die Grundsätze des Manifestos in konkrete Handlungsempfehlungen übersetzt
- perspektivisch: Übersetzung des Manifestos in mehrere Sprachen für die internationale Community und um dem Aspekt der barrierearmen Zugänglichkeit gerecht zu werden.

Bibliographie

- Morgan, Megan** (2021): *How To Write a Manifesto*. <https://www.wikihow.com/Write-a-Manifesto>
- Deutsche Forschungsgemeinschaft** (2015): *Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft*. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf
- Hong, N. P. C., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T.** (2021).

FAIR Principles for Research Software (FAIR4RS Principles). Research Data Alliance. DOI: 10.15497/RDA00065

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. (2016): "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>

Optimiertes Peer Reviewing in den Digital Humanities

Guhr, Svenja

svenja.guhr@tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Steyer, Timo

t.steyer@tu-braunschweig.de
Universitätsbibliothek Braunschweig, Deutschland

Scholger, Walter

walter.scholger@uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Dieckmann, Lisa

lisa.dieckmann@uni-koeln.de
Universität zu Köln, Deutschland

Reiter, Nils

nils.reiter@uni-koeln.de
Universität zu Köln, Deutschland

Wuttke, Ulrike

ulrike.wuttke@fh-potsdam.de
FH Potsdam, Deutschland

Beschreibung des Themas

Peer reviewing gilt auch in den Digital Humanities als die wissenschaftliche Qualitätssicherungsmaßnahme für Publikations- und Veranstaltungsformate. Die Umsetzung, ob als *open*, *blind* oder *double-blind* steht dabei schon seit einigen Jahren in der Diskussion und - ungeachtet des gewählten Verfahrens - in der Kritik. Bei der DHd-Jahrestagung 2022 wurde nach einigen Jahren des *blind peer reviewing* zum ersten Mal eine Variante des *open peer reviewing* für diese Tagungsreihe eingesetzt. In diesem neuen Ansatz erfahren die Autor:innen die Namen der Gutachter:innen, was

für mehr Transparenz, Qualität und Fairness des Begutachtungsprozesses sorgen soll.

Rund um das Begutachtungsverfahren waren aber nicht nur die Form des Reviewings, sondern auch die Begutachungskriterien der DHd-Jahrestagungen in den letzten Jahren immer wieder Gegenstand der Diskussion innerhalb der Community. Dies ergab sich nicht zuletzt daraus, dass die Kriterien seit der 1. DHd-Jahrestagung in Passau nicht umfassend weiter spezifiziert wurden sowie teilweise unscharfe Begriffe und Kategorien beinhalten. Eine weitere Besonderheit der Digital Humanities (im deutschsprachigen Raum) ist die Interdisziplinarität bzw. die Diversität der Fachtraditionen in der Wissenschaftskommunikation (z.B. Kritiklevel) und im Reviewprozess. Gleichzeitig ist die DHd-Community neuen experimentellen Formaten aufgeschlossen, weil sich noch keine Traditionen etabliert haben und dadurch die Gestaltungsfreiheit bei der Community liegt.

Um diesen Überlegungen Rechnung zu tragen, gründete sich die Task Force “Optimized Peer Review” (OPR), die durch einen bei der vDHd veranstalteten Workshop (Burghardt et al. 2021) um weitere Mitglieder ergänzt wurde. In Absprache mit dem DHd-Vorstand sowie dem aktuellen Programmkomitee entwickelte die Task Force eine Handreichung für das *peer reviewing* der DHd 2022 mit weiterführenden Erklärungen der Kriterien und Empfehlungen zur Begutachtung in der DHd-Community, die bereits für den Reviewprozess der DHd-Jahrestagung 2022 an die Reviewer:innen verteilt wird.

Ein globales Ziel der Handreichung ist es, eine größere Einheitlichkeit in der Anwendung der Begutachungskriterien für Beiträge zur DHd-Jahrestagung zu schaffen.

Diese neuen Entwicklungen und die damit verbundenen Diskussionen und Beschlüsse sind gerade für noch unerfahrene Reviewer:innen nicht immer leicht nachzuvollziehen, daher möchte die erwähnte Task Force mit diesem Workshop ein Format anbieten, um sich über die Begutachtungspraxis der DHd zu informieren, einen Ort für Austausch zwischen mehr und weniger erfahrenen Reviewer:innen zu schaffen, aber auch konkret das Schreiben von Gutachten zu erproben.

Ablauf

Der Workshop wird eine Mischung aus theoretischen Input- und interaktiven Diskussions- und Anwendungsphasen sein. In den Inputphasen wird ein Überblick über die unterschiedlichen Varianten des *peer reviewing* sowie eine Darstellung aktueller Diskurse aus verschiedenen wissenschaftlichen Perspektiven zum Begutachtungswesen gegeben. Hiermit wird ein möglichst einheitlicher Wissensstand geschaffen und Grundlage für die anschließenden Diskussionsrunden gelegt. Diese finden in Kleingruppen statt, in denen Erfahrungsaustausch und konstruktive Diskussion an erster Stelle stehen. Ziel ist es, Feedback zum derzeitigen Reviewprozess der DHd-Jahrestagungen sowie der Interpretation der Reviewkriterien zu bekommen. Daher versteht sich der Workshop nicht nur als Format für die Wissensvermittlung zum neuen Reviewverfahren, sondern auch als diskursiver Raum, um die weitere Gestaltung künftiger Begutachtungsverfahren transparenter zu gestalten und noch enger an die Bedarfe der Community zu koppeln. Zu Beginn des Workshops stellt das Organisationsteam das Thema, den Workshopablauf sowie die Ziele des Workshops vor. Darauf folgt eine Aktivierungsphase: Didaktisch organisiert als „Think-Pair-Share“-Aufgabe sind die Teilnehmenden dazu aufgefordert, zunächst in Einzelarbeit (*Think-Phase*) orientiert an einer Leitfrage ihre Erfahrungen mit dem *peer-reviewing*-Verfahren zu reflektieren.

Anschließend werden diese Reflektionen in Partnerarbeit (*Pair-Phase*) mit dem Ziel diskutiert, positive und negative Erfahrungen der Teilnehmenden zu sammeln und insbesondere Bedarfe der Teilnehmenden im Rahmen des Begutachtungsprozesses zu benennen: Welche Unterstützung wünschen sich die Teilnehmenden von Programmkomitee und Verband? Wie kann negativen Reviewerfahrungen aktiv vorgebeugt werden? Wie lassen sich künftige Reviewprozesse aufgrund der gemachten Erfahrungen optimieren?

In der *Share-Phase* soll der Fokus nach einer kurzen Schilderung der positiven Erfahrungen auf die negativen Erfahrungen gelenkt werden, zu denen sodann in Kategorien gebündelt die Ideen und Lösungsvorschläge der Teilnehmenden im Plenum diskutiert werden.

Diese erste Arbeitsphase dient der Aktivierung und Sensibilisierung der Teilnehmenden für das Thema Reviewing und für die Komplexität von negativen Erfahrungen im Reviewprozess sowie seiner Vereinheitlichung und Qualitätssicherung.

Die *Share-Phase* dient gleichzeitig als Überleitung zum Überblicksvortrag der Beitragenden, in dem die Ideen und ersten Umsetzungen vorgestellt werden, die die Task Force innerhalb des letzten Jahres erarbeitet haben:

- Handreichung: Empfehlungen für besseres Reviewing
- Handreichung: Schärfung der Reviewkriterien
- Community Maßnahmen: Ombudsstelle, Review Award, Community Engagement

Anschließend gibt es eine moderierte Diskussionsrunde zum Vorgehen und zu den Ideen der Task Force, aus der sich die Beitragenden konstruktives Feedback versprechen. Außerdem soll die Diskussionsrunde als Möglichkeit dienen, den Teilnehmenden Teilhabe an der weiteren Gestaltung des Reviewprozesses (inklusive flankierender Maßnahmen wie der Ombudsstelle und dem *best review award*) zu gewähren.

Nach der Diskussion wird in einem praxisorientierten Teil auf das Schreiben von konstruktiven Gutachten eingegangen. Anhand von Fallbeispielen geben erfahrene Gutachtende und die Beitragenden in kleinen Gruppen Tipps für das Formulieren von Lob und Kritik, für die generelle Bewertung von Einreichungen orientiert an der Handreichung und gehen auf die Fragen der Teilnehmenden ein. Auch wird anhand der Fallbeispiele konkret das Formulieren von konstruktivem Feedback erprobt. Durch diese Einheit wird das im Workshop vermittelte Wissen um exemplarische Beispiele ergänzt und so die Hürde für das Schreiben eines ersten Reviews gemindert.

Den Abschluss des Workshops bildet eine Ergebnisdokumentation, die zusätzlich zu einer Workshopdokumentation mit den Teilnehmer:innen und der Community über einen DHd-Blogpost geteilt wird. Zudem wird die Dokumentation in die weitere Arbeit der Task Force einbezogen werden.

Bei diesem Workshop profitieren die Teilnehmenden sowie die Beitragenden vom gemeinsamen Wissens-, Erfahrungs- und Ideepool und integrieren so aktiv die DHd-Community in die Gestaltung des *peer-reviewing*-Prozesses der zukünftigen Jahrestagungen.

Ablaufplan

Phase	Inhalt(e)	Sozialform und Methode(n)	Medien	Zeit in min
1. Begrüßung		Plenum		7
2. Vorstellung	Organisationsteam, Teilnehmende, Thema, geplanter Ablauf		Beamer-Präsentation	7
3. Aktivierungsphase	„Think-Pair-Share“-Aufgabe			45
a) <i>Think-Phase</i>	Reflektion <i>peer-reviewing</i> -Erfahrungen orientiert an Leitfragen	Einzelarbeit	Blatt DIN A4 + Stift	5
b) <i>Pair-Phase</i>	Positive Erfahrungen herausstellen, Lösungen für neg. Erfahrungen formulieren	Partner:innenarbeit	Blatt DIN A3 + Stifte	15
c) <i>Share-Phase</i>	Diskussion der Ideen und Lösungsvorschläge	Plenum, Gruppengespräch	Sammeln der Beiträge in Beamer-Präsentation	25
4. Pause				15

Phase	Inhalt(e)	Sozialform und Methode(n)	Medien	Zeit in min
5. Input				(77)
a) Vortrag A	Überblick zu Begutachtungsvarianten - Was ist <i>peer reviewing</i> ? (generell) - Was ist <i>open peer reviewing</i> ? (spezifisch) - verschiedene <i>peer-reviewing</i> -Varianten (Überblicksfolie mit Varianten in Spalten)	Vortrag	Beamer-Präsentation	10
b) kurze Interaktion	Zeit zur Klärung von Rückfragen	Plenum		7
c) Impulsvorträge	Reviewroutinen in verschiedenen Fachdisziplinen; allgemeine Erfahrungen aus drei verschiedenen Perspektiven (Informatik, Germanistik, DH2020)	Vortrag	Beamer-Präsentation	15
d) Interaktion	Frage ans Plenum: Wie läuft das <i>Reviewing</i> in Ihrer Disziplin? (Leitfrage für moderierten Austausch)	Plenum		10
e) Vortrag B	Infos zur <i>Task Force OPR</i> - spezifisch im DHd-Kontext: normale Begutachtungspraxis bis 2020 - Entscheidung in der Mitgliederversammlung 2020 - Gründung der Task Force - Aufgaben der Task Force - Zwischenergebnisse: Abschnitte der Handreichung, Ombudstelle, Coaching für neue Reviewer:innen	Vortrag	Beamer-Präsentation	20
6. Diskussion	moderierte Diskussionsrunde zu Vorgehen/Ideen der Task Force, Fokus: Handreichung	Plenum, Gruppengespräch		15
7. Pause				15

Phase	Inhalt(e)	Sozialform und Methode(n)	Medien	Zeit in min
8. Coaching				(55)
a) Impuls	Wie schreibt man ein gutes Review? Wie wird konstruktives Feedback formuliert? Was sind die Hürden beim Schreiben des ersten Reviews?	experimentelle Mischung: Plenum, Gruppengespräch, Vortrag	Sammeln der Beiträge in Beamer-Präsentation	15
b) Fallbeispielarbeit	Formulierung von konstruktivem Feedback anhand zur Verfügung gestellter Beispiele (3 Texte für 6 Gruppen) hinsichtlich der Kriterien in der Handreichung	Partner:innenarbeit	Laptops oder Blatt DIN A4 + Stift	25
c) Diskussion	Diskussion der Ergebnisse	Plenum	Sammeln der Beiträge in Beamer-Präsentation	15
9. Ergebnisse	Dokumentation der Ergebnisse	Plenum, Gruppengespräch	Sammeln der Beiträge in Dokument als Ergebnisdokumentation (Grundlage für Blogpost)	10
10. Abschluss	Feedbackblock und Ausblick	Plenum	Beamer-Präsentation	9
				240

Lernziele

Die Teilnehmenden erkennen die Komplexität des *peer reviewing* im DHd-Kontext.

Die Teilnehmenden reflektieren persönliche Reviewerfahrungen und formulieren konstruktive Ideen und Lösungsvorschläge zur Verbesserung des Reviewverfahrens.

Die Teilnehmenden erhalten Informationen über das Vorgehen der Task Force, diskutieren ihre Vorschläge und geben konstruktives Feedback.

Die Teilnehmenden üben das konstruktive Formulieren von Feedback und erhalten bewährte Formulierungen als Anregung für zukünftige Reviews.

Kontaktaten aller Beitragenden

Svenja Guhr (svenja.guhr@tu-darmstadt.de) ist wissenschaftliche Mitarbeiterin am Institut für Sprach- und Literaturwissenschaft der TU Darmstadt. Sie lehrt und forscht in der computationalen Literaturwissenschaft mit Forschungsinteresse in automatisierter Analyse von Prosatexten, Lautstärke als narratologisches Phänomen und Gender Studies.

Timo Steyer (t.steyer@tu-braunschweig.de) leitet das Referat Informationskompetenz an der Universitätsbibliothek der Technischen Universität Braunschweig. Er ist seit vielen Jahren in den Digital Humanities aktiv und ist Convenor der DHd-AG Digitales Publizieren. Seine Forschungsinteressen sind digitales Publizieren, (Meta)datenmodellierung und Data Literacy.

Walter Scholger (walter.scholger@uni-graz.at) ist Institutsmanager am Zentrum für Informationsmodellierung der Universität Graz, Sprecher des CLARIAH-AT Konsortiums und Co-Lead der DARIAH-EU Arbeitsgruppe zu ethischen und rechtlichen Aspekten der Digital Humanities (ELDAH) und bearbeitet die Themenfelder Open Science sowie Aspekte der digitalen Veröffentlichung und Nachnutzung von Forschungsdaten aus den Bereichen Wissenschaft und Kulturerbe.

Angaben zum Zielpublikum

Zahl der möglichen Teilnehmer:innen: < 30 Personen

Das Zielpublikum umfasst alle, die sich für den Begutachtungsprozess der Tagungen des DHd-Verbandes und ähnlicher Veranstaltungen interessieren. Es werden keine Vorkenntnisse benötigt. Angesprochen sind vor allem junge Wissenschaftler:innen, die planen erste Reviewanfragen anzunehmen und den Workshop als Orientierung nutzen wollen, um sich mit erfahrenen Reviewer:innen auszutauschen. Für den Erfahrungsaustausch freuen wir uns über die Teilnahme erfahrener Reviewer:innen, die *best-practice*-Erfahrungen und erfolgreiche Routinen mit den Teilnehmenden sowie den Beitragenden teilen möchten. Zusätzlich wollen wir Teilnehmende ansprechen, die den Workshop nutzen möchten, um ihre eigene Reviewpraxis zu reflektieren und aktiv an der Qualität ihrer Reviews zu arbeiten.

Benötigte technische Ausstattung

Benötigt werden ein Beamer für die Präsentation in der Einleitungsphase sowie WLAN im Seminarraum mit Zugang für die Teilnehmer:innen und die Workshopleitung. Die Teilnehmer:innen bringen ihren eigenen internetfähigen Laptop mit.

Bibliographie

Burghardt, Manuel / Dieckmann, Lisa / Guhr, Svenja / Reiter, Nils / Scholger, Walter / Steyer, Timo / Trilcke, Peer / Wuttke, Ulrike (2021): *Handreichung für den Begutachtungsprozess der DHd2022*. Zenodo doi:10.5281/zenodo.5093652.

Burghardt, Manuel / Dieckmann, Lisa / Reiter, Nils / Steyer, Timo / Scholger, Walter / Trilcke, Peer / Wuttke, Ulrike (2021): *Besseres Reviewing für die DHd (Version 1.0)*. Zenodo doi:10.5281/zenodo.4633633.

DFG (2010): *Hinweise zu Fragen der Befangenheit, DFG-Vordruck 10.201*, https://www.dfg.de/formulare/10_201/ [letzter Zugriff 14. Juli 2021].

Omer, Ahmad / Abdularhim, Mohhamed (2017): "The criteria of constructive feedback: The feedback that counts", in: *Journal of Health Specialties* 5(1): 45 <https://link.gale.com/apps/doc/A479274547/HRCA?u=anon~5c61e832&sid=bookmark-HRCA&xid=647edcaf> [letzter Zugriff 14. Juli 2021].

Ross-Hellauer, Tony (2017): "What is open peer review? A systematic review", in: *F1000Research* 6: 588 doi:10.12688/f1000research.11369.2.

Ross-Hellauer, Tony / Görögh, Edit (2019): "Guidelines for open peer review implementation", in: *Res Integr Peer Rev* 4(4) doi:10.1186/s41073-019-0063-9.

Parser bauen für domänenspezifische Notationen

Arnold, Eckhart

arnold@badw.de

Bayerische Akademie der Wissenschaften, Germany

Programmierworkshop: 4h

Benötigte Vorkenntnisse: Python und reguläre Ausdrücke

Domänenspezifische Notationen (DSL) sind auf ein jeweils bestimmtes Anwendungsfeld zugeschnitten und ermöglichen deshalb oft eine schnellere Dateneingabe und übersichtlichere Quelltexte als generalisierte Auszeichnungssprachen. Sie bilden deshalb in vielen Bereichen eine sinnvolle Ergänzung oder sogar Alternative zu XML. Beispiele dafür liefern: (Tinney 2014)(von Stockhausen 2020)(Arnold 2019).

Trotz dieser Beispiele werden DSLs in den Digital Humanities (DH) noch eher selten eingesetzt. Dies kann damit zusammenhängen, dass die Kenntnis der Technologien zum Bau von DSL in diesem Bereich noch wenig verbreitet ist. Ein Ziel des Workshops ist es genau diese Kenntnisse zu vermitteln. Mögliche Einsatzgebiete von DSLs gibt es in vielen Bereichen:

1. Bei der Eingabe bzw. Kodierung von Daten, wo DSLs sich als eine bequemere und übersichtlichere Alternative zu XML anbieten, insbesondere in Fällen, wo XML-Editoren die Eingabe nur begrenzt einfacher machen oder man nicht auf proprietäre Produkte wie Oxygen zurückgreifen möchte.
2. Bei der Daten-Extraktion. Da sich mit EBNF nicht nur reguläre sondern auch kontext-freie Grammatiken spezifizieren lassen, ist mit diesem Ansatz mehr möglich als mit regulären Ausdrücken. Dazu gehört insbesondere auch die Konvertierung von Alt-Daten, die in einem textbasierten Format kodiert sind, wie z.B. LaTeX-Manuskripte, nach XML, um sie für die Weiterverarbeitung mit dem Computer vorzubereiten.
3. Schließlich gibt es noch technischere Einsatzgebiete, die für die DH indirekt relevant werden können, wie z.B. die Übertragung von Datenstrukturen aus einer Programmiersprachenwelt in eine andere, etwa TypeScript-Interfaces zu Python TypedDicts: <https://ts2python.readthedocs.io>

Allerdings erfordert der Einsatz von DSLs die Programmierung von sog. „Parsern“, die Texte in der DSL in generische Datenbeschreibungssprachen wie XML oder SQL oder direkt in bestimmte Datenstrukturen übersetzen. Während es für kleinere Projekte noch genügt, die Grammatik einer DSL informell zu beschreiben und den Parser aus regulären Ausdrücken zusammenzusetzen, lassen sich größere Projekte oder DSLs, die sich mit der Zeit weiterentwickeln, ohne eine formale Spezifikation der Grammatik einer DSL und automatisierte Tests, die vor Fehlern bei späteren Ergänzungen der DSL schützen, kaum noch realisieren. Üblicherweise werden dafür Parser-Generatoren verwendet, die die Strukturdefinition bzw. Grammatik der DSL in einen ausführbaren Parser übersetzen. Die Grammatik wird dabei mit der Beschreibungssprache EBNF spezifiziert. (Eine EBNF-Spezifikation ist für eine DSL ungefähr das, was eine DTD für XML darstellt.)

Natürlich gibt es auch andere Möglichkeiten, Parser zu bauen, z.B. handgeschriebene Adhoc-Parser oder gestaffelte reguläre Ausdrücke, wie sie etwa von „TextMate-Grammatiken“ für die

farbliche syntaktische Hervorhebung von Text-Editoren verwendet werden. Für den Einsatz von Parser-Generatoren und EBNF sprechen diese Gründe:

1. Der EBNF-Formalismus ist seit einigen Jahrzehnten ein stabiler und vielfach genutzter Standard für die Spezifikation formaler Sprachen. Spezifiziert man die Grammatik einer DSL in EBNF, so kann man im Zweifelsfall mit relativ geringem Aufwand auf einen anderen Parser-Generator und eine andere Programmiersprache umziehen. Handgeschriebene Parser sind relativ stärker mit der einmal gewählten Technologie vermählt. Das allein ist ein guter Grund für den Einsatz von EBNF.
2. EBNF erlaubt eine absolut präzise und zugleich konzise Spezifikation der Grammatik einer formalen Sprache. Als Ergänzung zu einer Prosa-Beschreibung mit Beispielen, kann man damit die Regeln einer domänenspezifischen Notation missverständnisfrei kommunizieren.
3. Ab einer gewissen Komplexität erscheint mir der Einsatz eines Parser-Generators, den man mit der formalen Spezifikation (in EBNF) der eigenen DSL füttern kann, bequemer als andere Ansätze, wie etwa gestaffelte reguläre Ausdrücke oder vollständig handgeschriebene Parser. Auch, wenn eine DSL im Laufe der Zeit abgeändert bzw. ergänzt werden soll, zahlt sich eine explizit spezifizierte Grammatik (im Verein mit einer umfassenden Test-Suite) aus - so zumindest meine Erfahrung.

Dieser Workshop ist als Lehrveranstaltung auf 4 Stunden angelegt und vermittelt einen Einstieg in den Bau von Parsern für DSLs. Vermittelt werden:

1. Die formale Spezifikation von Grammatiken für DSLs in der Erweiterten Backus-Naur-Form (EBNF): EBNF ist so etwas wie Reguläre Ausdrücke auf Speed. Während reguläre Ausdrücke wie der Name sagt, nur die sog. „Regulären Sprachen“ verarbeiten können, können mit EBNF Grammatiken für vergleichsweise sehr viel ausdrucksreichere „kontextfreie Sprachen“ festgelegt werden. Insbesondere wird es dadurch sehr viel leichter, verschachtelte Strukturen zuzulassen.
2. Der Bau eines auf dieser Grammatik beruhenden Parsers für die DSL mit Hilfe eines Parser-Generators. Parser übersetzen DSL-Texte in „konkrete Syntaxbäume“. Konkrete Syntaxbäume enthalten in der Regel jedoch noch viele Spuren des Übersetzungsprozesses, die für die weitere Verarbeitung nicht mehr relevant sind. Die Kunst besteht darin, diese konkreten Syntaxbäume zu möglichst schlanken „abstrakten Syntaxbäumen“ zu vereinfachen. Während der Parser selbst automatisch generiert werden kann, hängt die Generierung des abstrakten Syntaxbaums von der Zieldomäne ab und muss daher von Hand festgelegt werden.
3. Die Extraktion von Daten aus Syntaxbäumen bzw. die Umformung von Syntaxbäumen in vorgegebene Zieldatenstrukturen. Auch die abstrakten Syntaxbäume entsprechen in der Regel noch nicht den Zieldatenstrukturen, sondern müssen noch einmal umgewandelt werden. Sollen die Zieldaten in XML vorliegen, so genügt eine einfache Serialisierung. Komplizierter wird es, wenn die Zieldaten gar keine Baumstrukturen haben, sondern z.B. Graphen sind.
4. Der Einsatz von Einheiten-Tests zum schrittweisen Aufbau und der kontinuierlichen Prüfung von Grammatiken. Insbesondere für spätere Änderungen und Ergänzungen einer DSLs sind Einheiten-Tests nahezu unverzichtbar, will man

die Rückwärtskompatibilität der sich weiterentwickelnden DSL sicher stellen.

Als Parser-Generator verwenden wir das Python-Rahmenwerk DHParse. Wer möchte kann sich dort im Vorfeld die „Schritt für Schritt“-Anleitung dazu durchlesen: <https://t1p.de/0l6l>.

Wohlbemerkt: In dem Kurs geht es um den Bau von Parsern mit Hilfe eines Parser-Generators, was etwas, aber nicht viel schwieriger ist als das Erlernen der Technologie der Regulären Ausdrücke. Es geht also um die Nutzung und nicht um die Entwicklung von Parser-Generatoren (auch wenn einige der Einträge in der Bibliographie sich damit beschäftigen), was ein sehr viel komplizierteres Thema ist. Wen das interessiert, dem lege ich als Einstieg die Blog-Serie von Guido van Rossum dazu ans Herz: t1p.de/y2ko

Bibliographie

Arnold, Eckhart (2021): DHParse. Toolchain for Domain Specific Languages in the Digital Humanities, gitlab.lrz.de/badw-it/DHParse.

Arnold, Eckhart (2019): Dokumentation der Notation für Artikel des Mittellateinischen Wörterbuchs, t1p.de/44x9.

Blaudeau, Clement / Shankar, Natarajan (2020): A verified packrat parser interpreter for parsing expression grammars, CPP 2020: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs January 2020 Pages 3–17, DOI: 10.1145/3372885.3373836, t1p.de/otay.

Ford, Brian (2004): Parsing Expression Grammars: A Recognition-Based Syntactic Foundation, Cambridge Massachusetts.

Mascarenhas, Fabio / Medeiros, Sérgio / Ierusalimsky, Roberto (2014): On the relation between context-free grammars and parsing expression grammars, Science of Computer Programming, Volume 89, Part C, Pages 235-250, t1p.de/v5uz.

van Rossum, Guido (2019): PEG Parsing, t1p.de/y2ko.

von Stockhausen, Annette (2020): Domain Specific Language (DSL) für Transkriptionen, t1p.de/r8yv.

Tinney, Steve et al. (2014): ORACC. Open Richly Annotated Cuneiform Corpus, t1p.de/uryr.

Voelter, Markus et al. (2013): DSL-Engineering. Designing, Implementing and Using Domain-Specific Languages, Stuttgart, t1p.de/wik.

Peer-to-Peer-Workshop zum Projekt Management in den Digital Humanities

Cremer, Fabian

Cremer@ieg-mainz.de

Leibniz-Institut für Europäische Geschichte

Dogunke, Swantje

swantje.dogunke@gmail.com

Thüringer Universitäts- und Landesbibliothek, Germany

Neubert, Anna

aneubert@uni-bielefeld.de
Universität Bielefeld

Wübbena, Thorsten

Wuebbena@ieg-mainz.de
Leibniz-Institut für Europäische Geschichte

Kontext und Einführung

Interdisziplinäre Wissensproduktion innerhalb der digitalen geistes- und kulturwissenschaftlichen Forschung benötigt verschiedene Planungs-, Koordinierungs- und Steuerungselemente, um eine erfolgreiche Zusammenarbeit zu ermöglichen. Dabei müssen unterschiedliche Aufgaben koordiniert, Wissen, Herausforderungen und Fortschritte kommuniziert und eine übergreifende Forschungsvision akzeptiert und umgesetzt werden, damit ein Projekt gelingt (Neubert 2020). Das Zusammenführen verschiedener Stakeholder, wie die Mitarbeitenden in Archiven, Bibliotheken, Infrastruktureinrichtungen, Museen oder Stiftungen mit Forschenden in Geisteswissenschaften und Informatik erfordert daher grundlegendes Wissen über Praktiken und Gegenstände und setzt Kenntnisse über die jeweiligen Vorgehensweisen voraus. Unterschiedliche Konzepte, Werkzeuge sowie ein Verständnis von Projektmanagement sind hier unabdingbare Voraussetzungen, um erfolgreich zusammenzuarbeiten und zu interessanten sowie innovativen Ergebnissen zu gelangen (Komprecht und Rösenstrunk 2016).

Gerade wenn es also auch um die „Kulturen des digitalen Gedächtnisses“ geht, trägt die sorgfältige Planung und Koordination von Projekten dazu bei, besser mit Möglichkeiten und Konsequenzen der digitalen Gedächtnisarbeit umzugehen. Dazu gehören neben effektivem Zeit- und Beziehungsmanagement auch die theoretische wie praktische Auseinandersetzung im Umgang mit verschiedenen Gruppen, Zeitschienen und Kulturen. Welche Chancen ergeben sich für die digitale geisteswissenschaftliche Gedächtnisarbeit durch gutes Projektmanagement? Welche Planungs- und Koordinierungselemente sind besonders wichtig und können in (digitalen) interdisziplinären Teams hilfreich sein? Wie können Potenziale erkannt und Risiken in solchen Projekten minimiert werden?

Neben den methodischen Fragestellungen beeinflussen auch die individuellen, strukturellen und organisatorischen Rahmenbedingungen die Projektarbeit in den DH. Welche Fähigkeiten und Kompetenzen sind für das Projektmanagement in den DH entscheidend und wie werden diese erlernt oder erfahren? Welche Rollen und welches Rollenverständnis bilden Grundlage für die Zusammenarbeit? Wie lässt im Wissenschaftssystem aus Aufgaben im Projektmanagement Anerkennung und Reputation gewinnen? In unserem Peer-to-Peer-Workshop wollen wir im Rahmen eines World Cafés diesen und anderen Fragen nachgehen und mit allen interessierten Forscher:innen und Projektmanager:innen diskutieren und dabei einen Raum schaffen, sich über methodische Fragen und persönliche Erfahrungen mit Projektplanung, Projektmanagement und Koordinierungsaufgaben in der digitalen interdisziplinären Wissensproduktion auszutauschen.

Bedarf und Genese

Die Vermittlung von Fähigkeiten und Methoden im Projektmanagement ist bisher kein weit verbreiteter Teil der Curricula in den DH-Studiengängen - Ausnahmen sind hier DH-Studiengänge in München oder Graz -, oder Gegenstand der verfügbaren Handbücher im deutschsprachigen Raum (vgl. Cremer 2019). Während in englischsprachigen DH Summer Schools regelmäßig Einführungen in das Projektmanagement stattfinden, steht dies in der DHd-Community noch aus. Jenseits dieses Desiderats fehlt es darüber hinaus auch an Möglichkeiten eines grundsätzlichen Wissensaustauschs über die Forschungspraxis und den Einsatz von Projektmanagementmethoden in den DH im deutschsprachigen Raum. [1] Es existiert eine Vielzahl an Digital Humanists, die sich die Fähigkeiten oft durch die „school of hard learned experience“ (Siemens 2016) aneignen mussten. Die Personen, die häufig als „Intermediaries“ (Edmond 2005) in DH-Projekten das Projektmanagement übernehmen (teilweise ohne Auftrag), finden jedoch nur wenig Austauschmöglichkeiten, weil sie im Projektkontext und auch in den Organisationen oft die einzigen Personen mit dieser Rolle und damit „strukturell isoliert“ sind (Cremer et al. 2018). Ein Konzeptionstreffen zu einer Interviewreihe zu Projektmanagement im Rahmen den vDHd2021 (Cremer et al. 2021) hat eine Vielzahl der Akteur:innen zusammengebracht, die hierbei den Bedarf nach einem Austauschformat und einer Netzwerkbildung festgestellt haben. Weiter wurden hier mehrere Themenkomplexe identifiziert, die sich für einen Erfahrungsaustausch und inhaltliche Auseinandersetzung eignen.

Ziele und Umsetzung

Der hier vorgeschlagene Workshop bietet den Teilgebenden Möglichkeiten der Auseinandersetzung und Weiterentwicklung der spezifischen Themen im Bereich des DH-Projektmanagements. Der hierfür notwendige Grad an Partizipation wird bereits im Vorfeld durch Beteiligung an der Programmgestaltung im Rahmen eines (*Call for Ideas*) ermöglicht. Die konkrete Bearbeitung der aufgeworfenen Fragen und Ideen erfolgt gemeinsam in der Gruppe in einem Peer-to-peer-Format im Rahmen eines World Café (Brown und Isaacs 2001). Die Prinzipien der Partizipation und Multiperspektivität sowie der gemeinsamen Ergebnis-sicherung und Dokumentation im World Café ermöglichen Wissensaustausch und Erkenntnisgewinn. Das World Café ist in der DHd-Community bereits als erfolgreiches Austausch- und Entwicklungsformat etabliert (vgl. Roeder et al. 2020 sowie Geiger und Pfeiffer 2020).

Zugleich wird mit der Durchführung und wissenschaftlichen Begleitung dieser Veranstaltung die Sensibilisierung für Projektmanagement in den DH erhöht. Ähnliche Effekte lassen sich bereits durch vorhergehende Veranstaltungen der DHd-Tagungen beobachten, wo insbesondere auch Studierende mit diesem Themenkomplex erstmals in Berührung gekommen sind. Der hier vorgeschlagene Workshop knüpft an diese Entwicklungen in der DHd-Community an, dazu gehören die Erfahrungen im Projektmanagement als Teil der Koordination von DH-Aktivitäten (Cremer et al. 2019 und Roeder et al. 2020) sowie die Interviewreihe zu Projektmanagement in den Digital Humanities auf der vDHd 2021 (Cremer et al. 2021). Die Dokumentation der an den Tischen entwickelten Ideen und die spätere Aufbereitung durch die Workshop-Veranstaltenden in Form von sowohl Blog- als auch wissenschaftlichen Beiträgen können eine hilfreiche Grundlage für

eine methodische Weiterentwicklung und organisatorische Strukturbildung sein.

Der strukturierte „Peer-to-Peer“-Austausch bietet neben dem Wissensaustausch und konzeptionellen Lösungsentwicklungen auch die Möglichkeit, weitergehende individuelle Formate (Bildung einer eigenen Peer Group etwa über eine Mailing-Liste) oder institutionelle Strukturen (Bildung eines Netzwerkes oder Arbeitsgruppe innerhalb des DHd-Verbandes) zu entwickeln, die über das sporadische und unregelmäßige Zusammenkommen hinausgehen. Hier soll der Workshop ergebnisoffen sowohl Impulse zur Entwicklung von Ideen geben, gemeinsame Interessen identifizieren und auch Raum für die Bildung der individuellen Netzwerke geben. Die Verortung dieser Teil-Community und die Gestaltung der Kommunikationsstrukturen kann sich im Rahmen des Workshops herauskristallisieren.

Themen und Ideenentwicklung

Über einen *Call for Ideas* möchten wir die Fragestellungen und Konzepte aus der Community bezüglich des weitgefächerten Themenkomplexes abfragen. Die eingegangenen Vorschläge werden von den Workshop-Veranstaltenden ausgewertet, ggf. zusammengefasst und ausgewählt, so dass in der Überarbeitungsphase dieses Beitrages die Themen bereits festgelegt sind. Der offene Call stellt sicher, dass die diskutierten Themen den Bedürfnissen der Community der Projektmanager:innen in den DH entsprechen. Die Rolle der Workshopleitung wird darin liegen, die eingereichten Themen zu strukturieren, die Moderation der Veranstaltung zu übernehmen und die Ergebnissicherung zu begleiten. Die folgenden Themen wurden aus der Community in vorhergehenden Veranstaltungen erarbeitet und geben einen Eindruck in das zu erwartende Programm ohne das diese Themen als gesetzt oder notwendig anzusehen sind:

- Managementkultur in den Geisteswissenschaften: Gibt es in den Geisteswissenschaften eine Management-Tradition nur unter anderem Namen? Was hat sich mit der Digitalen Transformation verändert und inwiefern bedürfen die DH ein spezifisches Projektmanagement? Werden Selbstbestimmung und Steuerung bzw. Selbstorganisation und Teamstrukturen als gegensätzliche Kulturen wahrgenommen?
- Projektmanagement gelernt und gelehrt: Welche Kompetenzen sind für ein Projektmanagement in den DH erforderlich? Wo sollte Projektmanagement in der Lehre verortet werden? Wie und wo haben die Teilgebenden die theoretische Ebene reflektiert und die praktische Ebene erfahren?
- Operationalisierung von Projektmanagement: Wie unterscheiden wir zwischen operativer, administrativer und strategischer Managementebene? Benötigt es dezidierte Stellen oder ist Projektmanagement eine (geteilte) Aufgabe für alle? Wie lässt sich Projektmanagement als eine „Teilaufgabe“ im Projekt operationalisieren und in den Einrichtungen institutionalisieren?
- Projektmanagement als Teil der Wissenschaft: Welche Managementkonzepte lassen sich als wissenschaftliches Arbeiten charakterisieren? Wie lässt im Wissenschaftssystem Anerkennung und Reputation gewinnen? Welche Formate und Bedingungen erlauben eine theoretische Reflexion der vorhandenen Managementtätigkeiten und Strukturen.

Organisation und Ablauf

Roadmap

- November 2021: Call for Ideas
- Dezember 2021: Redaktion der Themen
- Januar 2022: Endfassung des Abstracts mit allen Themen und Beitragenden
- März 2022: World Café auf der DHd 2022
- Juni 2022: Veröffentlichung der Dokumentation

Ablauf der Veranstaltung

1. Intro (15 min)
 1. Themeneinführung
 2. Erläuterung des Formats World Café
2. World Café Qualifying (25 min)
 1. Vorstellung der Thementische
 2. Gruppenbildung und informeller Austausch
3. World Café Runden (4-5 Runden) (4-5 x 20 min + je 5 min Pause)
 1. Diskussion und Dokumentation
 2. Rotation der Gruppen und Tische
4. World Café Endrunde (30 min + 15 min Pause)
 1. Präsentation der Tischplakate
 2. Zusammenfassung der Ergebnisse
5. Outro (25 min)
 1. Fortführung und Verstetigung des Austauschs
 2. Verabschiedung

Formalia

Zielpublikum: alle Interessierten, Vorerfahrung im Projektmanagement wünschenswert, aber keine Voraussetzung
 Zahl der möglichen Teilnehmerinnen und Teilnehmer: bis 30 (bis zu 5 Themen/Tische à 20 min mit je 4-7 Diskutant:innen)
 Benötigte technische Ausstattung: 5 Tischgruppen, A0-Papierbögen, Moderationskarten, dicke Filzstifte, Aufhängung für die Papierbögen (Wand oder Stellwand)

Workshopleitung

Fabian Cremer; (@fabian_cremer); Leibniz-Institut für Europäische Geschichte (IEG); cremer@ieg-mainz.de

Forschungsinteressen: Forschungsdatenmanagement, Digitale Transformation und Arbeitsteilung im Forschungsprozess, Soziale Forschungsinfrastrukturen

Swantje Dogunke; (@swagunke); Thüringer Universitäts- und Landesbibliothek (ThULB) Jena; swantje.dogunke@uni-jena.de

Forschungsinteressen: Partizipativer Aufbau von Infrastruktur und Services in den DH, digitale Editionen, Modellierung
 Anna Maria Neubert; (@annamneubert); Universität Bielefeld; aneubert@uni-bielefeld.de

Forschungsinteressen: Digitalisierung der Wissenschaften, Projektmanagement, Forschungsförderung, Wissenschaftspolitik

Thorsten Wübbena; (@ThWuebbena); Leibniz-Institut für Europäische Geschichte (IEG); wuebbena@ieg-mainz.de

Forschungsinteressen: Digitale Kunstgeschichte, Wissensrepräsentation, Forschungsinfrastrukturen, Offene Forschungsdaten

Bibliographie

Cremer, Fabian. 2019. „Gottes Werk und Teufels Beitrag: Ein Essay zu Digital Humanities und Projektmanagement“. Blog. *DHd-Blog* (blog). 19. März 2019. <https://dhd-blog.org/?p=11283>.

Cremer, Fabian, Swantje Dogunke, und Thorsten Wübbena. 2021. „Unfrequently Asked Questions: Interviewreihe zu Projektmanagement in den Digital Humanities“. Gehalten auf der *vdHd 2021 – Experimente*, Januar 28. <https://vdhd2021.hypotheses.org/187>.

Edmond, Jennifer. 2005. „The Role of the Professional Intermediary in Expanding the Humanities Computing Base“. *Literary and Linguistic Computing* 20 (3): 367–80. <https://doi.org/10.1093/lc/fqi036>.

Geiger, Jonathan, und Jasmin Pfeiffer. 2020. „Spielplätze der Theoriebildung in den Digital Humanities“. Gehalten auf der *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. 7. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2020), Paderborn, Februar 20. <https://doi.org/10.5281/zenodo.4621980>.

Komprecht, Anna Maria, und Daniel Röwenstrunk. 2016. „Projektmanagement in digitalen Forschungsprojekten. Ein Leitfaden für interdisziplinäre und kooperative Drittmittelprojekte im Umfeld Digitaler Editionen“. In *„Ei, dem alten Herrn zoll' ich Achtung gern“*. *Festschrift für Joachim Veit zum 60. Geburtstag*. <https://pub.uni-bielefeld.de/record/2912214>.

Neubert, Anna Maria. 2020. „Navigating Disciplinary Differences in (Digital) Research Projects Through Project Management“. In *Digital Methods in the Humanities*, herausgegeben von Silke Schwandt, 59–86. <https://doi.org/10.14361/9783839454190-003>.

Roeder, Torsten, Fabian Cremer, Swantje Dogunke, Frederik Elwert, Harald Lordick, Katrin Ott, Sibylle Söring, und Thorsten Wübbena. 2020. „Digital Humanities from Scratch 2020“. Gehalten auf der *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. 7. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2020), Paderborn, März 2. <https://doi.org/10.5281/zenodo.4621848>.

Siemens, Lynne. 2016. „Project Management and the digital humanist“. In *Doing digital humanities: practice, training, research*. New York.

Steyer, Timo, Fabian Cremer, Swantje Dogunke, Corinna Mayer, Katrin Neumann, und Thorsten Wübbena. 2018. „Peer-To-Peer statt Client-Server: Der Mehrwert Kollegialer Beratung und agiler DH-Treffen“. In *DHd2018: Kritik der digitalen Vernunft*. Köln. <https://doi.org/10.5281/zenodo.1186594>

Projektmanagement für die Digital Humanities

Frank, Markus

Markus.Frank@itg.uni-muenchen.de

Ludwig-Maximilians-Universität München, Germany

Themenbeschreibung

Einleitung

Unter 30% aller IT-Projekte werden erfolgreich abgeschlossen, mehr als 50% enden mit geringem Erfolg, und etwa 20% scheitern vollständig. Zu diesem Ergebnis kommt der viel beachtete CHAOS Report 2015, der in den Jahren 2011 bis 2015 über 25.000 Software-Projekte auf ihren Erfolg bzw. Misserfolg hin untersucht hat (siehe Standish Group/IKTM 2015). Erfolg wird im Report daran bemessen, ob ein Projekt den versprochenen Funktionsumfang liefert, sich im Budgetrahmen bewegt, und ob die geplante Bearbeitungszeit eingehalten wurde. Erfolgreiche IT-Projekte erfüllen oder übererfüllen alle Anforderungen, wenig erfolgreiche IT-Projekte liefern unvollständige Funktionalität oder benötigen zusätzliches Geld bzw. Zeit, und gescheiterte Projekte liefern keine Funktionalität ab. Es ist davon auszugehen, dass der Report keine IT-Projekte aus dem Bereich der Digital Humanities beinhaltet¹, und dennoch dürften die Gründe, die häufig für mangelnden Projekterfolg herangeführt werden, aus vielen DH-Projekten bekannt sein (vgl. Tiemeyer 2018: 2): IT-Projekte scheitern häufig an „weichen Faktoren“. Kommunikation zwischen den Projektteilen ist der Schlüsselfaktor für Erfolg, ebenso wie Transparenz und der professionelle Umgang mit Risiken, Unsicherheiten und Spannungen. Gerade das Verhalten der Projektleitung wirkt sich wesentlich auf den Erfolg aus, da sie die Rahmenbedingungen für erfolgreiche Projektabwicklung schaffen muss, wozu die IT-Strategie des Projektes, klare Rollen aber auch eine angemessene Ressourcenausstattung zählt. Je größer Projekte werden, je komplexer und interdisziplinärer, umso wichtiger wird die Wahl angemessener Arbeitsorganisationsformen und von Planungsstrukturen, die über die Erstellung eines Arbeitsplanes in der Projektantragsphase hinausgehen. Was den Erfolg eines Projektes behindert, sind in vielen Fällen nicht die „harten technischen Leistungsfaktoren“, sondern ein fehlendes Projektmanagement, welches die Rahmenbedingungen für den Projekterfolg schafft (vgl. Tiemeyer 2018: 2).

Aktuelle Situation in den Digital Humanities

Professionelles Projektmanagement spielt in den DH bis zum heutigen Zeitpunkt quasi keine Rolle, weder in den Projekten selbst, noch in der Ausbildung (siehe hierzu Cremer 2019). Standard in der Ausbildung von Geisteswissenschaftlern ist auch heute noch die Einzelarbeit (vgl. Siemens 2009: 225) und nicht die Kolaboration in inter- oder multidisziplinären Teams: „[...] humanists do not necessarily come to DH with the necessary skills and mindset for collaboration“ (Siemens 2016: 344). Obwohl die Thematik langsam an Bedeutung zu gewinnen scheint², taucht sie in den Anträgen auf Fördermittel bislang allenfalls peripher auf, es existieren kaum explizite Projektmanagement-Positionen in DH-Projekten³. Dedizierte Weiterbildungsangebote für das Projektmanagement in den DH oder ein Eingang dieser Themenbereiche in die Curricula der DH-Studiengänge sind bis zum heutigen Tag so selten, dass sich an Sharon Leons Feststellung aus dem Jahr 2011 über den Mangel an Absolventen und Nachwuchswissenschaftlern mit Vorbereitung auf die Zusammenarbeit in interdisziplinären Projekten bis heute kaum etwas geändert haben dürfte (siehe Leon 2011). Tatsächliche neue Arbeitsmodelle, die speziell für die DH entworfen wurden, sind ebenfalls kaum auffindbar (siehe als Ausnahmebeispiel Tabak 2017).

Der Bedarf

Grundsätzlich gilt: Je größer und komplexer ein IT-Projekt wird, je mehr Disziplinen zusammenarbeiten, umso größer wird die Wahrscheinlichkeit von Misserfolg (vgl. Standish Group/IKTM 2015: 3) und umso stärker steigt der Bedarf an Planung und Ablaufsteuerung. Werden die Projekte so groß, dass sie aus mehreren Teilprojekten bestehen, treten immer größere Koordinationsprobleme auf, die bis zum Auseinanderbrechen bzw. Abbruch der Projekte führen können (siehe hierzu ausführlich DFG 2008). Was also kann modernes Projektmanagement in einem Forschungsprojekt der DH bewirken? Projektmanagement dient dazu, das Ineinandergreifen von Ressourcen sowie die Abläufe im Projekt zu optimieren und Engpässe zu vermeiden. Es soll produktive Kommunikation über fachliche Domänen hinweg sicherstellen, Meilensteine definieren und über deren Einhaltung wachen. Es soll Konfliktpotentiale im Team reduzieren und bei der Konfliktlösung helfen. Zusammengefasst soll es dazu beitragen, dass Projekte erfolgreich abgeschlossen werden, die Ziele erfüllt oder übererfüllt werden, und dass die Arbeitskraft der Mitarbeiter geschont wird⁴. Gerade im Wettbewerb um Forschungsgelder kann ein professionelles Projektmanagement einen klaren Vorteil bei der Zielerreichung bedeuten, im Gegensatz zu Projekten ohne entsprechendes Management. Woran können sich die DH orientieren, wenn sie nach Management-Ansätzen für ihre Projekte sucht? Obwohl DH-Projekte einige Besonderheiten aufweisen, welche durch die Zusammenarbeit von Geisteswissenschaftlern und IT-Spezialisten entstehen, handelt es sich am Ende doch meistens um IT-Projekte, mit der zentralen Herausforderung, die Kommunikation zwischen den fachwissenschaftlichen Elementen und den IT-Elementen zu überbrücken. Kommunikation und interdisziplinäre Konflikte spielen eine bedeutende Rolle, im Kern benötigen die komplexen Softwareprojekte aber primär eine umfangreiche Planung und zugleich Organisationsformen, die flexibel auf Veränderungen reagieren können. Klassische sequenzielle Vorgehensmodelle (siehe Timinger 2017: 38-41) sind für IT-Projekte oft nicht zielführend, es braucht vielmehr iterative und inkrementelle Vorgehensmodelle (siehe Timinger 2017: 43-46) oder agile Methoden (siehe Timinger 2017: 161-240), die mit hoher Geschwindigkeit auf sich verändernde Rahmenbedingungen und neue methodische und fachwissenschaftliche Erkenntnisse reagieren können.

Der Workshop

Der Workshop soll an diesem Bedarf ansetzen, und soll eine erste Einführung in zentrale Bereiche des Projektmanagement im IT-Sektor geben, die an den Bedürfnissen der DH ausgerichtet sind. Dabei liegt ein besonderes Augenmerk darauf, den Teilnehmern die Möglichkeit zu geben, auf Basis ihrer eigenen Projekterfahrungen die Methoden zu reflektieren und exemplarisch anzuwenden.

Ein erster Teil des Workshops befasst sich mit der Projektplanung (siehe hierzu Tiemeyer 2018, Timinger 2017, Jakoby 2015): Es lässt sich wohl generell konstatieren, dass in DH-Forschungsprojekten zu wenig geplant wird, da zumeist eine Orientierung am groben Arbeitskonzept erfolgt, welches im Projektantrag skizziert wurde. Im Workshop werden Werkzeuge präsentiert, die aufbauend auf dem Projektantrag zur detaillierteren Planung komplexer Projekte eingesetzt werden können. Wie formuliert man Arbeitspakete, welche Abhängigkeiten zwischen Kernelementen existieren, wie evaluiert man Durchlaufzeiten und stellt Engpässe fest? Wie vermeidet man Zielantinomien?

Der zweite Teil befasst sich mit klassischem Projektmanagement: Welche Vorgehensmodelle existieren, und welche sind für DH-Projekte sinnvoll? Welche Instrumente hält das klassische Projektmanagement bereit, um die Arbeit zu steuern (Ist-Soll-Vergleiche etc.)?

Der dritte Teil führt in den Bereich des agilen Projektmanagements ein: Agilität ist ein zentrales Paradigma modernen F&E-Arbeit im Software-Segment (siehe Beck et al. 2001). Gezeigt werden grundlegende Konzepte von Scrum und Kanban (siehe z.B. Schwaber & Sutherland 2020). Welche Elemente daraus sind einsetzbar im Projektalltag und unter den Determinanten universitärer Organisationsverfassung?

Im letzten Teil des Workshops wird es um „Projektkatastrophen“ gehen: Wie kommt es zu sogenannten Death-March-Projekten (siehe Yourdon 1997), und wie vermeidet man, dass sich das eigene Projekt dazu entwickelt?

Weitere Angaben

Lernziele

- Die Teilnehmenden kennen gängige Ablauf- und Planungsstrukturen von Projekten, Grundlagen klassischer Managementansätze für IT-Projekte, agile Methoden der Projektrealisierung (SCRUM, KANBAN) sowie gängige Gefahren und zentrale Determinanten der Projektrealisierung in den Digital Humanities.
- Sie können passende Vorgehensmodelle auswählen, Projektziele formulieren und definieren, Zielhierarchien erstellen und operationalisieren, Risiken erfassen und einfache Projektstrukturpläne anfertigen. Darüber hinaus können sie Vorgangsknoten-Netzpläne (DIN 69900) lesen, in einfacher Form erstellen und Ressourcenplanung vornehmen.

Ablauf & Zeitplan

Tag 1 (4 Stunden):

- Projektmanagement Einleitung (Standards, Vorgehensmodelle, Phasen, Aufwand)
- Fragen und Übungen
- PAUSE
- Klassisches Projektmanagement (Arbeitsmodelle, Klassische Werkzeuge)
- Anwendungsfälle

Tag 2 (4 Stunden):

- Agiles Projektmanagement (Scrum, Kanban, Engpasstheorie)
- Anwendungsfälle
- PAUSE
- Zeit, Geld, Qualität und der „Death March“
- Fragen und Übungen

Zielpublikum

Der Workshop richtet primär an Personen, die als Geisteswissenschaftler oder IT-Spezialisten in DH-Forschungsprojekten arbeiten und/oder mit der Koordination bzw. Leitung von DH-Forschungsprojekten betraut sind. Sie sollten einige Erfahrungen in der Projektarbeit mitbringen, da die im Workshop vermittelten

Instrumente auf konkrete Management- und Planungsfragen der Projekte angewendet werden sollen.

Technische Ausstattung

Es wird keine besondere technische Ausstattung benötigt.

Beitragende

Dr. Markus Frank

(Ludwig-Maximilians-Universität, IT-Gruppe Geisteswissenschaften, Geschwister-Scholl-Platz 1, 80539 München)

Markus Frank ist an der IT-Gruppe Geisteswissenschaften im Bereich der Konzeption und Koordination von Digital Humanities Projekten tätig, zudem betreut er den Studiengang *Digital Humanities - Sprachwissenschaften*, in dem er zu geisteswissenschaftlichen, informatischen als auch zu Projektmanagement-Themen lehrt. Seine Arbeitsschwerpunkte liegen im Bereich der Korpuslinguistik, der Data Science und dem Scientific Programming. Seit einigen Jahren beschäftigt er sich mit dem Themenbereich professionelles Projektmanagement für Forschungsprojekte in den Digital Humanities. Nach der Promotion hat er berufsbegleitend *Wissenschaftsmanagement* im Master an der Deutschen Universität für Verwaltungswissenschaften in Speyer studiert.

Fußnoten

1. Der Schwerpunkt des Reports liegt auf privatwirtschaftlichen IT-Projekten, wobei die Softwareprojekte der Klassifikation „Government“ besonders schlecht abschneiden (21% Erfolg, 55% geringer Erfolg, 24% Scheitern, siehe Standish Group/IKTM 2015: 4).
2. Siehe hierzu beispielsweise die *Interviewreihe zu Projektmanagement in den Digital Humanities* der VDHD2021 (<https://vdhd2021.hypotheses.org/187>, aufgerufen am 27.05.2021) oder den Beitrag von Rodgers et al. 2016.
3. Aus Ausnahme kann hier das Projekt Verba Alpina an der LMU München gelten, welches über eine dedizierte Projektmanagement-Position verfügt (siehe Krefeld & Lücke 2014).
4. Für eine Studie zu den Überstunden im Arbeitsalltag des akademischen Mittelbaus siehe Ambrasat (2019).

Bibliographie

- Ambrasat, Jens** (2019): "Beahlt oder unbezahlt? Überstunden im akademischen Mittelbau." In: *Forschung & Lehre* 19 (2). 152–154.
- Beck, Kent / Grenning, James / Martin, Robert** et al. (2001): *Manifesto for Agile Software Development*. <http://agilemanifesto.org/> [zuletzt aufgerufen am 27.05.2021]
- Cremer, Fabian** (2019): *Gottes Werk und Teufels Beitrag: Ein Essay zu Digital Humanities und Projektmanagement*. DH-dBlog. <https://dhd-blog.org/?p=11283> [zuletzt aufgerufen am 27.05.2021]
- DFG** (2008): *Management von Forschungsverbünden - Möglichkeiten der Professionalisierung und Unterstützung*. Wiley-VCH.
- Jakoby, Walter** (2015³): *Projektmanagement für Ingenieure*. Springer Vieweg.

Krefeld, Thomas / Lücke, Stephan (2014): *Verba Alpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit*. <https://dx.doi.org/10.5282/verba-alpina> [zuletzt aufgerufen am 27.05.2021]

Leon, Sharon (2011): *Project Management for Humanists*. <http://mediacommons.org/alt-ac/pieces/preparing-future-primary-investigators-project-management-humanists> [zuletzt aufgerufen am 27.05.2021]

Rodgers, Stephanie et al. (2016): *Project Management for the Digital Humanities*. <https://scholarblogs.emory.edu/pm4dh/> [zuletzt aufgerufen am 05.11.2021]

Schwaber, Ken / Sutherland, Jeff (2020): *The Scrum Guide*. <https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf> [zuletzt aufgerufen am 27.05.2021]

Siemens, Lynne (2009): "It's a team if you use 'reply all': An exploration of research teams in digital humanities environments". In: *Literary and Linguistic Computin* 24 (2). 225–233.

Siemens, Lynne (2016): "Project management and the digital humanist". In: Crompton, Constance / Lane, Richard / Siemens, Ray (Hrsg.): *Doing Digital Humanities. Practice, Training, Research*. Routledge. 343–357.

Standish Group/IKMT (2015): *Chaos Report 2015*. https://www.standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf [zuletzt aufgerufen am 27.05.2021]

Tabak, Edin (2017): "A Hybrid Model for Managing DH Projects". In: *Digital Humanities Quarterly* 11 (1).

Tiemeyer, Ernst (2018): *Handbuch IT-Projektmanagement*. Hanser.

Timinger, Holger (2017): *Modernes Projektmanagement*. Wiley.

Yourdon, Edward (1997): *Death March. The Complete Software Developer's Guide to Surviving "Mission Impossible" Projects*. Prentice Hall PTR.

Repräsentativität in digitalen Archiven

Dziudzia, Corinna

corinna.dziudzia@ku.de

Forschungszentrum Gotha der Universität Erfurt

Hall, Mark

mark.hall@work.room3b.eu

The Open University, United Kingdom

Die „Kulturen des digitalen Gedächtnisses“ operieren vor dem Hintergrund eines Spannungsfelds, das man auf eine Dichotomie von Kanon und Archiv zuspitzen kann. Das digitale Archiv umfasst u.a. die digitalisierten Artefakte und darin nimmt der Kanon der aus verschiedenen Gründen besonders bewahrenswerten Artefakte einen spezifischen Stellenwert ein. Sie sind Bestandteil des kulturellen Gedächtnisses (vgl. Assmann, 2010). Das digitale Archiv umfasst eine epistemische Ebene, insofern darüber, auch für eine breitere Öffentlichkeit jenseits der klassischen Bildungsinstitutionen, Wissen bereitgestellt wird. Gleichmaßen liefert das digitalisierte Archiv auch die Grundlage für historisch arbeitende Wissenschaften, was etwa zur Entwicklung und Erprobung neuer wissenschaftlicher Methoden dient (Malazita et al., 2020). Das

Archiv als digitalisiertes Gedächtnis stellt einerseits altes Wissen dar und dient andererseits auch als Grundlage, von der aus, und mit der, neues Wissen produziert werden kann.

Das digitale Archiv reduziert dabei per se die Barrieren, die den Zugriff auf die Artefakte im physischen Archiv erschweren. Zugleich fügt das digitale Archiv zusätzliche Abstraktions- und Verarbeitungsschichten zwischen den physischen Artefakten und den digitalen Repräsentationen ein. Historische Artefakte müssen zuerst digitalisiert werden, dann mit Metadaten versehen und schlussendlich auffindbar gemacht werden. Wenn auf den digitalisierten Inhalten des digitalen Archivs aufbauend gearbeitet wird, dann repräsentiert jeder dieser Schritte einen Eingriff in die Datengrundlage mit dem Potential, die Arbeitsergebnisse zu verzerren. Welche Artefakte werden im Digitalisierungsprozess priorisiert? Wer erzeugt wie die Metadaten für die digitalisierten Artefakte? Wie werden die gesuchten Artefakte im digitalen Archiv gefunden? Die Antwort auf jede dieser Fragen beeinflusst die Repräsentativität des Archivs insgesamt und die Verlässlichkeit der darauf aufbauenden Ergebnisse.

Die Frage der Repräsentativität wird insbesondere im Rahmen der Intersektionalitätsforschung aktuell stark thematisiert, nicht selten verbunden mit einer Kritik, u.a. an den Digital Humanities (vgl. Risam, 2015), wobei die zugrundeliegende Debatte zum Kanon und der Kritik daran, dass er zu 'weiß' und zu 'männlich' sei, bis in die 1960er Jahre zurückreicht. Das Bewusstsein wächst dabei bei verschiedenen Akteuren, dass Auswahlen, und insbesondere auch Digitalisierungsprozesse, Ergebnisse hegemonialer Machtstrukturen sein können und es wird mit entsprechenden Erklärungen und Selbstverpflichtungen darauf reagiert (z.B. Ziegler, 2019). Das wird etwa auch vor dem Hintergrund des postcolonial computing thematisiert (Irani et al., 2010; Chan, 2017; Kizhner et al., 2021).

Der zentrale Ansatzpunkt der Kritik am möglichen Quellenbias der Archive und Digitalisate (Hall, 2020; Inwood & Stewart, 2020; Kizhner et al., 2021) ist, dass in den meisten Fällen viel zu wenig über die Zusammensetzung und die Auswahlkriterien in den digitalen Archiven bekannt ist. Wenngleich positive Beispiele digitalisierender Institutionen existieren, die darüber Auskunft geben, was sie digitalisieren, warum und in welchen Bereichen sie wie weit fortgeschritten sind oder was sie woher kommend darüber hinaus bereitstellen, ist Transparenz generell kaum vorhanden.

Die Komplexität in dieser Frage ergibt sich nicht zuletzt daraus, dass die Interessen und Anforderungen einer Reihe von Akteuren zusammenreffen, insofern es sich bei der Digitalisierung um ein 'boundary object' handelt (Star & Griesmer, 1989). Geisteswissenschaftler_innen, die mit den digitalen Archiven arbeiten wollen, Archivar_innen, die Digitalisierungs- und Metadatenprozesse leiten und durchführen, Geldgeber_innen, die Mittel für die Digitalisierungsprozesse bereitstellen, und Informatiker_innen, die Archivierungs- und Suchsoftware entwickeln – jede dieser Gruppen hat unterschiedliche Sichtweisen darauf, was Repräsentativität für sie bedeutet, aber was fehlt, ist ein Forum, in dem ein Austausch über diese Fragen über die Gruppengrenzen hinweg möglich ist.

Zu der Problematik haben wir im März 2021 den internationalen Workshop "Digital Archive and Canon" organisiert. Dieser stieß auf reges Interesse mit Beiträgen von Forscher_innen und Archivar_innen und der daraus resultierende Austausch ermöglichte ein tieferes Verständnis für die Wünsche und Positionen der jeweils anderen Gruppe. Aus der Abschlussdiskussion ergab sich, dass großes Interesse an einem Forum für einen regelmäßigen Austausch über die Repräsentativitätsfrage besteht, insbesondere in einer Struktur, welche die Interaktion zwischen den interessierten

Gruppen ermöglicht. Der hier vorgeschlagene ganztägige Workshop für die DHd 2022 stellt einen ersten Schritt in diese Richtung dar, mit zwei geplanten Outputs: einem Positionspapier und der Gründung einer DHd-Arbeitsgruppe "Repräsentativität". Es bestehen zwar schon akteursspezifische AGs (Datenzentren, DH Theory, OCR), die AG "Repräsentativität" zielt allerdings darauf, Akteur_innen über Gruppengrenzen hinweg zusammenbringen, was unsere Vorarbeit als notwendig identifiziert hat.

Themen und Fragen

Im Rahmen des Workshops sollen eine Reihe von Themen und Fragen diskutiert werden, wobei diese an die Interessen der Teilnehmer_innen angepasst werden. Die zentralen Themen, welche den Workshop potentiell strukturieren werden, sind:

- **Repräsentativität in der Digitalisierung:** Was ist der Status quo hinsichtlich der Digitalisierung? Werden besonders kanonische Artefakte digitalisiert? Werden perspektivisch alle Artefakte digitalisiert und die gesamten Bestände aller bewahrenden Institutionen? Wer entscheidet, welche relevant sind, welche aber nicht, und auf Grundlage welcher Kriterien erfolgt dies? Wer stellt diese Artefakte zur Verfügung, wer digitalisiert sie? Nach welchen Kriterien erfolgt die Priorisierung der Reihenfolge?
- **Repräsentativität in der Annotation:** Werden die Metadaten aus dem physischen Archiv übernommen oder neu erstellt? Wenn Metadaten übernommen werden, wie wird gewährleistet, dass diese nicht historische Voreingenommenheiten widerspiegeln?
- **Koordination der Digitalisierung:** Sollten Digitalisierungsarbeiten über Länder- und Archivgrenzen hinaus koordiniert werden? Wie beeinflusst das Schwerpunktsetzung, Auswahl und Aspekte der Nachhaltigkeit (Barats et al., 2020)? Könnte die Digitalisierung kanonischer Werke zwischen Archiven aufgeteilt werden? Wie wäre das praktisch umsetzbar? Sind Erfahrungen aus dem WorldCat System hier anwendbar?
- **Digitale Systeme:** Welche Verarbeitungsschritte werden für die Bereitstellung digitaler Archive über Suchsysteme angewandt? Insofern Suchhilfen wie Topic-Modelle bereitgestellt werden, wie können Verzerrungen in diesen vermieden bzw. minimiert werden? Wie kann eine Suche über die Archivgrenzen hinweg ermöglicht werden, im Sinne der Deutschen Digitalen Bibliothek und der Europeana?
- **Finanzierung:** Wie sollen Fragen der Repräsentativität in den Geldgeberprozess eingebunden werden? Wie viel sollte explizit durch die Geldgeber vorgegeben werden, ähnlich den Anforderungen bezüglich der Lizenzierung der Digitalisate?
- **Dokumentation:** Was muss während aller Prozesse beachtet und dokumentiert werden, um die Repräsentativität eines Archivs zu verbessern? Welche Informationen braucht der/die Nutzer_in des Archivs mindestens? Welche Informationen sind für welche Nutzergruppen relevant?

Durchführung

Vor dem Workshop

Im Vorfeld des Workshops werden Teilnehmer_innen gebeten ein Statuspapier zuzusenden, in dem sie darlegen, was die jeweils

eigene Perspektive ist, was ihre Interessen an dem Thema sind, und welche Aspekte für relevant erachtet werden. Unmittelbar vor dem Workshop werden diese Papiere, sowie gegebenenfalls weitere Materialien, für alle Teilnehmer_innen zur Verfügung gestellt.

Programm Workshop

Der Workshoptag selbst wird in zwei große Blöcke geteilt.

1. Am Vormittag werden die aus den Statuspapieren identifizierten Themen innerhalb der jeweiligen Interessensgruppen (Archivar_innen/Bibliothekar_innen, Wissenschaft, Technik/IT, Förderer_innen/Geldgeber_innen, ...) in separaten Gruppen besprochen. Jede Gruppe verständigt sich über den Status quo: was ist der Ist-Stand, was ist die zukünftige Perspektive, was wäre wünschenswert. Die Ergebnisse dieser Gruppenarbeit werden gesammelt und dann im Plenum besprochen. Die sich aus dieser Diskussion herauskristallisierenden, zentralen Themen definieren dann den Inhalt des Nachmittags.

2. Im Nachmittagsblock werden die Gruppen neu zusammengestellt, diesmal mit dem Ziel, dass in jeder Gruppe Teilnehmer_innen aus allen Interessensgruppen vertreten sind. In dieser Arbeitsphase werden die am Vormittag identifizierten Themen diskutiert, mit dem Ziel erste skizzenhafte Positionspapiere je Gruppe zu entwickeln. Diese Entwürfe werden dann im Plenum zusammengeführt, mit dem Ziel ein gemeinsames Positionspapier zu formulieren.

Abschließend wird am Ende des Workshoptages in einer optionalen Session die formale Gründung der DHd AG besprochen und beschlossen.

Nach dem Workshop

Nach dem Workshop wird das gemeinsame Statuspapier mit den Teilnehmer_innen geteilt und kollaborativ fertiggestellt. Zusätzlich werden die formalen Schritte der DHd AG Gründung durchgeführt.

Call for Position Statements

Im Rahmen des DHd 2022 findet ein ganztägiger Workshop statt, der sich der Frage der Repräsentativität im digitalen Archiv widmet. Diese Frage wird insbesondere im Rahmen der Intersektionalitätsforschung stark thematisiert, nicht selten verbunden mit einer Kritik, u.a. an den Digital Humanities. Welche Artefakte werden im Digitalisierungsprozess priorisiert? Wer erzeugt wie die Metadaten für die digitalisierten Artefakte? Wie werden die gesuchten Artefakte im digitalen Archiv gefunden? Das Bewusstsein wächst dabei, dass Auswahlen, und insbesondere auch Digitalisierungsprozesse, Ergebnisse hegemonialer Machtstrukturen sein können, wie es u.a. das postcolonial computing thematisiert. Der zentrale Ansatzpunkt der Kritik am möglichen Quellenbias der Archive und den Digitalisaten ist zudem, dass oftmals viel zu wenig über die Zusammensetzung, die Auswahlkriterien usw. einzelner Korpora in digitalen Archiven bekannt ist und es an Transparenz mangelt. Die Komplexität ergibt sich nicht zuletzt daraus, dass die Interessen und Anforderungen verschiedener Akteur_innen zusammentreffen, die jeweils unterschiedliche Sichtweisen darauf haben, was Repräsentativität für sie bedeutet.

Der Workshop "Repräsentativität in digitalen Archiven" bietet ein Forum um Fragen hinsichtlich Auswahl, Annotation, Metada-

ten, Koordination, Nachhaltigkeit, Suchsystemen, Finanzierung, Dokumentation, Transparenz usw. vor diesem Hintergrund übergreifend und aus den verschiedenen Perspektiven (Archiv/Bibliothek, IT, Wissenschaft, Förderung usw.) zu thematisieren. Dafür wird vorab um kurze Positionspapiere von interessierten Teilnehmer_innen gebeten (max. 500 Worte), die mind. folgende Informationen enthalten sollten: was ist die jeweils eigene Perspektive, was sind die Interessen an dem Thema, welche Aspekte werden für relevant erachtet. Im Verlauf des Workshoptags wird ein gemeinsames Positionspapier erarbeitet und eine DHd-Arbeitsgruppe gegründet, um den Austausch und die Arbeit zum Thema "Repräsentativität" zu verstetigen.

Organisation

Die Teilnehmer_innenzahl ist vor Ort begrenzt, die Veranstaltung wird allerdings hybrid geplant. Die benötigte technische Ausstattung beschränkt sich auf einen Laptop. Spezifische Vorkenntnisse sind nicht nötig. Bedingung zur Teilnahme ist die Einreichung des Statuspapiers via Mail an die Organisator_innen bis zum 09.02.2022.

Corinna Dziudzia

Corinna.dziudzia@uni-erfurt.de
Forschungszentrum Gotha der Universität Erfurt
Schlossberg 2
99868 Gotha

Corinna Dziudzias Arbeitsschwerpunkte liegen im Bereich der Begriffs- und Wissenschaftsgeschichte, insbesondere zur Tradierung von literarischem Wissen und Kanonbildung. Wesentlich ist dafür die Digitalisierung und die dabei beobachtbaren Effekte.

Mark Hall

mark.hall@open.ac.uk
School of Computing & Communications
The Open University
Milton Keynes, MK7 6AA (UK)

Mark Halls Interessenschwerpunkte liegen auf der Öffnung des digitalen Kulturguts für die interessierte Öffentlichkeit und methodischer Klarheit in den Digital Humanities. Aus beiden Schwerpunkten ergibt sich ein Interesse an der Repräsentativität des digitalen Archives.

Bibliographie

Assmann, Aleida(2010): "Canon and Archive", in: Erll, Astrid / Nünning, Ansgar (eds.): *A Companion to Cultural Memory Studies*. Berlin / New York: De Gruyter 97-107.

Barats, Christine(2020): "Fading Away...The Challenge of Sustainability in digital studies", in: *digital humanities quarterly*, Volume 14 Number 3, <http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html>[letzter Zugriff 10.11.2021].

Chan, A.S.(2018): "Decolonial computing and networking beyond digital universalism", in: *Catalyst: Feminism, Theory, Technology*, 4(2), 1-5.

Hall, Mark M. (2020): "Opportunities and risks in digital humanities research", in: Carius, Hendrikje / Prell, Martin / Smolar-

ski, René (eds.): *Kooperationen in den digitalen Geisteswissenschaften gestalten*, Göttingen: Vandenhoeck & Ruprecht 47-66.

Inwood, Kris / Maxwell-Stewart, Hamish(2020): "Selection Bias and Social Science History", in: *Social Science History*44, Nr. 3: 411-416 <https://doi.org/10.1017/ssh.2020.18>.

Irani, L. / Vertesi, J. / Dourish, P., Philip K. / Grinter, R.E.(2010): "Postcolonial computing: a lens on design and development", in: *Proceedings of the SIGCHI conference on human factors in computing systems*:1311-1320.

Kizhner, Inna [et.a.] (2021): "Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture", in: *Digital Scholarship in the Humanities*36, Nr. 3: 607-40. <https://doi.org/10.1093/llc/fqaa055>.

Malazita, James H. / Teboul, Ezra J. / Rafah, Hined(2020): "Digital Humanities as Epistemic Cultures: How DH Labs make Knowledge, Objects, Subjects", in: *digital humanities quarterly*, Volume 14 Number 3, <http://www.digitalhumanities.org/dhq/vol/14/3/000465/000465.html>[letzter Zugriff 10.09.2021].

Risam, Roopika(2015): "Beyond the Margins: Intersectionality and the Digital Humanities", in: *digital humanities quarterly*, Volume 9 Number 2, <http://digitalhumanities.org:8081/dhq/vol/9/2/000208/000208.html>[letzter Zugriff 10.09.2021].

Star, Susan Leigh / Griesemer, James R. (1989): "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39", in: *Social Studies of Science*19, Nr. 3: 387-420. <https://doi.org/10.1177/030631289019003001>.

Ziegler, S.L.(2019): "Digitization Selection Criteria as Anti-Racist Action", in: *Code{4}lib Journal*, Issue 45 <https://journal.code4lib.org/articles/14667>[letzter Zugriff 10.09.2021].

Textexplorationen in der digitalen Literaturwissenschaft Eine kritische und angewandte Auseinandersetzung mit Repräsentations- und Interpretationsansätzen von Text

Brandl, Stephanie

stephanie.brandl@tu-berlin.de
Københavns Universitet; Technische Universität Berlin

Lassner, David

lassner@tu-berlin.de
Technische Universität Berlin

Krömer, Cora

cora.kroemer@fau.de
Universitätsbibliothek Erlangen-Nürnberg

Baillot, Anne

anne.baillot@univ-lemans.fr
Le Mans Université

Anforderungen an den Ort

Maximalanzahl Teilnehmende: 25

Räumliche Anforderungen:

- Beamer
- Whiteboard/Tafel
- Stromversorgung für Laptops der Teilnehmenden
- Wifi

Anforderungen an die Teilnehmenden

Idealerweise bringen Teilnehmende ihre eigenen Laptops mit, die bestenfalls schon die nötige Software vorinstalliert haben. Wir werden kurz vor der Konferenz eine Willkommens-E-Mail mit einer Liste der relevanten Software verschicken. Die praktischen Sitzungen werden mithilfe von Jupyter Notebooks (Python3, Jupyter) abgehalten. Wir planen zusätzlich als Absicherung einen Online-Zugang zu einem JupyterHub Server mit vorinstallierten Paketen für Teilnehmende, bei denen die Installation Schwierigkeiten macht.

Wir ermutigen die Teilnehmenden ausdrücklich einen eigenen Datensatz mitzubringen, an Beispiel dessen die Aufgaben ausgeführt werden können. Wir gehen davon aus, dass dies aufschlussreich für die jeweiligen Teilnehmenden ist und zusätzlich den Workshop bereichert. Wir haben außerdem bereits mit zwei Dateninstitutionen Kontakt aufgenommen. Beide haben Interesse bekundet sowohl Daten für den Workshop zur Verfügung zu stellen als auch am Workshop teilzunehmen. Damit stellen wir sicher, dass auch für Teilnehmende, die keine Daten mitbringen, Material vorhanden ist, und im Zweifel auch Expert:innen vor Ort sind, die etwaige Zwischenergebnisse der verwendeten Methoden evaluieren und in Kontext setzen können.

Beschreibung

Traditionelle Methoden und Theorien der Literaturwissenschaft können in vier Typen unterteilt werden: autor-, text-, leser- und kontextorientiert [Köppe & Winko, 2013]. Die Literaturwissenschaft zeichnet sich durch einen Methodenpluralismus aus [Nünning & Nünning 2020], der in den letzten Jahren vor allem durch kognitionswissenschaftliche [Zunshine, 2015] und empirische Ansätze [Kuiken & Jacobs, 2020] weiter ausdifferenziert wurde. Gerade diese haben zu neuen Erkenntnissen in leserorientierten Ansätzen geführt. Umgekehrt lässt sich auch vermuten, dass neue Lektüreweisen, –in unserem Fall maschinengenerierte Lektüren– literaturwissenschaftliche Innovationen direkt oder mittelbar als Folgewirkung hervorbringen [Parr & Honold, 2018]. Nicht nur die Literaturwissenschaft an sich, sondern auch ihre leserorientierten Ansätze zeichnen sich durch Theorien- und Methodenpluralismus aus [Rautenberg & Schneider, 2015], z. B. rezeptionstheoretische Ansätze [Willand, 2014].

Die Verbreitung von digitalen Lesemedien führt zu einer neuen Beschäftigung mit Repräsentationsformen von Texten und Textorganisation [Saemmer, 2015], da der Sinn eines Textes nicht davon unabhängig gedacht werden kann. Dies führt zu neuen und zu erlernenden Lesestrategien, die es zu erforschen gilt.

Auf der anderen Seite prägen visuelle Repräsentationen von Literatur schon lange ihre Institutionalisierung: Literarische Bewegungen werden in Schulbüchern in Form von Zeitleisten dargestellt, dramatische Handlungen mit der Freytagschen Pyramide abgebildet, und dies schon seit dem 19. Jahrhundert. Bereits diese Ansätze setzen sich zum Ziel, übergeordnete Strukturen anschaulich zu machen, die sich aus dem literarischen Text selbst ableiten lassen. Mit der bourdieuschen Literatursoziologie etablierte sich in den 60er Jahren die Vorstellung der Literatur als ein Feld, das von diversen Dynamiken durchzogen wird. Die digitalen Methoden, spätestens seit Morettis *Graphs Maps Trees* (2005), gehen in der räumlich-visuellen Exploration von Literatur weiter. Der Einsatz von Machine Learning-Methoden seit den 2010er Jahren führte einerseits zur Diversifizierung der Visualisierungsmethoden (WordClouds als Visualisierung von Topic Models, Netzwerke als Ergebnis von Named Entity Recognition oder Distanzgraphen aus stilometrischen Analysen), andererseits zur Repräsentation von immer abstrakteren Literaturphänomenen, bei denen der unmittelbare Zusammenhang mit dem literarischen Text nur mittels komplexer Rechenmethoden nachzuvollziehen ist. In diesem Zusammenhang haben sich Word Embeddings als eine mögliche vektorielle Repräsentationsmethode etabliert, die sich unmittelbar an die Semantik des literarischen Textes anlehnt. Wörter werden hier, basierend auf ihrem Kontext in Bezug auf andere Wörter im Text, als Vektoren dargestellt. Der resultierende Vektorraum bildet so semantische wie syntaktische Beziehungen zwischen Wörtern ab.

Word Embeddings bilden häufig die Brücke zwischen dem Text als Zeichenfolge und der Darstellung, mit der digitale Literaturwissenschaftler:innen ihr Korpus lesen, also bspw. als Input für ein ML-Modell.

In *Reading Machines* rückt Stephen Ramsay eine *Radical Transformation* [Ramsay 2011] der Lesart in den Fokus: kein sequentieller Leseprozess sondern z.B. Kapitel in zufälliger Reihenfolge, aber auch etwa 'nur die häufigsten 100 Worte lesen'. Damit stellt sich in hermeneutischer Perspektive die Frage der Angemessenheit der gewählten Transformation. Wenn wir die Überführung in Word Embeddings ebenfalls als solch eine Radical Transformation betrachten, wie wirkt sich das auf das Lesen aus? Ist einem Word Embedding, das auf dem Werk eines Autors oder einer Autorin trainiert wurde, die Angemessenheit für sie oder ihn in gewisser Weise durch den Lesehorizont dessen einprogrammiert? Zieht man Erkenntnisse aus neurowissenschaftlicher Studien heran, lässt sich sogar feststellen, dass bereits eine kleine Veränderung der visuellen Darstellung eines Gedichts einen messbaren Einfluss auf das Lesen hat [Fechino et al, 2020].

Eine weitere große Herausforderung bei der Verwendung von Word Embeddings ist die zuverlässige Evaluation ihrer Qualität, das heißt, wie sehr sie mit der gewünschten Bedeutung übereinstimmen. Es gibt zwar eine Menge von Evaluationsmethoden, wie Analogy Tests oder die Applikation von Downstream Tasks, häufig klopfen diese allerdings nur Teile der zugrundeliegenden Repräsentation ab und geben kein umfassendes Bild über ihre Qualität.

Ein in diesem Zusammenhang viel diskutierter Aspekt von Word Embeddings sind 'Biases', der zugrundeliegenden Daten, die durch die Modellarchitekturen reproduziert, oder sogar verstärkt werden können. Dies ist insbesondere bei Systemen ein Problem, die auf so großen und heterogenen Datenmengen trainiert

wurden, dass es sich im Nachhinein schwer dokumentieren lässt, welche Biases dem jeweiligen Modell innewohnen (Bender et al., 2021). In Bezug auf die häufig kleineren, gut mit Metadaten versehenen Korpora in den DH, steckt in diesen reproduzierten Biases aber tatsächlich ein interessantes Forschungsfeld (Gebru et al., 2018). So lässt sich beispielsweise sehen, in welchen Zeiträumen welche Biases besonders vorherrschend waren, bzw. wann diese möglicherweise wieder verschwinden. Hierfür könnten mehrere Word Embeddings für verschiedene Zeiträume trainiert, und dann miteinander verglichen werden.

Diese Art von Vergleich ist aber nicht nur auf Biases beschränkt, man kann sie ebenfalls als eine Form des 'Realitätschecks' durchführen, wenn man Dynamiken zwischen Word Embeddings betrachtet: Bei einem Korpus, das einen größeren Zeitraum umfasst, verändern sich darin allgemein die Repräsentationen der Worte so, wie man es von der Bedeutung der Worte auch erwarten würde?

In dem Workshop möchten wir uns der beschriebenen Thematik von zwei Seiten nähern, wir geben (1) einen Überblick über traditionelle Theorien und Methoden der Literaturwissenschaft, insbesondere auch der Leseforschung und kontextualisieren letztere mit Erkenntnissen aus neurowissenschaftlichen Laborstudien. Wir geben (2) eine Zusammenfassung der aktuellen Methoden zu Textrepräsentation [z.B. Glove, Bert] und Beispiele für DH-Projekte, in denen diese bereits verwendet werden, insbesondere fokussieren wir uns hier auch auf gängige Visualisierungsmethoden und die damit zusammenhängenden Limitationen. Beides wollen wir durch praktische Übungen an eigenen oder bereitgestellten Datensätzen näherbringen. Ziel ist es, dass im Laufe des Workshops Verknüpfungen und Schnittmengen der unterschiedlichen Lesebegriffe aus den verschiedenen Fachrichtungen hergestellt und gefunden werden.

Zum Abschluss bieten wir konkrete Möglichkeiten an, welche Visualisierungen auf welche Weise ihren Platz als literaturwissenschaftliche Methode finden können.

Ablauf

Der Ablauf des Workshops gliedert sich in Vorträge sowie praktische Sitzungen, bei denen das im Vortrag zuvor besprochene direkt von den Teilnehmenden ausprobiert werden soll. Im folgenden soll ein kurzer Überblick über die einzelnen Vortragsteile gegeben werden.

Traditionelle Theorien und Methoden der Literaturwissenschaft

(Cora Krömer)

Theorien und Methoden der Literaturwissenschaft können allgemein in vier Typen unterteilt werden: text-, autor-, leser-, kontextorientiert [Köppe & Winko, 2013]. Als Einführung in den Workshop werden wir uns insbesondere einen Überblick über text- und leserorientierte Theorien und Methoden verschaffen, um die nachfolgenden Einheiten besser einordnen zu können. Digitale Texte (z. B. Hypertexte), neue empirische Quellen (z. B. Online-Leserkommentare) und digitale Methoden laden dazu ein, über das Lesen und Interpretieren von Texten und ihren Visualisierungen nachzudenken.

Einführung Textrepräsentationen ML

(David Lassner)

Hier werden verschiedene für DH relevante Textrepräsentationen besprochen. Dabei wird zum einen projektbezogen der typische Ablauf (Buch-Scan-OCR-TEI-NLP) durchgegangen und die verschiedenen verwendeten Repräsentationen verglichen, zum anderen die Entwicklung von Textrepräsentationen in NLP der letzten Jahren nachvollzogen, also von Bag-of-Words, über Word2Vec bis hin zu Bytepair Encoding und kontextualisierten Embeddings und der Frage, wie man aus diesen wieder generelle Wortrepräsentationen erhält.

Neuroscience

(Stephanie Brandl)

In welcher Form und Geschwindigkeit Menschen Texte lesen wird bereits seit Jahrzehnten untersucht [Rayner, 1998] und modelliert [Reichle et al., 2003]. Oberflächliche Eigenschaften wie Worthäufigkeiten oder Wortlängen beeinflussen unsere Lesegeohnheiten. Neuere Forschung zeigt zudem, dass auch die Darstellung des Textes eine Rolle spielt, ob ein Gedicht beispielsweise in der ursprünglichen Form oder als Prosatext gezeigt wird [Fechino et al., 2020]. Dementsprechend ist es auch wichtig, Word Embeddings so darzustellen, dass der sehr dichte Informationsgehalt verarbeitet und evaluiert werden kann.

Limitationen von Word Embeddings

(Stephanie Brandl)

Word Embeddings sind stark abhängig vom zu Grunde liegenden Datensatz, beispielsweise lernen statische Methoden wie GloVe genau eine Repräsentation pro Wort, so dass zeitliche oder andere Entwicklungen im Datensatz nicht beachtet werden und in einem Vektor verschmelzen. Diese Abhängigkeit führt auch dazu, dass Verzerrungen (Biases) im Datensatz möglicherweise in den Wort-Vektoren auftauchen. Wenn beispielsweise alle Personen des medizinischen Personals im Datensatz weiblich sind, wird ein Algorithmus einen männlichen Bewerber möglicherweise für eher ungeeignet halten, um als Arzt zu arbeiten. Verschiedene Methoden wurden bereits veröffentlicht, um solche problematischen Strukturen (zeitliche Abhängigkeit, Biases uvm) in Word Embeddings aufzuzeigen [Basta et al., 2019; Brandl & Lassner, 2019; Gonen et al., 2020] und auch aufzulösen [Gonen et al., 2019; Manzini et al., 2019]. Wir werden einige davon besprechen und im Anschluss auch an den gelernten Word Embeddings anwenden.

Lesen von Textrepräsentationen und Visualisierungen

(Anne Baillot)

Zum Schluss wird auf die Entwicklung literaturwissenschaftlichen Umgangs mit Analyse bzw. Interpretation von Text eingegangen: zuerst auf das Faszinosum Netzwerk, dann auf die Herausforderung der Definition von Beziehungen zwischen Textelementen und ihrer Bedeutung, schließlich auf veröffentlichungstechnische Fragen, die mit der Einbettung dieser graphischen Repräsentationen einhergehen.

Bibliographie

Basta, C. R. S., Ruiz Costa-Jussà, M., & Casas Manzanares, N. (2019). "Evaluating the underlying gender bias in contextualized word embeddings." In *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 33-39).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

Brandl, S., & Lassner, D. (2019). "Times Are Changing: Investigating the Pace of Language Change in Diachronic Word Embeddings". In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (pp. 146-150).

Fechino M., Jacobs A.M., Lüdtke J. **Following** (2020). "Jakobson and Lévi-Strauss' footsteps: A neurocognitive poetics investigation of eye movements during the reading of Baudelaire's 'Les Chats'". In *Journal of Eye Movement Research*, ff10.16910/jemr.13.3.4ff. Ffhal-02749759f

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for datasets*. arXiv preprint arXiv:1803.09010.

Gonen, H., & Goldberg, Y. (2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 609-614).

Gonen, H., Jawahar, G., Seddah, D., & Goldberg, Y. (2020). "Simple, interpretable and stable method for detecting words with usage change across corpora". In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 538-555).

Köppe, S. & Winko, S. (2013). "Theorien und Methoden der Literaturwissenschaft". In Anz, T. (Hrsg.) *Handbuch Literaturwissenschaft*. Bd. 2. Stuttgart: J.B. Metzler. DOI 10.1007/978-3-476-01271-5. (pp. 285-371).

Kuiken, D. & Jacobs, A. (2021). *Handbook of Empirical Literary Studies*. Boston: De Gruyter.

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). "Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings". In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Moretti, F. (2005). *Graphs, Maps, Trees. Abstract Models for a Literary History*. London/New York.

Nünning, V. & Nünning, A. (2020). *Methods of Textual Analysis in Literary Studies Approaches, Basics, Model Interpretations*. Trier: Wissenschaftlicher Verlag Trier.

Parr, R. & Honold A. (2018). *Grundthemen der Literaturwissenschaft: Lesen*. Berlin; Boston: De Gruyter.

Ramsay, S. (2011). *Reading Machines: Toward and Algorithmic Criticism*. University of Illinois Press.

Rautenberg, U. & Schneider, U. (2015). *Lesen: ein interdisziplinäres Handbuch*. Berlin; Boston: De Gruyter.

Rayner, K. (1998). "Eye movements in reading and information processing: 20 years of research". *Psychological bulletin*, 124(3), 372.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). "The EZ Reader model of eye-movement control in reading: Comparisons to other models". *Behavioral and brain sciences*, 26 (4), 445-476.

Saemmer, A. (2015). *Rhétorique du texte numérique: Figures de la lecture, anticipations de pratique*. Villeurbanne.

Willand, M. (2014). *Lesermodelle und Lesertheorien. Historische und systematische Perspektiven*. Berlin und Boston.

Zunshine, L. (2015). *The Oxford Handbook of Cognitive Literature Studies*. Oxford.

Vom Begriff über das Phänomen zur Analyse Ein CRETA-Workshop zur Operationalisierung in den DH

Andresen, Melanie

melanie.andresen@ims.uni-stuttgart.de
University of Stuttgart, Germany

Krautter, Benjamin

Benjamin.Krautter@gs.uni-heidelberg.de
Heidelberg University, Germany

Pagel, Janis

janis.pagel@ims.uni-stuttgart.de
University of Stuttgart, Germany

Pichler, Axel

axel.pichler@fu-berlin.de
FU Berlin, Germany

Der Workshop stellt eine weiterentwickelte und personell anders besetzte Version des auf der DHd 2020 abgehaltenen, beinahe gleichnamigen Workshops dar. Er adressiert eine der zentralen Herausforderungen für Arbeiten in den Digital Humanities – die Operationalisierung geisteswissenschaftlicher Konzepte und Fragestellungen für computergestützte Forschungsansätze (vgl. Jannidis 2010: 109–132; Moretti 2013; Flanders/Jannidis 2015; Jacke 2014: 118–139; Pichler/Reiter 2020, Pichler/Reiter 2021). Während Geisteswissenschaftler*innen vor allem mit komplexen, häufig mehrere Textphänomene umfassenden Konzepten arbeiten und als relevant erachtete Kontexte zu deren Deutung heranziehen, ist die computergestützte Arbeit an identifizierbare Phänomene auf der Textoberfläche gebunden. Die hieraus erwachsende Diskrepanz zwischen theoretischen Erwartungen und konkreten Ergebnissen gilt es über eine adäquate Operationalisierung zu überbrücken (vgl. Moretti 2013: 1). Ziel ist es also, Verfahren zu entwickeln, die theoretische Begriffe über potenziell mehrere Teilschritte auf Textoberflächenphänomene zurückführen. Oder kurz gesagt: **die Erkennung und Messbarmachung von Instanzierungen theoretischer Konzepte**. Mit unserem Workshop wollen wir genau diese Schnittstelle in den Fokus rücken. Anhand ausgewählter Anwendungsfälle zeigen wir, welche Herausforderungen sich aus dem Einsatz computergestützter Methoden für geisteswissenschaftliche Fragestellungen ergeben und wie mit ihnen umgegangen werden kann. In einem praktischen Teil haben die Teilnehmenden die Möglichkeit, an der Operationalisierung vorgegebener exemplarischer Fragestellungen der Textanalyse und der dafür relevanten Konzepte zu arbeiten. Hierfür

stellen wir Anwendungsfälle mit geeigneten Tools und Technik-„Baukästen“ zur Verfügung. Programmierkenntnisse werden dabei nicht vorausgesetzt. Ziel des Workshops ist es, das Bewusstsein für die Differenzen zwischen geisteswissenschaftlicher und computergestützter Arbeitsweise zu schärfen, typische Herausforderungen zu adressieren und Herangehensweisen zur Operationalisierung geisteswissenschaftlicher Konzepte aufzuzeigen. Denn nur durch die reflektierte Auseinandersetzung mit den Operationalisierungsannahmen kann ein angemessener Umgang mit den Ergebnissen gewährleistet werden.

Use Cases

Als Anwendungsfälle stellen wir Phänomene vor, zu denen wir im Rahmen des „Center for Reflected Text Analytics“ e.V. (CRETA) ¹ umfangreiche Erfahrungen gesammelt haben. Die gewählten Beispiele decken verschiedene Aufgabentypen ab: Wir behandeln erstens die Extraktion bestimmter Instanzen aus einem Text und zweitens ein holistisches Textphänomen.

Entitäten und Entitätenreferenzen

In einem ersten Anwendungsfall befassen wir uns mit dem Konzept der Entität und ihrer Referenz in literarischen Texten (vgl. Reiter u.a. 2017: 19–22; Blessing u.a. 2020). Dabei fassen wir den Begriff der Entität sehr weit: „Alles, was man als Einheit denken kann, kann als Entität behandelt werden“ (Jannidis 2017: 103). Zu den Entitäten zählen dementsprechend Personen/Figuren, Orte, Organisationen sowie Ereignisse. Das Konzept ist also für verschiedene Forschungsfragen anschlussfähig. Auf Entitäten kann auf verschiedene Weise referiert werden, etwa über Eigen- und Gattungsnamen (z. B. „Angela Merkel“, „die Kanzlerin“). Um Entitäten in einem Text zu extrahieren, müssen folglich die Entitätenreferenzen annotiert und kookkurrenente Ausdrücke aufgelöst werden. Die Herausforderungen bestehen vor allem in der Festlegung der Referenzausdrücke (welche Ausdrücke werden berücksichtigt?), in der Abgrenzung von Entitätenreferenzen gegenüber generischen Ausdrücken sowie im Umgang mit Verschachtelungen, Metonymien und textspezifischen Besonderheiten.

Protagonisten im Drama

Der zweite Anwendungsfall setzt sich mit der Identifikation von Protagonisten im Drama auseinander, fokussiert also ein holistisches Textphänomen. Die verschiedenen Perspektiven der Literaturwissenschaft auf Protagonisten, Hauptfiguren und Helden von Dramen (vgl. die Ausführungen in Krautter u.a. 2018: 6–16 und Wulff 2002: 431–448) haben zur Folge, dass eine Reihe von Definitionen und Identifikationsstrategien koexistieren, die häufig an historische Normvorstellungen geknüpft sind. Diese historische Gebundenheit erschwert die operationale Definition von Protagonisten, wenn man auf größere Abschnitte der Literaturgeschichte blickt.

Direkt anschlussfähig für die Methoden der Digital Humanities erscheint die in den späten 1970er Jahren von Manfred Pfister skizzierte Annahme, dass „quantitative[] Dominanzrelationen“ (Pfister 2001: 227) hilfreich für die Differenzierung des Bühnenpersonals seien. Pfister nennt zwei Kriterien, die dabei helfen können, dramatische Figuren schon aufgrund quantitativer Eigenschaften als Haupt- oder Nebenfiguren zu identifizieren: nämlich

die Zeitdauer, die sie auf der Bühne stehen, und ihr Anteil an der gesamten Figurenrede (vgl. Pfister 2001: 226–227). Diese Auffassung Pfisters lässt sich mit digitalen Methoden der Dramenanalyse um weitere Eigenschaften der Figuren, etwa durch Netzwerkmetriken oder Topic Modeling, zu einem multidimensionalen Ansatz ergänzen. Die größte Herausforderung stellt hierbei die Validierung der Ergebnisse dar, da diese an die Gültigkeit der operationalen Definition für die manuelle Annotation gebunden ist.

Ansätze zur Operationalisierung

Im Workshop stellen wir zwei Ansätze zur Operationalisierung vor, die sich – in verschiedenen Phasen des Forschungsprozesses – sehr gut gegenseitig ergänzen. Der erste Ansatz besteht in der operationalen Definition theoretischer Konzepte durch manuelle Annotationen. Die Ergebnisse sind also keine Skripte oder Funktionen, sondern klare(re), als Abfolge von Operationen bestimmte Definitionen der fraglichen Konzepte, die von Menschen mit hoher intersubjektiver Übereinstimmung umgesetzt werden, und zudem die theoretische Diskussion bereichern können (vgl. Gius/Jacke 2017; Pagel u.a. 2020; Reiter 2020, Pichler/Reiter 2021). Daneben führt der Annotationsprozess auch zu einer intensiven und kritischen Beschäftigung mit den Texten und den textuellen Indikatoren des Konzeptes und liefert damit auch Ideen für eine computergestützte Operationalisierung.

Als zweiten Ansatz stellen wir eine Vorgehensweise vor, die Zielphänomene indirekt operationalisiert. Da sich viele geisteswissenschaftliche Konzepte nicht direkt messen lassen, sie also nicht unmittelbar durch mögliche Instanzen auf der Textoberfläche repräsentiert werden, müssen die Konzepte schrittweise auf messbare Eigenschaften zurückgeführt werden. In diesem Fall werden also mehrere messbare Eigenschaften in den Blick genommen, die mit dem Zielkonzept verwandt, aber nicht deckungsgleich sind und das Konzept somit indirekt operationalisieren. Aufschlussreich ist dabei in erster Linie die Gesamtschau der verschiedenen als relevant erachteten Einflussfaktoren (vgl. „instrumental variables“ in Sack 2011; „indirekte Operationalisierung“ in Reiter/Willand 2018). Bei textbasierten Phänomenen lassen sich so insbesondere linguistische und strukturelle Eigenschaften, die größtenteils mit hoher Reliabilität automatisch extrahierbar sind, in die Operationalisierung integrieren.

Ablauf

In einem Theorieteil führen wir in Geschichte und Praxis der Operationalisierung von geisteswissenschaftlichen Fragestellungen und Konzepten für die computergestützte Analyse ein. Anhand der oben genannten Beispiele aus der CRETA-Praxis thematisieren wir die Problematik und stellen Ansätze zur Operationalisierung im Detail vor. Je nach Interesse kann anschließend, im praktischen Teil, einer dieser Anwendungsfälle ausgewählt und bearbeitet werden. Dabei haben die Teilnehmenden die Möglichkeit, beide Operationalisierungsansätze an ihrem gewählten Anwendungsfall zu erproben. Hierfür befassen sie sich zunächst mit dem Konzept, indem sie es anhand eines Textauszugs manuell annotieren und parallel stichpunktartig die Richtlinien schärfen. In einer ersten Diskussionsrunde werden die verschiedenen Ergebnisse gesammelt und diskutiert. Zur Erprobung des zweiten Ansatzes stellen wir für jeden Anwendungsfall einen Operationalisierungs-„Baukasten“ vor. Dieser besteht aus einer Sammlung von Python-Skripten in einem Jupyter-Notebook,

das auf das jeweilige Untersuchungsvorhaben zugeschnitten ist und den Teilnehmenden die Möglichkeit gibt, sich dem zu untersuchenden Phänomen über computergestützte Verfahren anzunähern. Die Teilnehmenden können in Kleingruppen in diesem Baukasten verschiedene Parameter einstellen sowie manuell Eigenschaften an- oder abwählen, wobei sie auf ihr Vorwissen über den Untersuchungsgegenstand aus der ersten Praxisrunde zurückgreifen können. Nachdem die Teilnehmenden die Eigenschaften ausgewählt und ggf. parametrisiert haben, können sie die Ergebnisse visualisieren und mit den Texten abgleichen. Damit erhalten die Teilnehmenden ein direktes Feedback zu den ausgewählten Parametern und können prüfen, ob das Untersuchungsvorhaben mit den festgelegten Einstellungen angemessen umgesetzt wird. Der Baukasten ist zur iterativen Nutzung vorgesehen, sodass der Einfluss verschiedener verwandter Eigenschaften auf die Ausgaben sichtbar wird und die Teilnehmenden sich einer geeigneten technischen Umsetzung sukzessiv annähern können. In einer abschließenden Diskussion werden die Ergebnisse gesammelt und es wird ausgewertet, wie adäquat sich die jeweiligen Zielphänomene mittels der gewählten Annahmen abbilden haben lassen.

Lernziele

Ziel unseres Workshops ist es, die Teilnehmenden für die Wichtigkeit der Operationalisierung in den Digital Humanities zu sensibilisieren und ihnen Wege zu ihrer erfolgreichen Realisierung vorzustellen. Durch die interdisziplinäre Ausrichtung von DH-Arbeiten kommt der Operationalisierung eine Schlüsselposition zu, da sie eine Brücke zwischen geisteswissenschaftlichen Konzepten und computergestützter Umsetzung schlägt (vgl. Moretti 2013: 1). Mit den gewählten Anwendungsfällen wollen wir den Teilnehmenden ein „Repertoire“ für die Operationalisierung verschiedener Aufgabentypen mitgeben. Wir zeigen zum einen, dass die Annotation eines Phänomens als Methode seiner Operationalisierung dienen kann (vgl. Gius/Jacke 2017: 233–254); zum anderen führen wir für textbasierte Phänomene eine indirekte Operationalisierung ein (vgl. Reiter/Willand 2018). Beide Verfahrensweisen sind auf andere Anwendungsfälle übertragbar. Gleichzeitig möchten wir deutlich machen, dass es für jedes Untersuchungsvorhaben nicht nur eine, sondern verschiedene Wege der Operationalisierung gibt. Die Spielräume, die bei der Operationalisierung geisteswissenschaftlicher Fragestellungen entstehen, machen es notwendig, Entscheidungen reflektiert zu treffen, sie offenzulegen und ihren Einfluss auf die Ergebnisse als Voraussetzung für eine angemessene Interpretation zu bedenken.

Anhang

Zeitplan

(insgesamt 3 Stunden + 30 Min. Pause)

1. Einführung und Ablauf (10 Min.)

2. Theoretischer Teil (insgesamt 30 Min.)

- Erläuterung der Problemstellung

- Vorstellung der Anwendungsfälle

- 3. Praktischer Teil

- Einführung in die Primärtexte und Tools, Ausgabe der skizzierten Guidelines (10 Min.)

- Erste Praxisrunde (Kleingruppen): Manuelle Annotation eines Phänomens, parallele Erweiterung/Überarbeitung der Guidelines, iterativ (30–40 Min.)

- Kaffeepause (30 Min.) –
- Sammeln der Ergebnisse und Diskussion der Herangehensweisen (20 Min.)
- Zweite Praxisrunde (Kleingruppen): Arbeit am Operationalisierungsbaukasten, Feedback über Ausgabedatei, iterativ (30-40 Min.)
- 4. Abschlussdiskussion: Sammeln der Ergebnisse, Diskussion der Erfahrungen und Lernziele (30 Min.)

Die Durchführung des Workshops auf der DHd 2020 hat gezeigt, dass das gestraffte dreistündige Format gute didaktische Resultate zeitigt. Der Fokus auf die praktischen Dimensionen der Operationalisierung ist dabei gewollt: Aus der konkreten praktischen Arbeit heraus lässt sich unserer Ansicht nach am besten der theoretische Rahmen und die theoretischen Probleme bei der Operationalisierung reflektieren.

Zahl der möglichen Teilnehmenden

Zwischen 15 und 25

Angaben zur technischen Ausstattung

Abgesehen von Beamer und ausreichend Steckdosen ist keine besondere technische Ausstattung erforderlich. Die Teilnehmenden arbeiten im praktischen Teil an ihrem eigenen Laptop. Informationen zu eventuellen Vorab-Installationen werden rechtzeitig mitgeteilt.

Beitragende

Der Workshop wird von Mitgliedern des Center for Reflected Text Analytics (CRETA) e.V. veranstaltet, die bereits erfahrene Workshop-Leiter*innen im DH-Bereich sind (DHd 2017, DHd 2018, ESU 2018, DHd 2019, HCH 2019, DHd 2020).

CRETA konzentriert sich auf die Entwicklung von Methoden zur kritisch-reflektierten Textanalyse im Forschungsbereich der Digital Humanities. Die Methoden werden fachübergreifend für textanalytische Fragestellungen aus der Literatur-, Sprach-, Geschichts- und Sozialwissenschaft sowie Philosophie erarbeitet und eingesetzt. Das bis 2020 vom BMBF geförderte eHumanities-Zentrum ist Ende 2020 mit der Gründung eines Vereines in eine neue Phase übergegangen. Mit der Vereinsgründung wird der Tatsache Rechnung getragen, dass über Stuttgart hinaus inzwischen Wissenschaftler*innen an ähnlichen Zielen arbeiten und CRETA in vielfältiger Weise verbunden sind, etwa durch gemeinsame Projekte.

Melanie Andresen

melanie.andresen@ims.uni-stuttgart.de
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart

Melanie Andresen ist Postdoc am Institut für Maschinelle Sprachverarbeitung an der Universität Stuttgart. Sie hat Germanistische Linguistik an der Universität Hamburg studiert und ist dort 2020 im Bereich der Korpuslinguistik promoviert worden. Aus den Projekten *herma* (Universität Hamburg) und *Q:TRACK* (Universität Stuttgart, Universität zu Köln) bringt sie viel Erfahrung mit der Operationalisierung geistes- und sozialwissenschaftlicher Fragestellungen mit.

Benjamin Krautter

Benjamin.Krautter@uni-koeln.de
Universität zu Köln
Institut für Digital Humanities
Albertus-Magnus-Platz
50931 Köln

Benjamin Krautter ist Promotionsstudent am Germanistischen Seminar der Universität Heidelberg und Mitarbeiter im Projekt *Q:TRACK*. Dort arbeitet er u. a. an der Operationalisierung literaturwissenschaftlicher Kategorien für die quantitative Dramenanalyse. Im Zentrum seines Forschungsinteresses steht dabei die mögliche Verbindung quantitativer und qualitativer Methoden für die Analyse und Interpretation literarischer Texte.

Janis Pagel

janis.pagel@uni-koeln.de
Universität zu Köln
Institut für Digital Humanities
Albertus-Magnus-Platz
50931 Köln

Janis Pagel ist Promotionsstudent am Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart und Mitarbeiter am Institut für Digital Humanities der Universität zu Köln. Er studierte Germanistik und Linguistik in Bochum, sowie Computerlinguistik in Stuttgart und Amsterdam. Er forscht zu Anwendungen von computerlinguistischen Methoden auf literaturwissenschaftliche Fragestellungen und Koreferenzresolution auf literarischen Texten.

Axel Pichler

axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart

Axel Pichler studierte Philosophie und Germanistik in Wien und Graz. Im Sommersemester 2021 war er Gastprofessor für Digital Humanities am EXC "Temporal Communities" der FU Berlin. Zurzeit arbeitet er als Postdoc unter anderem an der Entwicklung und Reflexion von Methoden der computergestützten Textanalyse am Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart.

Fußnoten

1. <https://cretaverein.de/>

Bibliographie

Blessing, André / Echelmeyer, Nora / John, Markus / Reiter, Nils (2017): „An end-to-end environment for research question-driven entity extraction and network analysis“, in: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver.

Flanders, Julia / Jannidis, Fotis (2015): *Knowledge Organization and Data Modeling in the Humanities*. https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/11127/file/flanders_jannidis_datamodeling.pdf

Gius, Evelyn / Jacke, Janina (2017): „The Hermeneutic Profit of Annotation. On Preventing and Fostering Disagreement in Literary Analysis“, in: *International Journal of Humanities and Arts Computing* 11, S. 233–254.

Jacke, Janina (2014): „Is There a Context-Free Way of Understanding Texts? The Case of Structuralist Narratology“, in: *Journal of Literary Theory* 8, S. 118–39.

Jannidis, Fotis (2017): „Grundlagen der Datenmodellierung“, in: *Digital Humanities. Eine Einführung*, hg. v. Fotis Jannidis, Hubertus Kohle und Malte Rehbein, Stuttgart, S. 99–108.

Jannidis, Fotis (2010): „Methoden der computergestützten Textanalyse“, in: *Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze – Grundlagen – Modellanalysen*, hg. v. Vera Nünning, Ansgar Nünning und Irina Bauer-Begerow, Stuttgart, Weimar, S. 109–132.

Ketschik, Nora / Overbeck, Maximilian / Murr, Sandra / Pichler, Axel und André Blessing (2020): „Interdisziplinäre Annotation von Entitätenreferenzen“, in: *Reflektierte algorithmische Textanalyse*, hg. v. Nils Reiter, Axel Pichler und Jonas Kuhn, Berlin. <https://doi.org/10.1515/9783110693973-010>

Krautter, Benjamin / Pagel, Janis / Reiter, Nils / Willand, Marcus (2018): „Titelhelden und Protagonisten – Interpretierbare Figurenklassifikation in deutschsprachigen Dramen“, in: *Litlab Pamphlet* 7.

Moretti, Franco (2013): „Operationalizing”: or, the function of measurement in modern literary theory“, in: *Literary Lab* 6, S. 1–13.

Pagel, Janis / Reiter, Nils / Rösiger, Ina / Schulz, Sarah (2020): „Annotation als flexibel einsetzbare Methode“, in: *Reflektierte algorithmische Textanalyse*, hg. v. Nils Reiter, Axel Pichler und Jonas Kuhn, Berlin. <https://doi.org/10.1515/9783110693973-006>

Pichler, Axel / Reiter, Nils (2020): „Reflektierte Textanalyse“, in: *Reflektierte algorithmische Textanalyse*, hg. v. Nils Reiter, Axel Pichler und Jonas Kuhn, Berlin. <https://doi.org/10.1515/9783110693973-003>

Pichler, Axel / Reiter, Nils (2021): „Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists Das Erdbeben in Chili“, in: *Journal of Literary Theory* 15, S. 1–29. <https://doi.org/10.1515/jlt-2021-2008>

Pfister, Manfred (2001): *Das Drama. Theorie und Analyse*, München.

Reiter, Nils (2020): „Anleitung zur Erstellung von Annotationsrichtlinien“, in: *Reflektierte algorithmische Textanalyse*, hg. v. Nils Reiter, Axel Pichler und Jonas Kuhn, Berlin. <https://doi.org/10.1515/9783110693973-009>

Reiter, Nils / Gius, Evelyn / Willand, Marcus (Hg.) (2019): *A Shared Task for the Digital Humanities. Special issue of Cultural Analytics*. November 2019.

Reiter, Nils / Blessing, André / Echelmeyer, Nora / Kremer, Gerhard / Koch, Steffen / Murr, Sandra / Overbeck, Maximilian / Pichler, Axel (2017): CUTE: CRETA Unshared Task zu Entitätenreferenzen“, in: *DHd 2017 Bern*, Conference Abstracts, S. 19–22.

Reiter, Nils / Willand, Marcus (2018): „Poetologischer Anspruch und dramatische Wirklichkeit: Indirekte Operationalisierung in der digitalen Dramenanalyse“, in: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*, hg. v. Toni Bernhart, Marcus Willand, Sandra Richter und Andrea Albrecht, Stuttgart, S. 45–76.

Sack, Graham Alexander (2011): „Simulating Plot: Towards a Generative Model of Narrative Structure“, in: *Papers from the AAAI Fall Symposium* (FS-11-03).

Wulff, Hans Jürgen (2002): „Held und Antiheld, Prot- und Antagonist: Zur Kommunikations- und Texttheorie eines komplizierten Begriffsfeldes. Ein enzyklopädischer Aufriß“, In: *Weltent-*

würfe in Literatur und Medien. Phantastische Wirklichkeiten realistische Imaginationen. Festschrift für Marianne Wünsch, hg. v. Hans Krah und Claus-Michael Ort. Kiel, S. 431–448.

Wahrnehmungsstrukturen und User Experience des digitalen Kulturerbes Ein Blick auf museale Online Sammlungen

Kienbaum, Janna

jkienbau@uni-potsdam.de
Universität Potsdam, Germany

Kreiseler, Sarah

sarah.kreiseler@leuphana.de
Leuphana Universität Lüneburg, Germany

Heidmann, Frank

frank.heidmann@fh-potsdam.de
Fachhochschule Potsdam, Germany

Der Workshop hinterfragt partizipativ die Visualisierungskultur und Wahrnehmungs- bzw. Nutzungspraktiken digitaler Kulturerbe-Daten am Beispiel musealer Sammlungen im Web. Im Zentrum steht der Einsatz von Eyetracking-Verfahren sowie Methoden der Heuristischen Usability-Evaluation. Ausgehend von den Webseiten „explorativ“ gestalteter Online Sammlungen werden die grafischen Benutzungsoberflächen als visuelle Schnittstelle zwischen interner Objektdatenbank und Rezipientenschaft kritisch beleuchtet. In engem Austausch mit den Teilnehmer*innen soll die User Experience praktisch erforscht werden. Dazu gehören einerseits Fragen nach den Wahrnehmungsstrukturen und der intuitiven Nutzung der Seiten, ihrer Zeichen und Navigation. Andererseits wird explizit ein Augenmerk auf die angebotene Auswahl der Informationen von v.a. Einzelobjektseiten und das Vernetzungspotential verlinkter Daten als Möglichkeit der Kunstvermittlung gelegt.

Gegenstand des Workshops

Im Zuge der Digitalisierungsstrategien von Museen spielt die Veröffentlichung und Repräsentation des kulturellen Erbes im Web eine essenzielle Rolle¹, deren Dringlichkeit durch die pandemiebedingte Schließung noch einmal verstärkt wurde. Untersuchungen des Erfahrungspotentials und der vermittelnden Wirkung von musealen Online Sammlungen gewinnen daher ebenso an Aktualität.² An die Stelle von physischer Materialität, Originalität, Größenformat und sinnlicher Erfahrung der Objekte treten in den Online Sammlungen die Faktoren Information, Zweidimensionalität der Bildschirmfläche und systematischer Vergleich. Online Sammlungen beziehen ihre Informationen meist aus den intern genutzten Datenbanken, welche als digitaler Zugang für Expert*innen die physischen Bestandskataloge ersetzen soll. Neben

der Veröffentlichung von wissenschaftlicher Dokumentation sollen Online Sammlungen auch ein kuratiertes Angebot darstellen (Vgl. Krämer 2001: 181).

Die grafischen Benutzungsoberflächen (GUI) dienen dabei als Schnittstelle und Scharnier der Vermittlung des digitalen Kulturerbes. Während tabellarisch strukturierte GUIs den Datenbank-Charakter des museumsinternen Dokumentationssystems übernehmen und den Rezipient*innen in erster Linie zur gezielten Recherche dienen, versuchen „explorative“ Zugänge ein freieres Erkunden zu ermöglichen (Vgl. Kreiseler et al. 2017). Dieses soll durch digitale Darstellungsmöglichkeiten (Zoom, Farb- und Bildvergleiche usw.) sowie durch die Vernetzung ihrer (Meta-)daten, Schlagworte, Ähnlichkeiten oder literarische Zusatzinformationen gefördert werden. Die verknüpften Interaktionsmöglichkeiten markieren einen zentralen Aspekt innerhalb der von Museen veröffentlichten Strategien zur digitalen Vermittlung³.

Das „explorative“ Erkunden zielt also auf eine bestimmte Form der Navigation ab. Das Prinzip des Browsens als ein „genussvolles“ Bewegen in vernetzten Strukturen wird dabei immer wieder dem vorherrschenden Prinzip der Datensuche bzw. dem Information Retrieval entgegengesetzt: „As an interface, search fails to match the ample abundance of our digital collections and the generous ethos of the institutions that hold them“ (Whitelaw 2015). Die Metapher des großstädtischen Flaneurs um 1900 aufgreifend, bewegt sich auch der „Informationsflaneur“ umherschweifend durch den digitalen Raum ohne dabei ein konkretes Ziel zu verfolgen (Dörk et al. 2011: 1). Vielmehr wird die Bewegung durch ein Interesse und durch andere Akteur*innen geleitet. So sollen auch Besucher*innen von Online Sammlungen zum Flanieren angeregt werden, ohne dass eine Suchanfrage gestellt werden muss.

Ein Umherschweifen wird begünstigt durch das Angebot verschiedener Ansichten, vor allem Distanzverfahren ermöglichen dem „Informationsflaneur“ zunächst den Blick schweifen zu lassen (Vgl. Glinka 2016; Vgl. Dörk 2017).⁴ Diverse digitale Visualisierungen setzen hier auf ein Wechselspiel aus Distant- und Close-Viewing-Methoden vernetzter musealer Daten. Sie basieren zunehmend auf computationalen Analysealgorithmen, die einer bestimmten Kuration folgen.⁵ Das Prinzip der Distanzansicht wird dabei auch von den s.g. Objektübersichtsseiten der Online Sammlungen ermöglicht. Sie bilden Zusammenstellungen der digitalen Kunstobjekte als Thumbnails unter Ähnlichkeitskriterien wie „Gattung“, „Zeitlichkeit“, „Künstler*in“ oder „Bildmotiv“ und dienen deren Vergleichbarkeit.

Als Untersuchungsgegenstand des Workshops sollen „explorative“ Online Sammlungen von Museen dienen. Kunstmuseen wie das Städel Museum Frankfurt a.M., das Museum für Kunst und Gewerbe Hamburg, die Kunsthalle Mannheim oder das Lenbachhaus aus dem deutschen sowie die Tate Gallery, das Rijksmuseum, das Museo del Prado oder das Belvedere Wien aus dem europäischen Raum sind hier als Fallbeispiele zu nennen. Deren „Explorationspotential“ soll im Workshop praktisch anhand von Methoden der Heuristischen Usability Evaluation sowie Eyetracking-Verfahren (s.u.) erfahrbar gemacht und analysiert werden. Die Gestaltung der Webseiten entscheidet – in Kombination mit Vorwissen, Motivation etc. der Nutzer*innen – über die Qualität der Wahrnehmungs- und Interpretationsprozesse: Das Interface Design unterliegt dabei hinsichtlich seiner psychologischen Wahrnehmung Gestaltungsgesetzen der Ähnlichkeit, Nähe oder Distanz, die über Gruppierungstendenzen der Elemente entscheiden. So verdeutlicht das Design z.B. anhand ähnlicher Farbgestaltungen oder räumlicher Nähe der Bestandteile der GUI inhaltliche Zusammenhänge (Thesmann 2016: 225ff.; Vgl. Detel 2014: 35ff.). Die Gestaltungsgesetze folgen Mustern der Rezeption. Es han-

delt sich um Konzepte, die im User Interface Design praktisch umgesetzt werden.

Ziel

Das Ziel des Workshops ist es, gemeinsam mit den Teilnehmer*innen auf Online Sammlungen von Museen als Repräsentationsformate des kulturellen Erbes kritisch zu schauen, sie zu evaluieren und mögliche Lücken der User Experience zu erkunden. Nach einer Einführung zum Gegenstand der explorativen Sammlung sowie zur Heuristischen Evaluation von GUIs nach Jakob Nielsen, folgt ein aktiver Erkundungsprozess: Unter Einbezug bildschirmbasierter Eyetracking-Verfahren können die Teilnehmer*innen eigene (Blick)Bewegungen durch die Sammlungen aufzeichnen und analysieren.

Fragen, die den Workshop leiten, sind: Inwiefern unterstützen die Gestaltung und das Navigationssystem „explorativer“ musealer Online Sammlungen das Prinzip des Browsens (Vgl. Whitelaw 2015) und Flanierens (Dörk et al. 2011)? An welchen Stellen entsteht Frustration seitens der Nutzer*innen, z.B. durch virtuelle Sackgassen, unzureichende Informationen oder eine mangelhafte formal-ästhetische Umsetzung? Welche Typen an verlinkten Daten werden von den Besucher*innen vorrangig wahrgenommen, welche erhalten keine oder wenig Aufmerksamkeit? Welchen Einfluss haben dabei die spezifischen Interface-Elemente (GUI-Patterns), z.B. Startseite, Objektübersichtsseite und Einzelobjektseite einer Online Sammlung? Wie lange halten sich Rezipient*innen auf den Webseiten auf? Welche Zeichenelemente werden am intensivsten wahrgenommen? Welchen Spielraum zwischen „offenem“ Umherschweifen und Führung lassen die GUIs zu?

Methodik

Als methodische Herangehensweise werden verschiedene Ansätze gewählt, die sich zum einen auf die Evaluation von Wahrnehmungsstrukturen, Aufmerksamkeitsverteilungen und Navigationsprozessen (= Usability) beziehen, zum anderen emotionale Aspekte der Nutzung und die Bewertung der visuellen Ästhetik (= User Experience⁶) in den Fokus nehmen. Obwohl der Prozess der Konzeption und Gestaltung von User Interfaces auf ein standardisiertes Set von Methoden des Human-Centred Design zurückgreift, ist das konkrete Artefakt – ob als Website, Mobile App oder interaktives Exponat – einer Vielzahl von systemischen Fehlerquellen hinsichtlich Usability und User Experience unterworfen. Zur Erreichung einer hohen User Experience stehen eine Vielzahl von Evaluationsverfahren – Expert*innen-Methoden (z.B. Heuristische Evaluationen) sowie Verfahren mit Nutzer*innenbeteiligung (z.B. Usability Tests, ggf. mit Eyetracking) – zur Verfügung.

Zwei dieser Verfahren werden in Kleingruppen in Hands-On-Sessions zum Einsatz kommen, darunter die Heuristische Expert*innen-Evaluationen nach Jakob Nielsen (Vgl. Nielsen (1994) [2020]) sowie das Eyetracking-Verfahren.

Heuristische Evaluationen umfassen ein generisches Set von Usability und User Experience Kriterien, die auf bestimmte Problemkategorien bei der Gestaltung von User Interfaces hinweisen. Sie ermöglichen beispielsweise eine schnelle Bewertung von Interaktions- und Navigationselementen hinsichtlich Erwartungskonformität, Selbstbeschreibungsfähigkeit und Konsistenz. Das Ziel einer Heuristischen Evaluation ist es, möglichst vollständig Usability- und User Experience-Probleme einer interaktiven Anwendung aufzudecken. Im Workshop dient die Heuristik von

Nielsen lediglich als Ausgangspunkt. Sie soll gemeinsam mit den Teilnehmer*innen im Hinblick auf domainspezifische Anforderungen für museale Online Sammlungen erweitert werden.

Als sinnvolle Ergänzung heuristischer Verfahren und beispielhafte Methode für Usability Evaluationen mit Nutzer*innenbeteiligung werden Eyetracking-Verfahren eingeführt. Mit ihnen können Blickbewegungen (Sakkaden) und Fixationen bei der Exploration und Aufgabebearbeitung – z.B. auf einer Website – erfasst und analysiert werden. Auf diese Weise lassen sich u.a. folgende Fragen beantworten:

- Auffälligkeit: Wird ein Objekt als solches überhaupt wahrgenommen?
- Betrachtungsdauer: Wie lange wird ein Objekt insgesamt wahrgenommen?
- Fixationsorte: Welche Stellen des Objektes werden fixiert?
- Fixationsreihenfolge: In welcher Reihenfolge werden die verschiedenen Objekte/Regionen fixiert?

Über die Visualisierung der Fixationspfade und -dauer werden auf diese Weise Schwierigkeiten bei der Exploration und Navigation sichtbar gemacht. Daraus lassen sich z.B. Rückschlüsse über die Anordnung und Gestaltung der Elemente auf einer Website ziehen.

Im Workshop wird die aktuelle Generation miniaturisierter Remote Eyetracker vorgestellt und eingesetzt. Sie ermöglicht die schnelle Kalibrierung von Proband*innen sowie die anwendungsfreundliche Aufzeichnung, Analyse und Visualisierung von Fixationsmustern (Vgl. Pelz 2019).

Format

Der Workshop richtet sich an alle Interessierte, insbesondere der Digitalen Geisteswissenschaften, des (digitalen) Museumswesens, des Informationsdesigns und der digitalen Kunstgeschichte. Vorwissen wird nicht benötigt.

Es ist ein Halbtages-Workshop (4h). Maximale Teilnehmer*innenzahl: 20.

Workshopleiter*innen

Janna Kienbaum

Universitätsbibliothek
Wissenschaftliche Mitarbeiterin DFG-Projekt FDNext
Campus II - Golm
+49 331/977-230130
jkienbau@uni-potsdam.de

Janna Kienbaum ist studierte Kulturwissenschaftlerin. Ihren Bachelorabschluss der Kulturwissenschaft und Italienischen Philologie absolvierte sie an der Universität Potsdam, worauf ein Masterstudium der Kulturwissenschaft an der Humboldt-Universität zu Berlin folgte. Sie ist als wissenschaftliche Mitarbeiterin in dem DFG-Projekt "FDNext" zum Ausbau des Forschungsdatenmanagements an der Universität Potsdam tätig. Zuvor arbeitete sie von 2017 und 2020 am Institut für Künste und Medien in dem Mixed-Methods-Projekt "New potentials for analyzing networked images", gefördert durch die Volkswagen Stiftung. Ihr Forschungsschwerpunkt kreist um Fragen zur digitalen Museumsarbeit und digitalen Bildwissenschaft bzw. Kunstgeschichte. Sie ist aktives Mitglied im "Netzwerk für digitales Geisteswissenschaften" der Universität Potsdam. In ihrem Promotionsvorhaben untersucht Janna Kienbaum die musealen Möglichkeiten und

Grenzen einer digitalen Kunstvermittlung angesichts interaktiv gestalteter Online Sammlungen im Web. Die Untersuchung erfolgt anhand einer theoretischen Fundierung über die Perspektive der Diagrammatik.

Sarah Kreiseler

Leuphana Universität
Institut für Philosophie und Kunstwissenschaft
Universitätsallee 1
21335 Lüneburg
sarah.kreiseler@leuphana.de

Sarah Kreiseler ist Kultur-, Medienwissenschaftlerin und Kuratorin. Sie war wissenschaftliche Mitarbeiterin im Programm „Promovieren im Museum“ an der Leuphana Universität Lüneburg, das u.a. in Kooperation mit dem Museum für Kunst und Gewerbe Hamburg umgesetzt wurde. Sie erschloss historische Kunstreproduktionsfotografien (Glasnegative) des ersten Museumsmitarbeiters, untersuchte deren Funktionsspektrum und Einfluss innerhalb eines Archiv- und Museumgefüges. Abschließend kuratierte sie in Zusammenarbeit mit Dr. Esther Ruelfs die Ausstellung „Das zweite Original. Fotografie neu ordnen: Reproduktionen“, u.a. mit der in Zusammenarbeit mit dem UCLAB Potsdam entstandenen digitalen Visualisierung namens "Close-Up Cloud". Sie hat einen M.A. in Europäischer Medienwissenschaft (Universität Potsdam) und beendet derzeit ihre objektbasierte Dissertation.

Frank Heidmann

Fachhochschule Potsdam
Fachbereich Design
Kiepenheuerallee 5
14469 Potsdam
frank.heidmann@fh-potsdam.de

Frank Heidmann (Dr. rer. nat.), ist seit 2005 Professor für das Themenfeld „Design of Software Interfaces“ im Studiengang Interface Design an der Fachhochschule Potsdam. Nach dem Studium der Angewandten Physischen Geographie an der Universität Trier war er Leiter des Competence Center „Human-Computer Interaction“ am Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO) in Stuttgart. Neben der Lehrtätigkeit leitet Frank Heidmann das "IDL // Interaction Design Lab", einen Dienstleister im Bereich der anwendungsorientierten Forschung und Entwicklung des Studienganges Interface Design. Es unterstützt Unternehmen und öffentliche Institutionen bei der Planung, Einführung und Umsetzung innovativer interaktiver, digitaler Produkte und Systeme. Seine Interessen in Forschung und Lehre umfassen neben der partizipativen Gestaltung und Evaluation von Benutzungsschnittstellen, die Visualisierung raumbezogener Daten (Geovisualisierung), sowie die Frage wie Informations- und Kommunikationstechnologien individuelle Einstellungs- und Verhaltensänderungen hinsichtlich nachhaltiger Lebensstile fördern können.

Fußnoten

1. Seit 2015 ist eine breite Welle an Digitalisierungsstrategien und -papieren der Museumshäuser und Kultureinrichtungen wahrzunehmen. Vgl. u.a. das Verbundprojekt „Museum4punkt0“. Entwickelt werden seit Mai 2017 „digitale Prototypen, um neue Formen der Kommunikation, Partizipation, Bildung und Vermittlung in Museen zu ermöglichen“: <http://www.museum4punkt0.de/> [letzter Zugriff 13. Juli 2021]. Erwähnenswert ist ebenso das Förderprojekt der Landesregierung Baden-Württemberg „Digitale Wege ins Museum I bzw. II“, das seit Oktober 2017 „die Entwicklung innovativer digitaler Vermittlungs-

programme in sechs Landesmuseen und dem Zentrum für Kunst und Medien Karlsruhe (ZKM)“ für die nächsten zwei Jahre unterstützt: <https://mwk.baden-wuerttemberg.de/de/service/presse/pressemitteilung/pid/kunstministerium-foerdert-digitale-wege-ins-museum/> sowie <https://www.mfg.de/ueber-die-mfg/portfolio/detailansicht/93-digitale-wege-ins-museum-ii/> [letzter Zugriff 13. Juli 2021].

2. Vgl. u.a. die Studie von dem Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO und dem Innovationsverbund »Future Museum«. Angesichts der Corona-Pandemie wurden innovative virtuelle Formate der Museumsvermittlung erfragt und erste Ergebnisse nun publiziert. Die Umfrage ergab u.a., dass „zwei Drittel der Befragten bereits an virtuellen Museumsbesuchen teilgenommen haben, jedoch nur 35 Prozent diese als zufriedenstellend beschreiben würden.“ <https://www.iao.fraunhofer.de/de/presse-und-medien/aktuelles/ist-der-virtuelle-museumsbesuch-zukunftsaehig.html> [letzter Zugriff 13. Juli 2021].

3. Das Konzept der „digitalen Vermittlung“ bezieht sich in erster Linie auf die vielfältigen Zugangsmöglichkeiten zum Objektbestand im digitalen Raum und wird z.B. in der „Digitalen Strategie“ des Städel Museums als „neuartige, umfassende Wissensvermittlung [...], die verstärkt auf interaktive, partizipative und narrative Elemente setzt“, beschrieben. <https://www.staedelmuseum.de/de/digitale-strategie> [letzter Zugriff 13. Juli 2021].

4. Als Beispiel kann der „Vikus Viewer“ genannt werden: <https://vikusviewer.fh-potsdam.de/> [letzter Zugriff 13. Juli 2021].

5. Gegenüber einer reinen Digitalisierung von Museumsobjekten im Sinne einer „digitalisierten Kunstgeschichte“ spricht Johanna Drucker hier von der digitalen Kunstgeschichte: „But a clear distinction has to be made between the use of online repositories and images, which is digitized art history, and the use of analytic techniques enabled by computational technology that is the proper domain of digital art history“ (Drucker 2013: 7; Vgl. Zweig 2015).

6. Mit dem Begriff der User Experience (UX) wird das Nutzungserleben als die Wahrnehmung und Bewertung einer Person beschrieben, die sich vor, während und nach der Benutzung eines interaktiven Systems ergeben. Im Unterschied zur Usability ist User Experience der deutlich weiter gefasste Begriff, der sich nicht allein auf die Effektivität und Effizienz der Aufgabenerfüllung beschränkt, sondern explizit hedonische und pragmatische Qualitätsaspekte des User Interface umfasst (Vgl. Diefenbach / Hassenzähl 2017: 8).

Bibliographie

- Brüggemann, Viktoria / Dörk, Marian / Kreiseler, Sarah** (2016): „Museale Bestände im Web: Eine Untersuchung von acht digitalen Sammlungen“ in *EVA Berlin: Proceedings of the Electronic Media and Visual Arts conference*: 227-236 <http://mariandoerk.de/papers/evaberlin2016.pdf> [letzter Zugriff 12. Juli 2021].
- Detel, Wolfgang** (2014): *Erkenntnis- und Wissenschaftstheorie, Grundkurs Philosophie, Bd.4*. Ditzingen: Reclam.
- Diefenbach, Sarah / Hassenzähl, Marc** (2017): *Psychologie in der nutzerzentrierten Produktgestaltung: Mensch-Technik-Interaktion-Erlebnis*. Berlin: Springer.
- Dörk, Marian et al.** (2011): „The Information Flaneur: A Fresh Look at Information Seeking“, in: *CHI 2011: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, 1215-1224. <https://mariandoerk.de/information-flaneur/chi2011.pdf> [letzter Zugriff 12. Juli 2021].
- Dörk, Marian** (2017): „One view is not enough: High-level visualizations of a large cultural collection“, in: *Information Design Journal* 23(1) © 2017 John Benjamins Publishing Company. <http://mariandoerk.de/papers/idj2017.pdf> [letzter Zugriff 12. Juli 2021].
- Drucker, Johanna** (2013): „Is There a ‘Digital’ Art History?“, in: *Visual Resources* 29 (1-2): 5-13, DOI: 10.1080/01973762.2013.761106.
- Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO Presseinformation** (2021): Ist der virtuelle Museumsbesuch zukunftsfähig? <https://www.iao.fraunhofer.de/de/presse-und-medien/aktuelles/ist-der-virtuelle-museumsbesuch-zukunftsaehig.html> [letzter Zugriff 13. Juli 2021].
- Glinka, Katrin et al.** (2016): „Von sammlungs-spezifischen Visualisierungen zu nachnutzbaren Werkzeugen“, in: *DHQ: Digital Humanities Quarterly*. 11:2, 204-207. <https://uclab.fh-potsdam.de/wp/wp-content/uploads/dhd2017-aus-programm.pdf> [letzter Zugriff 13. Juli 2021].
- Krämer, Harald** (2001): *Museumsinformatik und Digitale Sammlung*. Wien: WUV | Universitätsverlag.
- Kreiseler, Sarah et al.** (2017): „Tracing exploratory modes in digital collections of museum Web sites using reverse information architecture“, in: *First Monday*. 22:4, <https://firstmonday.org/ojs/index.php/fm/article/view/6984/6090> [letzter Zugriff 13. Juli 2021].
- Nielson, Jakob** (1994) [2020]: „10 Usability Heuristics for User Interface Design“, in: *NN/g Nielsen Norman Group*. <https://www.nngroup.com/articles/ten-usability-heuristics/> [letzter Zugriff 13. Juli 2021].
- Pelz, Jeff B.** (2019): „Eyetracking Research“, in: Edlund, J.E. & Nichols, A.L. (eds.): *Advanced Research Methods for the Social and Behavioral Sciences*. Cambridge: Cambridge University Press 168-190.
- Thesmann, Stefan** (2016): *Usability, User Experience und Accessibility im Web gestalten*. Wiesbaden: Springer Vieweg.
- Schweibenz, Werner** (2020): „Wenn das Ding digital ist...Überlegungen zum Verhältnis von Objekt und Digitalisat“, in: Andraschke, Udo / Wagner, Sarah (eds.): *Objekte im Netz: Wissenschaftliche Sammlungen im digitalen Wandel*. Bielefeld: transcript 15-27.
- Städelmuseum**: Digitale Strategie <https://www.staedelmuseum.de/de/digitale-strategie> [letzter Zugriff 13. Juli 2021].
- Whitelaw, Mitchell** (2015): „Generous Interfaces for Digital Cultural Collections“, in: *Digital Humanities Quarterly* 9, 1 <http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html> [letzter Zugriff 30. Juni 2021].
- Zweig, Benjamin** (2015): „Forgotten Genealogies: Brief Reflections on the History of Digital Art History“, in: *International Journal for Digital Art History*, no. 1 (June). <https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/21633/15405> [letzter Zugriff 12. Juli 2021].

Index der Autorinnen und Autoren

Achmann, Michael	349, 363
Adamczak, Katarzyna	296
Agt-Rickauer, Henning	29
Ahmed, Sajawel	323
Akkermann, Miriam	43
Alschner, Stefan	14
Altenhöner, Reinhard	19
Althage, Melanie	241, 338
Amrhein, Kilian	362
Andorfer, Peter	333
Andresen, Melanie	186, 408
Arndt, Nadine	24
Arnold, Eckhart	396
Arnold, Frederik	162
Arnold, Matthias	268
Asabidi, Ruslan	349
Baddack, Cornelia	24
Baedke, Jan	360
Baglatzi, Alkyoni	334
Baierer, Konstantin	86, 381
Baillet, Anne	405
Balck, Sandra	117, 220
Barbot, Laure	72
Barth, Florian	340
Barzen, Johanna	329
Bateman, John	17
Bauer, Bernhard	335
Baumgarten, Marcus	14
Büdenbender, Stefan	326
Beck, Julia	325
Bentz, Isabelle	9
Beretta, Francesco	387
Berg, Mia	297
Bernhart, Toni	124
Bühler, Fabian	329
Böhm, Alexander	360
Bigalke, Jan	344
Bischoff, Eva	262
Bläß, Sandra	54, 120
Bleier, Roman	333
Blessing, André	314
Blessing, Andre	264
Blümel, Ina	266
Bludau, Mark-Jan	216, 220
Bönisch, Dominik	356
Boenig, Matthias	86
Borgards, Roland	82
Borges, Rebekka	283
Brandl, Stephanie	405
Brüggemann, Viktoria	216
Börner, Ingo	272, 373
Brottrager, Judith	244
Bruschke, Jonas	179
Bunout, Estelle	308
Burghardt, Manuel	21, 82, 113, 393
Burkhard, Fabienne	288
Busch, Anna	11, 259, 280, 347
Cantera, Alberto	278

Charvat, Vera Maria	272
Clematide, Simon	308
Colditz, Iris	278
Contreras Saiz, Mónica	252
Cremer, Fabian	347, 397
Czmiel, Alexander	21, 256
Debbeler, Anke	283
Dennerlein, Katrin	93, 107
Diecke, Josephine	17
Dieckmann, Lisa	19, 21, 393
Dietz, Katharina	320
Dietz, Nadine	326
Dönicke, Tillmann	340
Dogunke, Swantje	397
Drach, Sviatoslav	344
Dreyer, Malte	338
Düring, Marten	308
Dörk, Marian	216, 220
Du, Keli	104, 316
Dudar, Julia	316
Dumont, Stefan	359
Dunkelmann, Lena	326
Dziudzia, Corinna	402
Ehrmann, Maud	308
Eide, Øyvind	278
Elwert, Frederik	267
Emanuel, Chagai	278
Engel, Alexander	37
Engl, Elisabeth	86
Erler-Fridgen, Katharina	320
Ernst, Felix	75
Ewerth, Ralph	17, 143
Fangerau, Heiner	220
Fábregas Tejeda, Alejandro	361
Feldmüller, Tim	239
Fetz, Bernhard	11
Fickers, Andreas	308
Fiechter, Benjamin	162
Fischer, Frank	72, 89, 373
Fischer-Nebmaier, Wladimir	24, 353
Fleischmann, Florian	238
Flüh, Marie	120, 155
Flueh, Marie	54
Frank, Markus	400
Franke, Claus	273
Franken, Lina	101
Freyberg, Linda	78
Fritze, Christiane	391
Gödeke, Luisa	340
Gebhard, Henning	306
Geestmann, Mareen	86
Genzel, Kristina	280
Gerstenberg, Annette	371
Gerstorfer, Dominik	120, 160, 307, 384
Gfrereis, Heike	9
Göggelmann, Michael	40
Gius, Evelyn	120, 147
Glawion, Anastasia	199
Gleixner, Sebastian	24
Gnau-Franké, Birte	326
Goldberg, Jan Michael	47
Grabsch, Sascha	359
Gradl, Tobias	60
Gray, Edward	72

Grebe, Anja	317	Klemstein, Franziska	351
Grisot, Giulia	166	Klindt, Marco	362
Götzelmann, Germaine	75	Knecht, David	387
Guescini, Rolf	338	Knierim, Aenne	363
Guhr, Svenja	21, 393	Kohle, Hubertus	143
Guido, Daniele	308	Konle, Leonard	126
Gutiérrez De la Torre, Silvia Eunice	247	Korwisi, Kristof	358
Hadden, Richard	337	Krautter, Benjamin	186, 408
Hagen, Thora	209	Kreß, Hannah	138
Hall, Mark	402	Kreiseler, Sarah	411
Halling, Thorsten	220	Kreuzmair, Elias	251
Hampel, Lisa	349	Krewet, Michael	75
Hannessschläger, Vanessa	282	Krüger, Katharina	358
Heßbrüggen-Walter, Stefan	57	Krömer, Cora	405
Heckelen, Malte	288	Kröncke, Merten	126
Heftberger, Adelheid	17	Kroeber, Cindy	179
Hegel, Philipp	75	Krug, Markus	34
Heidmann, Frank	411	Kruse, Carl	323
Heilmann, Juliane	280	Köster, Jan	273
Heiniger, Anna Katharina	40	Kuper, Heinz-Günter	362
Helling, Patrick	193, 283, 286	Kurek, Sarah	65
Henke, Konstantin	268	Kurz, Stephan	24, 353
Henninger, Christine	325	Kuzman-Šlogar, Koraljka	282
Henny-Krahmer, Ulrike	183, 203, 313	Lampert, Marcus	389
Herrmann, J. Berenike	166, 354	Lang, Sarah	175
Hess, Jan	314	Lange, Felix	362
High-Steskal, Nicole	229, 317	Langer, Lars	82
Hildenbrandt, Vera	9	Langhanki, Florian	381
Hiltmann, Torsten	338	Lassner, David	405
Hinrichsen, Lena	86, 381	Laubrock, Jochen	98
Hinzmann, Maria	170, 320	Lüdeling, Anke	338
Hüllermeier, Eyke	143	Lecroq, Axelle	300
Hofmann, Anna Mareike	340	Leh, Almut	371
Holly, Eva Maria	220	Lein, Richard	353
Hoppe, Stephan	179	Lemaire, Marina	376
Horstmann, Jan	14	Lemitz, Bastian	138
Horvath, Aliz	345	Lepper, Marcel	11
Hotz, Gerhard	387	Leymann, Frank	329
Howanitz, Gernot	17	Leyrer, Katharina	245
Illmayer, Klaus	72, 325	Liem, Johannes	317
Imeri, Sabine	376	Liesenfeld, Nina	326
Jacke, Janina	21	Loertscher, Miriam	17
Jander, Melina	378	Lordick, Harald	347
Jannidis, Fotis	126	Lorenz, Andrea	297
Jettka, Daniel	203	Mahmutovic, Edin	323
Jügel, Thomas	278	Maiwald, Ferdinand	179
Jüngerkes, Sven	24	Maus, David	54
Jung, Kerstin	193, 274, 314	Mayr, Eva	9, 317
Jurish, Bryan	289	Möbus, Dennis	371
Kacsuk, Zoltan	151	Meding, Holle Ameriga	252
Kalmer, Silke	294	Meier-Vieracker, Simon	251
Kalyakin, Roman	308	Meister, Malte	120, 307, 384
Kamocki, Pawel	282	Menzel, Sina	117
Kampkaspar, Dario	294	Messerli, Thomas	354
Kaplan, Frederic	309	Meyer, Peter	342
Katsiadakis, Helen	345	Milling, Carsten	373
Keck, Jana	264	Mischke, Dennis	131, 347
Kelm, Holden	298	Müller, Anja	362
Köhring, Esther	82	Müller, Christiane	14
Kienbaum, Janna	411	Möller, Klaus-Peter	280
Kindler, Sebastian	277	Müller-Laackman, Jonas	359
Klaes, Jan Sebastian	358	Müller, Melissa	327
Klappenbach, Lou	291, 298	Müller, Sophie	294
Klee, Anne	170, 320	Münzmay, Andreas	19

Moeller, Katrin	47, 376	Saric, Sanja	275
Mollenhauer, Elisabeth	303	Schöch, Christof	170, 316, 320
Mondaca, Francisco	278	Scheel, Christian	289
Mrugalski, Michał	272	Scheltjens, Werner	350
Muenster, Sander	179	Schenk, Torsten	75
Muessemann, Hannah	252	Scherer, Thomas	29
Nantke, Julia	14, 54, 121	Schlesinger, Claus-Michael	288
Neuber, Frederike	256, 313	Schlögl, Matthias	333, 337
Neubert, Anna	398	Schlieder, Christoph	350
Neudecker, Clemens	86	Schlomske-Bodenstein, Nadine	329
Neuefeind, Claes	278, 344	Schmidt, David	34
Nicholson, Daniel J.	361	Schmidt, Thomas	65, 93, 107
Niebling, Florian	179	Schmunk, Stefan	326
Nieländer, Maret	289	Schnaitter, Hannes	117, 220
Nowicki, Anna-Lena	362	Schneider, Philipp	242
Odebrecht, Carolin	272, 338	Schneider, Sophie	50
Offenbert, Eva	9	Schneider, Stefanie	142, 225
Pagel, Janis	186, 408	Scholger, Martina	313, 341
Pagenstecher, Cord	371	Scholger, Walter	21, 282, 393
Pattee, Aaron	179	Schroeder, Paul	309
Pöckelmann, Marcus	253	Schrott, Maximilian	24
Pestov, Paul	86	Schubert, Zoe	266
Peter, Ulrike	273	Schulz, Christoph	19
Petras, Vivien	117	Schumacher, Mareike	121, 155, 232, 384
Pfeffer, Magnus	151	Schwarz, Anja	262
Pichler, Axel	408	Seifert, Sabine	280
Pielström, Steffen	194, 274	Seltmann, Melanie E.-H.	294, 326
Pietsch, Andreas	138	Serif, Ina	37
Pietsch, Christopher	216	Seung-Bin, Yim	72
Plakidis, Melina	220	Shaked, Shaul	278
Poetis, Panoria	225	Siebold, Anna	248
Pons, Jessie	267	Sikora, Uwe	138
Pratschke, Margarete	19	Sluyter-Gäthje, Henny	131, 190, 373
Primavesi, Patrick	19	Smiatek, Katharina	225
Puppe, Frank	34	Spence, Paul	345
Purschwitz, Anne	212	Spiegel, Simon	17
Radisch, Erik	367	Springstein, Matthias	142
Radmacher, Emilia	225	Söring, Sibylle	75, 376
Rahnama, Javad	143	Stadler, Peter	183
Rainer, Simon	367	Stallmann, Marco	138
Rapp, Andrea	326	Standl, Bernhard	329
Rath, Brigitte	331	Stäcker, Thomas	326
Rau, Felix	286, 303	Steffes, Moritz	320
Rebiger, Bill	253	Stegmeier, Jörn	294
Rebora, Simone	354	Steiner, Elisabeth	275, 341
Rehm, Georg	220	Stellmacher, Martha	19
Rehman, Misbahur	323	Steyer, Timo	347, 393
Reichert, Nils	376	Stratil, Jasper	29
Reimann, Anna	37	Ströbel, Philip	309
Reiners-Selbach, Stefan	360	Testori, Marinella	345
Reiter, Nils	21, 40, 186, 393	Tiefenbacher, Sara	325
Rettinghaus, Klaus	293	Tischlik, Joshua	323
Reul, Christian	134, 358, 381	Tomasek, Stefan	134
Rezania, Kianoosh	278	Tonne, Danah	75
Richter, Sandra	11	Towara, Nadine	358
Richts-Matthaei, Kristina	19	Trilcke, Peer	11, 131, 190, 280
Rißler-Pipka, Nanette	369	Tropper, Eva	9
Roeder, Torsten	259, 369	Tseng, Chiao-I	98
Rok, Cora	316	Tóth-Czifra, Erzsébet	345
Romanello, Matteo	309	Tu, Ngoc Duyen Tanja	342
Rossenova, Lozana	265	Utescher, Ronja	179
Roth, Martin	151	Varachkina, Hanna	340
Röttgermann, Julia	170, 320	Vasold, Gunter	337, 341
Röwenstrunk, Daniel	19	Vauth, Michael	147, 307, 384

Velissaropoulos, Georgios	334
Voß, Franziska	325
Vock, Richard	266
Vogeler, Georg	337
Vogeltanz, Maximilian	275
Volk, Martin	309
Walkowski, Niels-Oliver	113
Wübbena, Thorsten	14, 398
Wehner, Maximilian	134, 381
Weidling, Michelle	86
Weimer, Lukas	378
Weis, Joëlle	14
Werner, Stephanie	326
Werthmann, Antonina	60
Wieloch, Jasmin	220
Wieneke, Lars	309
Wierzoch, Jan	291
Windhager, Florian	9, 317
Winko, Simone	126
Wirth, Christian	82
Wirtz Eybl, Irmgard	11
Wittenbecher, Maxim	325
Wünsch, Lukas	138
Wolf, Katrin	277
Wolff, Christian	93, 107, 349, 363
Wunsch, Kevin	294
Wuttke, Ulrike	21, 345, 393
Wyss, Eva L.	326
Zaagsma, Gerben	207
Zarei, Alireza	72
Zarriess, Sina	179
Zedlitz, Jesper	212
Zeini, Arash	278
Zinck, Josefine	117
Zirker, Angelika	40
Çakir, Dılan Canan	89
de Günther, Sabine	78
Štanzel, Arnošt	296
Đurčo, Matej	72, 272
van Beek, Thijs	309
von Hindenburg, Barbara	24
Özsoy, Ömer	323

