

DHd2023

OPEN HUMANITIES

OPEN CULTURE



KONFERENZABSTRACTS



13.-14. MÄRZ
BELVAL
LUXEMBURG

15.-17. MÄRZ
TRIER
DEUTSCHLAND

DHd2023

OPEN HUMANITIES

OPEN CULTURE



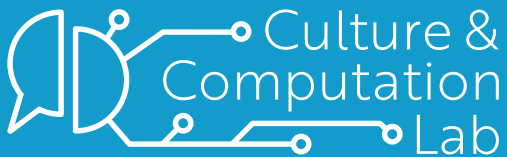
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY



UNIVERSITÉ DU
LUXEMBOURG

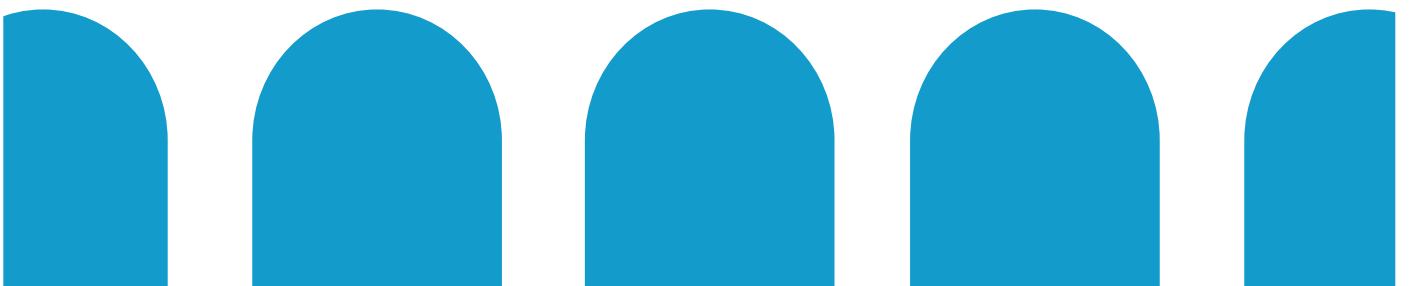


digital humanities im
deutschsprachigen raum



Kompetenzzentrum

Trier Center for Digital Humanities



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

DHd2023: Open Humanities, Open Culture

Universität Luxemburg & Universität Trier

13. bis 17. März 2023

Partner und Sponsoren

GERDA HENKEL STIFTUNG



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.

Herausgeber:innen: Anna Busch, Peer Trilcke
Redaktion und Korrektur der Auszeichnungen:
Alistair Plum, Vivien Wolter, Hendrik Chudoba, Joëlle Weis
Konvertierung TEI nach PDF: Alistair Plum
<https://github.com/plumaj/DHd2023-BoA>

Historie der Autorinnen und Autoren sowie Versionen der Konversionsskripte:

Ingo Börner (2022)
https://github.com/ingoboerner/dhd2022_boa
Nina Seemann (2020)
<https://github.com/NinaSeemann/DHd2020-BoA>
Attila Klett (2019)
<https://github.com/texttechnologylab/DHd2019BoA>
Claes Neuefeind (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 und Gestaltung des Covers: Julia Hennemann (Trier Center for Digital Humanities)
Gestaltung der Konferenz-Website: Kirill Mitsurov (Luxembourg Centre for Contemporary and Digital History)

Luxemburg und Trier 2023
Online verfügbar: <https://doi.org/10.5281/zenodo.7688632>

Vorwort

Dieses Jahr steht die 9. Jahrestagung des Verbands „Digital Humanities im deutschsprachigen Raum“ unter dem Motto „Open Humanities, Open Culture“ und wird gemeinsam von der Université du Luxembourg und der Universität Trier ausgerichtet. Verschiedene Tagungsbeiträge greifen zentrale Aspekte des Tagungsmottos der DHd2023 auf, denn Offenheit berührt alle Aspekte der wissenschaftlichen Praxis, sei es im Bereich der Daten (Open Data), der Software (Open Source), der Publikationen (Open Access), der Lehrmaterialien (Open Educational Resources), des kulturellen Gedächtnisses (Open Heritage) oder des Verhältnisses zur Gesellschaft (Public Humanities).

Das Programmkomitee der DHd2023 stand vor der Herausforderung, aus 174 Einreichungen eine Auswahl zu treffen, was bei der hohen Qualität der Beiträge nicht einfach war. Ohne die Unterstützung von 165 Gutachter:innen, die insgesamt 643 Gutachten erstellt haben, wäre das nicht möglich gewesen. Mit dem nun vorliegenden Tagungsprogramm und diesem Book of Abstracts hoffen wir, Ihnen im Jahr 2023 wieder facettenreiche Einblicke in die aktuelle Forschung der digitalen Geisteswissenschaften bieten zu können.

Zur wissenschaftlichen Qualitätssicherung wurde der zweistufige Begutachtungsprozess beibehalten. Wie bereits im Jahr zuvor, wurde nach dem sogenannten Open-Peer-Review-Verfahren begutachtet, bei dem die Namen der Gutachter:innen den Autor:innen offengelegt werden. Neben der fachlichen Zuordnung bei der Begutachtung achtete das Programmkomitee auch auf das akademische Alter der Einreichenden und Begutachtenden. Die bei der letzten DHd-Konferenz eingeführte Rückmeldephase, in der die Gutachten von den Einreichenden kommentiert werden können, sodass die Begutachtenden im Anschluss darauf reagieren können, wurde beibehalten. Dieser Schritt soll den wissenschaftlichen Diskurs im Begutachtungsprozess fördern, zum Austausch zwischen Beiträger:innen und Gutachter:innen beitragen und das Verhältnis zwischen Gutachter:innen und Autor:innen damit weniger kompetitiv, stattdessen kommunikativer machen. Der Dialog zwischen Einreichenden und Begutachtenden trägt nicht nur dazu bei, die inhaltliche Qualität der Beiträge zu erhöhen und den Begutachtungsprozess transparenter zu machen, sondern unterstützt auch das Programmkomitee in der endgültigen Entscheidungsfindung. In vielen Fällen ist es zu einem intensiven Austausch über die eingereichten Texte, die Gutachten und die Erwidern gekommen – ein beeindruckender Community-Diskurs!

Die Arbeiten des Programmkomitees wurden zudem von der vom Verband eingesetzten Task Force „DHd Abstracts“ flankiert, die eine Dokumentation des Einreichungsprozesses der Konferenzbeiträge zu den DHd-Jahrestagungen vorgelegt hat, um den Beiträger:innen entsprechende Hilfestellung an die Hand zu geben.

Wir danken allen Autor:innen, die Beiträge für die DHd2023 eingereicht haben. Sie alle ermöglichen es uns, bei der Jahrestagung gemeinsam die Vielfalt und Qualität der Forschung in den digitalen Geisteswissenschaften aufzuzeigen. Dank gebührt natürlich auch den unermüdlichen Gutachter:innen. Auch wenn das Programmkomitee der DHd2023 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. Die Gutachten – sowohl in ihren ausformulierten Textteilen als auch in ihrer Punktevergabe – sind essentiell für die Programmherstellung. Die Arbeit der Gutachter:innen stellt die wissenschaftliche Qualität der eingereichten Beiträge sicher und erweist sich damit als zentraler Aspekt des Konferenz Erfolgs.

Zu den Mitgliedern des Programmkomitees zählen diesmal Anne Baillot, Noah Bubenhofer, Estelle Bunout (Koordination Luxemburg), Anna Busch (stellvertretender Vorsitz), Alexander Czmiel, Lisa Dieckmann, Evelyn Gius, Katrin Glinka, Andreas Henrich, Andreas Münzmay, Patrick Sahle, Martina Scholger, Silke Schwandt, Peer Trilcke (Vorsitz) und Joëlle Weis (Koordination Trier). Allen Kolleg:innen danken wir sehr herzlich für ihr Engagement. Unser persönlicher Dank gilt den lokalen Organisator:innen in Luxemburg und Trier: Sie haben die Arbeit des Programmkomitees entscheidend unterstützt.

Wir freuen uns auf eine bereichernde und „offene“ Konferenz.

Potsdam, im März 2023
Anna Busch und Peer Trilcke
für das Programmkomitee der DHd2023

Reviewer:innen der DHd2023

Review Award 2023

Hinzmann, Maria

Nominierte für den Review Award 2023

Brunner, Annelen

Elwert, Frederik

Glinka, Katrin

Helling, Patrick

Hertling, Anke

Heßbrüggen-Walter, Stefan

Hinzmann, Maria

Howanitz, Gernot

Illmayer, Klaus

Klammt, Anne

König, Mareike

Krautter, Benjamin

Kröger, Bärbel

Pfeffer, Magnus

Roller, Ramona

Steyer, Timo

Veit, Joachim

Wagner, Andreas

Weis, Joelle

Wuttke, Ulrike

Reviewer:innen

Achmann, Michael

Acquavella-Rauch, Stefanie

Akkermann, Miriam

Andresen, Melanie

Andrews, Tara

Arnold, Eckhart

Baillet, Anne

Barabucci, Gioele

Barzen, Johanna

Bernhart, Toni

Biber, Hanno

Blaschitz, Edith

Blumtritt, Jonathan

Börner, Ingo

Brunner, Annelen

Bubenhofer, Noah

Bunout, Estelle

Burch, Thomas

Burckhardt, Daniel

Bürgermeister, Martina

Busch, Anna

Busch, Hannah

Capelle, Irmlind

Casties, Robert

Cremer, Fabian

Czmiel, Alexander

Deicke, Aline

Dieckmann, Lisa

Du, Keli

Duan, Tinghui

Dunst, Alexander

Düring, Marten

Eggert, Lisa

Eide, Øyvind

Elwert, Frederik

Ernst, Thomas

Fechner, Martin

Fischer, Frank

Flüh, Marie

Frank, Markus

Franken, Lina

Freyberg, Linda

Fritze, Christiane

Gasser, Sonja

Geiger, Jonathan

Gengnagel, Tessa

Gerber, Anja

Gerstenberg, Annette

Gerstorfer, Dominik

Gius, Evelyn

Glawion, Anastasia

Glinka, Katrin

Gradl, Tobias

Grallert, Till

Grote, Brigitte

Guhr, Svenja

Gülden, Svenja A.

Hahn, Udo

Haider, Thomas Nikolaus

Hall, Mark

Hegel, Philipp

Helling, Patrick

Henny-Krahmer, Ulrike

Henrich, Andreas

Hermes, Jürgen

Hertling, Anke

Hess, Jan

Heßbrüggen-Walter, Stefan

Heyer, Gerhard

High-Steskal, Nicole	Mischke, Dennis	Serif, Ina
Hiltmann, Torsten	Molitor, Paul	Stadler, Peter
Hinzmann, Maria	Münzmay, Andreas	Staecker, Thomas
Hodel, Tobias	Nantke, Julia	Stede, Manfred
Hohmann, Georg	Nerbonne, John	Steyer, Timo
Homburg, Timo	Neuber, Frederike	Thomas, Christian
Horstmann, Jan	Neudecker, Clemens	Trilcke, Peer
Howanitz, Gernot	Neuefeind, Claes	Trippel, Thorsten
Illmayer, Klaus	Niekler, Andreas	Tu, Ngoc Duyen Tanja
Jannidis, Fotis	Nunn, Christopher	Veit, Joachim
Janz, Nina	Oberbichler, Sarah	Vertan, Cristina
Jäschke, Robert	Offert, Fabian	Viehhauser, Gabriel
Jung, Kerstin	Pfeffer, Magnus	Vogeler, Georg
Jünger, Jakob	Pielström, Steffen	Wagner, Andreas
Kampkaspar, Dario	Proisl, Thomas	Weis, Joelle
Keck, Jana	Puppe, Frank	Wettlaufer, Jörg
Kepper, Johannes	Rehm, Georg	Wieneke, Lars
Klammt, Anne	Reiners-Selbach, Stefan	Windhager, Florian
Klemstein, Franziska	Reiter, Nils	Wübbena, Thorsten
Kleymann, Rabea	Rißler-Pipka, Nanette	Wuttke, Ulrike
Klinke, Harald	Roeder, Torsten	Zaagsma, Gerben
Koch, Walter	Roller, Ramona	Zehe, Albin
Kocher, Ursula	Röwenstrunk, Daniel	
König, Mareike	Rüdiger, Jan Oliver	
Konle, Leonard	Sahle, Patrick	
Krautter, Benjamin	Schaßan, Torsten	
Kröger, Bärbel	Schmidt, Thomas	
Kurz, Stephan	Schmunk, Stefan	
Lang, Sarah	Schneider, Stefanie	
Langner, Martin	Schöch, Christof	
Leinen, Peter	Scholger, Walter	
Lemaire, Marina	Scholz, Martin	
Lüschow, Andreas	Schommer, Christoph	
Mandl, Thomas	Schumacher, Mareike	
Mayr, Eva	Schwandt, Silke	
Meier-Vieracker, Simon	Seifert, Sabine	
Meister, Jan Christoph	Seltmann, Melanie Elisabeth	

Inhaltsverzeichnis

Workshops

Algorithmen anwenden – algorithmisch denken. „Algorithmizität“ als Brücke zwischen Geisteswissenschaften und Informatik?	
Burghardt, Manuel; Geiger, Jonathan D.; Horstmann, Jan; Kleymann, Rabea; Schmitz, Jascha; Schwandt, Silke	13
Data Driven Storytelling zu kulturellen Objekten und Biographien	
Liem, Johannes; Kusnick, Jakob; Jänicke, Steffan; Doppler, Carina; Passecker, Markus; Mayr, Eva; Windhager, Florian	16
Data Feminism in DH: Hackathon und Netzwerktreffen	
Lang, Sarah; Borek, Luise; Probst, Nora	19
Die perfekte digitale Open-Access-Publikation	
Baum, Constanze; Dahnke, Michael; Dinger, Patrick; Fadeeva, Yuliya; Horstmann, Jan; Seltsmann, Melanie Elisabeth-H.; Steyer, Timo	22
3D- und 4D-Modellierung in den Digital Humanities. Eine praktische und theoretische Einführung in Blender	
Hunziker, Manuel; von Pippich, Waltraud; Rensinghoff, Berenike	25
Forschungssoftware rezensieren – Konzeption, Durchführung und Umsetzung	
Homburg, Timo; Klammt, Anne; Offert, Fabian; Thiery, Florian	27
GitMA oder CATMA für Fortgeschrittene	
Schumacher, Mareike; Meister, Malte; Gerstorfer, Dominik	30
Greening DH: individuelle Handlungsspielräume und institutionelle Perspektiven	
Baillet, Anne; Feidicker, Charlotte; Gerber, Anja; Roeder, Torsten	33
Hands-on-Workshop Datendokumentation	
Lemaire, Marina; Moeller, Katrin; Schulz, Julian; Söring, Sibylle; Wettlaufer, Jörg	37
Hands-on-Workshop Wissenschaftsbloggen mit de.hypotheses Halbtägiger Workshop	
König, Mareike	40
Offen für Professionalisierung? Wie Software und Entwickler*innen in den Digital Humanities gestärkt werden können	
Czmiel, Alexander; Henny-Krahmer, Ulrike; Jettka, Daniel	42
Pipelines für Natural Language Processing und digitale Literaturanalyse in spaCy	
Varachkina, Hanna; Barth, Florian; Dönicke, Tillmann; Biermann, Johannes; Altmann, Friederike; Neitzke, Thorben; Sporleder, Caroline	45
Semantic Web und Linked Open Data in den Geschichts-wissenschaften	
Kröger, Bärbel; Störko, Johanna; Wettlaufer, Jörg	48
Skalierbare Blicke auf Leben und Werk: Visuelle Analyse und Kuratierung von kulturellen Objekten und Künstler*innen-Biographien	
Windhager, Florian; Liem, Johannes; Mayr, Eva; Schlögl, Matthias; Ebel, Carla; Probst, Stefan; Beck, Samuel; Koch, Steffen	51
SPARQL für (digitale) Geisteswissenschaftler:innen – Querying Wikidata und die MiMoTextBase	
Hinzmann, Maria; Klee, Anne; Konstanciak, Johanna; Röttgermann, Julia; Schöch, Christof; Steffes, Moritz	54
»textklang« – Ein Mixed-Methods-Workshop zu Lyrik in Text und Ton	
Ketschik, Nora; Bernhart, Toni; Gärtner, Markus; Koch, Julia; Schaffler, Nadja; Kuhn, Jonas	58
Wunsch und Wirklichkeit – Forschungsinfrastrukturen in den Computational Literary Studies: interdisziplinär, modular, vernetzt?	
Jung, Kerstin; Helling, Patrick; Pielström, Steffen; Kababgi, Daniel	62

Panels

Digitalisierung kulturellen Erbes und postkoloniale Perspektiven	67
Forschungsdaten-infrastruktur als offene Werkstatt: Community Building zwischen generischen und datenspezifischen Praktiken	69
Herausforderung, Lesson Learned oder Chance? Der Zusammenhang zwischen Kulturen des Scheiterns und Open-Bewegungen in den Digital Humanities	72
Living Handbook “Digitale Quellenkritik”	74

Open DH? Mapping Blind Spots	77
Open Humanities in der Filmwissenschaft – zwischen Wunsch und Wirklichkeit	80
Opening Sources – modulare Wege zur Quellenbereitstellung und -edition	83

Vorträge

Algorithmen-gestützte Analyse visuell-materieller Eigenschaften von Briefen <i>Nantke, Julia; Reul, Christian; Bläß, Sandra; Flüh, Marie</i>	88
„Auch heute war die Stimmung im Allgemeinen fest.“ Zero-Shot Klassifikation zur Bestimmung des Media Sentiment an der Berliner Börse zwischen 1872 und 1930 <i>Wehrheim, Lino; Borst, Janos; Burghardt, Manuel; Niekler, Andreas</i>	90
Bilder im Kontext: Die Entwicklung des Corpus Vitrearum vom Bildarchiv zu Born-Digitals <i>Pittroff, Sarah; Gerber, Anja; Steller, Jonatan</i>	95
Bullingers Briefwechsel zugänglich machen: Stand der Handschriftenerkennung <i>Ströbel, Phillip; Hodel, Tobias; Fischer, Andreas; Scius, Anna; Wolf, Beat; Janka, Anna; Widmer, Jonas; Scheurer, Patricia; Volk, Martin</i>	98
Coding editions. Computational approaches to the editing of pre-modern texts. <i>Cugliana, Elisa</i>	102
Datenaufbereitung und -kuration im Spannungsfeld von Reproduzierbarkeit und Wiedernutzung als Leitideen der Open Sciences. Eine Fallstudie aus der Kunstgeschichte <i>Niemann, Klara; Klammt, Anne</i>	105
Die besonderen Herausforderungen multimodaler heterogener Daten- und Quellentypen an die Datenverwaltung. Ist eine Forschungsdateninfrastruktur ohne eine Datenbank umsetzbar? <i>Gerber, Anja</i>	109
»Die Greta Garbo der Leichtathletik« – Eine systematische Analyse der Modifier vossianischer Antonomasien mithilfe von Word Embeddings <i>Schwab, Michel; Fischer, Frank</i>	114
Die historische Konfliktsimulation als wissenschaftliches Modellierungsproblem in der Lehre <i>Wintjes, Jorit; Pielström, Steffen; Bock, Sina</i>	118
Die offene Edition. Vernetzung, Datenpublikation und Transparenz in der edition humboldt digital <i>Dumont, Stefan; Kraft, Tobias; Seifert, Sabine; Thomas, Christian; Wierzoch, Jan</i>	121
Disko: Zur Einbindung von Citizen Humanities beim Aufbau eines Diversitäts-Korpus <i>Schumacher, Mareike; Marie, Flüh; Peter, Leinen</i>	125
EGRAPHSEN. Von einem Nebenprodukt des Supervised Machine Learnings zu einer evidenzbasierten Malerzuweisung auf attischen Vasen <i>Kipke, Marta</i>	129
Einfluss des häufigen Lesens auf Textwahrnehmung: Ergebnisse eines Leseexperiments <i>Glawion, Anastasia; Weitin, Thomas</i>	134
FakeNarratives – First Forays in Understanding Narratives of Disinformation in Public and Alternative News Videos <i>Tseng, Chiao-I; Liebl, Bernhard; Burghardt, Manuel; Bateman, John</i>	138
Forschung, Informationswissenschaft und Archiv = drei Perspektiven auf eine Aufgabe <i>Grundig de Vazquez, Katja; Krefft, Annett; Thoden, Klaus</i>	142
Forschungsperspektiven zur Interaktion mit Musiknotation <i>Nowakowski, Matthias; Berndt, Axel; Hadjakos, Aristotelis</i>	146
From the Secret Archive to open and fair access. Ways of modelling legal ecclesiastical data from the XVI and XVII centuries <i>Albani, Benedetta; Anokhina, Alexandra; Park, Yohan</i>	151
Gattungen und Emotionen in der Lyrik des Realismus und der frühen Moderne <i>Kröncke, Merten; Konle, Leonard; Winko, Simone; Jannidis, Fotis</i>	156
GND und Normdaten für europäische Literatur? Personen und Werke in den multilingualen Korpora von ELTeC <i>Calvo Tello, José; Rißler-Pipka, Nanette; Barth, Florian</i>	160
Grenzen der Offenheit: eine digitale Sammlung zur Erforschung historischer Arzneimittelrezepte <i>Dinger, Patrick; Horstmann, Jan; Schellhammer, Stefan; Troglauer, Patrick</i>	165
How to Open Heritage? Digitale Erschließungskonzepte für Provenienzforschung am Museum für Naturkunde Berlin <i>Wagner, Sarah; Dubova, Alona; Marquart, Aron</i>	170

Increasing the visibility of Tyrol's cultural heritage through historical newspapers – the triple-open approach of the Zeit.shift project	
<i>Lyding, Verena; Franzini, Greta</i>	174
Internationale Autor*innen zu Gast in der DDR: Die Einreisekartei des Schriftstellerverbandes und ihre digitale Aufbereitung	
<i>Fischer, Frank; Illmer, Viktor Jonathan; Regeler, Lukas Nils; Müller-Tamm, Jutta; von Berenberg-Gossler, Luise; Diehr, Franziska</i>	177
Knowledge Graph-basierte Forschungsdatenintegration in NFDI4Culture	
<i>Tietz, Tabea; Bruns, Oleksandra; Fliegl, Heike; Posthumus, Etienne; Schrader, Torsten; Sack, Harald</i>	181
Knowledge Graph Design in der Forschungspraxis. Beschreibung, Interpretation und Kontextualisierung heraldischer Quellen mit der Digital Heraldry Ontology	
<i>Schneider, Philipp; Hiltmann, Torsten</i>	185
Konflikte als Theorie, Modell und Text – Ein kategorientheoretischer Zugang zur Operationalisierung von Konflikten	
<i>Gerstorfer, Dominik; Gius, Evelyn</i>	189
Korpora modular, verteilt, vernetzt in Text+	
<i>Leinen, Peter; Trippel, Thorsten; Weimer, Lukas; Witt, Andreas</i>	194
Korpuszusammensetzung und Verlässlichkeit des deutschsprachigen Google Ngram-Viewers	
<i>Jannidis, Fotis</i>	198
Korrespondenzen der Frühromantik: Ein kontrolliertes Vokabular zur Analyse von Kommunikation und Wissenstransfer für das Semantic Web	
<i>Súdzek Cronauer, Elena; Fath, Laura; Deicke, Aline; Strobel, Jochen; Weyand, Sandra; Burch, Thomas</i>	203
Lemmbasierte Publikationsformate weiter denken mit dem "Open Encyclopedia System"	
<i>Grote, Brigitte; Strobl, Maren</i>	207
"Mind the Gap": Von Lücken in der Provenienzforschung und ihrer Präsenz im digitalen Raum	
<i>Lang, Sabine</i>	212
Minimal Editing: Die Hyperdiplomatische Transkriptionsplattform	
<i>Galka, Selina; Klug, Helmut W.</i>	217
Musikgeschichte im distant reading: Präsentation der Musikverlagsdatenbank mvdb	
<i>Rosenthal, Maximilian; Richter, Matthias</i>	221
Narrativität und Handlung: Zum Verhältnis von Handlungszusammenfassungen und relevanten Ereignissen	
<i>Hatzel, Hans Ole; Gius, Evelyn; Stierner, Haimo; Biemann, Chris</i>	227
Offene Daten für die digitale Philosophie: Anforderungen an eine Datensammlung zur Philosophie und ihrer Geschichte	
<i>Heßbrüggen-Walter, Stefan</i>	231
Offene Werkgenesen, Editionen und Archive. Versuch einer generischen Datenmodellierung	
<i>Bürgermeister, Martina; Pektor, Katharina; Steindl, Christoph; Eigner, Johanna</i>	234
Offenheit durch Dokumentation: Lose Forschungsfäden im "Online-Compendium der deutsch-griechischen Verflechtungen" zusammenführen	
<i>Soethaert, Bart; Pechlivanos, Miltos</i>	237
Open and Closed AI. Eine Kunstkritik künstlich generierter Bilder	
<i>Bell, Peter</i>	242
Open Culture im Museum: „Skandal-KULTUR reloaded: Literarische Affären INTERAKTIV erkunden“ als digitales Ausstellungsexperiment	
<i>Bamberg, Claudia; Lambert, Michael; Petkov, Radoslav</i>	244
Open Data in musikphilologischen Projekten: Herausforderungen, Strategien, Potenziale	
<i>Kepper, Johannes; Münzmay, Andreas</i>	249
Open Science Prinzipien und interdisziplinäre Kollaboration in D-WISE: Zwischen Hermeneutik und Digitaler Methode in der Diskursanalyse	
<i>Eiser, Isabel; Fischer, Tim; Schneider, Florian; Koch, Gertraud; Biemann, Chris; Petersen Frey, Fynn</i>	252
Oral History auf dem Weg zu Big Data: menschliche und maschinelle Annotation lebensgeschichtlicher Interviews im Vergleich	
<i>Egger, Nils; Franken, Lina; Möbus, Dennis; Schmid, Florian</i>	257
PhiloBERTa: Ein multilinguales Sprachmodell zur Beantwortung philosophiehistorischer Fragestellungen	
<i>Noichl, Maximilian; Panzer, Lukas</i>	261
Provenienzforschung und ihre Quellenbestände. Aktuelle Nutzungsszenarien zwischen Open Access und Inaccessibility	
<i>Hopp, Meike; von dem Bussche, Ruth</i>	265

#PublicDH oder doch nur #WissKomm?	269
<i>Seltmann, Melanie Elisabeth-H.</i>	
Re: ARTigo. Neuentwurf eines Social-Tagging-Frameworks aus funktionalen Programmbausteinen	273
<i>Schneider, Stefanie; Kristen, Maximilian; Vollmer, Ricarda</i>	
Sammlungsdaten mit Wikidata anreichern und für die Vernetzung öffnen. Konzepte und praktische Erprobungen	278
<i>Schelbert, Georg; Müller, Michael</i>	
Schrifttradition digital: Rituell reine Torarollen in der jüdischen Diaspora	283
<i>Frank, Laura; Eichhorst, Dana; Ullrich, Rebecca; Hadassah Wendl, Katharina; Martini, Annett; Tonne, Danah</i>	
Selbstoptimierung vs. Selbstliebe? Eine vergleichende Inhaltsanalyse von Fitspiration- und Bodypositivity-Bildern auf Instagram mit Methoden der automatischen Bildklassifikation	286
<i>Glas, Julia; Wolff, Christian; Ludwig, Bernd; Achmann, Michael</i>	
Skalierungspraktiken in der computergestützten Analyse von literarischen Texten	291
<i>Krautter, Benjamin</i>	
Synoptische Interfaces Digitaler Editionen	296
<i>Herbst, Yannik; Roeder, Torsten; Reul, Christian</i>	
Textliche Relationen maschinenlesbar formalisieren: Systeme der Intertextualität	301
<i>Horstmann, Jan; Lück, Christian; Normann, Immanuel</i>	
Tool Studies 2.0 – Zum Potenzial von Transformern für die Erkennung und Klassifikation von Software-Tools in DH-Publikationen	304
<i>Burghardt, Manuel; Ruth, Nicolas; Niekler, Andreas</i>	
Understanding the impact of three derived text formats on authorship classification with Delta	309
<i>Du, Keli</i>	
Vom Heben verborgener Schätze – Literarische Blogs als Ressource	312
<i>Schenk, Nicolas; Blessing, André; Hein, Pascal; Hess, Jan; Jung, Kerstin; Schlesinger, Claus-Michael</i>	
Vom sprachlichen Indikator zum komplexen Phänomen?	317
<i>Jacke, Janina</i>	
Von A bis Z: Überlegungen zur Erstellung eines Wissensgraphen aus historischen Enzyklopädiën	321
<i>Hagen, Thora</i>	
„Werktitel als Wissensraum“ – über die Potentiale von Werknormdaten für die Digitalen Geisteswissenschaften	326
<i>Dietrich, Elisabeth; Kolbe, Ines</i>	
„Zu Rande kommen“: Phänomen und Präsentation von Randnotizen am Beispiel der digitalen Ferdinand-Tönnies-Briefedition	329
<i>Bamberg, Claudia; Dörk, Uwe; Wierzock, Alexander; Trautmann, Tatjana; Burch, Thomas; Petkov, Radoslav</i>	

Doctoral Consortium

Constructing Multicultural Germany: Narratives on the Germany Men's National Football Team from 2006 to 2018	336
<i>Kou-Herrema, Tianyi</i>	
Diagramme edieren – zur kritischen Repräsentation visueller Narrative	337
<i>Sutor, Nadine</i>	
Erweiterbare, interaktive Softwareplattform für die Anwendung von Sprachtechnologie in großen Textkorpora zur Unterstützung von Such- und Analyseworkflows in den Digital Humanities	339
<i>Petersen-Frey, Fynn</i>	
Feministische Filmgeschichte als Linked Open Data: Ein Thesaurus für das Women Film Pioneers Project (WFPP)	341
<i>Junginger, Pauline</i>	
Listen in historischen Zeitungen: Herausforderungen und Potenziale der digitalen Analyse einer vernachlässigten Textsorte	343
<i>Rastinger, Nina C.</i>	
Mixed Methods in der Genozidforschung	345
<i>Schirmer, Miriam</i>	
Stilometrie in der Diplomatie: Ein neues Forschungsfeld?	346
<i>Geißel, Pia</i>	

Theorising the Aesthetic Properties of Reading in a Digital Social Reading (DSR) Environment: Exploring DSR Practices in India <i>Ghosh, Sharanya</i>	348
Von Wissensdingen und Werkräumen. Graph-basierte Modellierung von Denk- und Arbeitsspuren in Nachlässen <i>Stahn, Lena-Luise</i>	350

Posterpräsentationen

Analyse, Produktion, Reflexion: Nachnutzungsszenarien für Forschungsdaten am Beispiel der Daten des Projekts Dehmel digital <i>Bläß, Sandra; Flüh, Marie; Nantke, Julia; Reul, Christian</i>	355
A Quantitative Analysis of Digital Scholarly Editions <i>Kurzmeier, Michael; O'Sullivan, James; Murphy, Órla; Pidd, Michael; Wessels, Bridgette</i>	356
ARS - Architecture Research Stage <i>Dürfeld, Michael; Stein, Christian; List, Ferdinand; Rahman, Zead; Dias, Renata; Thran, Niklas; Marschner, Michèle</i>	357
Barockpoetik als Wikibase: Eine Datenbank zu konfessions-geschichtlichen Aspekten in deutschen Barockpoetiken <i>Haider, Thomas Nikolaus; Schennach, Stephanie; Thelen, Julius; Wesche, Jörg</i>	358
Beginnen in Köln: Von der Textdatenbank zur zeitgemäßen digitalen Auszeichnung und Analyse <i>Bigalke, Jan; Blumtritt, Jonathan; Gengnagel, Tessa</i>	360
Buddhist Murals of Kucha on the Northern Silk Road. Ein Versuch der Semi-Automatisierung der Annotierung <i>Radisch, Erik</i>	362
Building a virtual research environment to move from digital to distant Diplomats (ERC project DiDip) <i>Vogeler, Georg; Luger, Daniel; Nicolaou, Angelos; Kovacs, Tamas; Atzenhofer-Baumgartner, Florian; Lamminger, Florian; Aoun, Sandy; Decker, Franziska</i>	364
Das Deutsche Kunstarchiv auf neuen Wegen <i>Fischeidl, Kathrin</i>	366
Das preußische Kriegsspiel als Forschungsobjekt <i>Henning, Pia</i>	368
Das QhoD-Projekt: Digitale Edition von Quellen zur habsburgisch-osmanischen Diplomatie 1500-1918 <i>Mayer, Manuela; Kurz, Stephan; Yilmaz, Yasir; Sonnberger, Jakob</i>	369
Das Thüringische Flurnamenportal <i>Aehnlich, Barbara; Kunze, Petra</i>	370
Der NFDI4Culture Helpdesk – ein Beratungsangebot für die Kulturwissenschaften <i>Mayer, Desiree; Kailus, Angela</i>	372
Die digitale Schulbuch-Bibliothek GEI-Digital im neuen Gewand: Ein modernes Präsentationssystem öffnet digitalisierte Schulbücher für die Open Humanities <i>Klaes, Jan Sebastian; Krüger, Katharina; Leonhardt, Susann; Nieländer, Maret; Sommer, Kai; Towara, Nadine</i>	373
Die Wahl der Mittel – Jupyter-Notebooks als Forschungsinfrastruktur <i>Jung, Kerstin; Hein, Pascal; Blessing, André; Hess, Jan; Kushnarenko, Volodymyr</i>	375
DigEdTnT - Digital Edition Creation Pipelines: Tools and Transitions <i>Pollin, Christopher; Strutz, Sabrina; Steiner, Christian; Klug, Helmut</i>	377
Digitale Editionen von historischen Reiseberichten öffnen: Open Text und Open Data mit einheitlicher Textauszeichnung, semantischer Annotation und ontologiebasierter Datenmodellierung <i>Balck, Sandra; Frank, Ingo</i>	378
Digitale Interaktion auf Augenhöhe – drei Wege zu partizipativer Forschung und FAIRer Lehre an der UB Kiel <i>Christ, Andreas; Diebel, Richard; Henzel, Katrin; Petersen, Britta; Vetter, Angela</i>	381
Digitale Methoden kritisch reflektieren – Die Erweiterung des Werkzeugkastens der Historiker:innen <i>Althage, Melanie</i>	383
duerer.online - Virtuelles Forschungsnetzwerk Albrecht Dürer <i>Große, Peggy</i>	384

Dunkelgrün, blassgrün, fenchelgrün oder: Über die Konkretisierung des Vokabulars im deutschsprachigen Roman (1760–1920)	
<i>Hilger, Agnes</i>	386
D-WISE - Digitale Wissenssoziologische Diskursanalyse	
<i>Fischer, Tim; Eiser, Isabel; Schneider, Florian; Petersen-Frey, Fynn; Biemann, Chris; Koch, Gertraud</i>	388
Fabrikation von Erkenntnis: Experimente in den Digital Humanities	
<i>Dieckmann, Lisa; Steyer, Timo; Walkowski, Niels-Oliver; Weis, Joëlle; Wuttke, Ulrike</i>	390
Fanfiction Semantics - Eine quantitative Analyse sensibler Themen in deutscher Fanfiction	
<i>Häußler, Julian</i>	391
Kaleidoskopische Muster des Protests. Visuelle und textuelle (Selbst-)Repräsentationen osteuropäischer Protestkulturen aus qualitativer und quantitativer Perspektive	
<i>Howanitz, Gernot; Kaltseis, Magdalena</i>	393
Klassifikation von Figurenauf- und -abtritten in XML-kodierten Dramen	
<i>Ehlers, Lena; Andresen, Melanie</i>	395
KoMuX - Der Kompositamuster-Explorer	
<i>Brunner, Annelen; Katrin, Hein</i>	397
Metaphors of Religion	
<i>Gebhard, Henning; Jha, Vandana; Tögel, Philipp; Dipper, Stefanie; Elwert, Frederik; Tonne, Danah</i>	399
Netzwerk Offenes Mittelalter	
<i>Borek, Luise; Busch, Hannah; Ketschik, Nora</i>	401
NFDI4Culture und Text+ - Kartierung einer Zusammenarbeit	
<i>Schrade, Torsten; Stein, Regine; Tolkendorf, Julia; Vater, Christian; Weimer, Lukas</i>	402
Offene Editionen - Die Task Area Editionen im NFDI-Konsortium Text+	
<i>Blumtritt, Jonathan; Cugliana, Elisa; Geißler, Nils; Hegel, Philipp; Hensen, Kilian; Hörnschemeyer, Jörg; Kudella, Christoph; Lemke, Karoline; Lordick, Harald; Neuber, Frederike; Neuefeind, Claes; Schulz, Daniela; Seltmann, Melanie Elisabeth-H.; Sievers, Martin; Gengnagel, Tessa</i>	404
Onboard onto DraCor. Prototyping Workflows to Homogenize Drama Corpora for an Open Infrastructure	
<i>Börner, Ingo; Fischer, Frank; Giovannini, Luca; Lu, Christopher; Milling, Carsten; Skorinkin, Daniil; Sluyter-Gäthje, Henny; Trilcke, Peer</i>	406
Open Archives VR. Ein 3D-Modell des Theodor-Fontane-Archivs als interaktiver Erlebnis- und Kommunikationsraum	
<i>Brandes, Vanessa; Busch, Anna; Trilcke, Peer; Zimmermann, Ronny</i>	408
Opening a Journal. Erfahrungen bei der Gründung des Journal of Computational Literary Studies	
<i>Gius, Evelyn; Schöch, Christof; Trilcke, Peer; Gerstorfer, Dominik; Guhr, Svenja; Ripoll, Elodie; Sluyter-Gäthje, Henny</i>	410
Open Jean Paul. Funktionen und Potentiale offener Editionsdaten	
<i>Neuber, Frederike; Lecroq, Axelle</i>	412
OWIDplusLIVE - Tagesaktuelle N-Gramm-Analysen	
<i>Rüdiger, Jan Oliver; Wolfer, Sascha; Cotgrove, Louis</i>	413
Projektvorstellung - Sprachanfragen. Empirisch gestützte Erforschung von Zweifelsfällen	
<i>Lang, Christian; Tu, Ngoc Duyen Tanja; Schneider, Roman; Volodina, Anna</i>	415
Schlendern im Digitalen Museum	
<i>Hall, Mark; Walsh, David</i>	417
Sharing the CROWN - Von Sammlungsdaten zu Linked Open Research Data	
<i>Grießer, Martina; Hanzer, Helene; Kirchweyer, Franz; Kloser, Peter; Lamers, Teresa; Pollin, Christopher; Scholger, Martina; Steiner, Elisabeth; Vasold, Gunter</i>	419
Standoff-Tools - Generische Dienste für die automatische Annotation von XML-Dokumenten mit Plain-Text-Werkzeugen	
<i>Lück, Christian</i>	421
Urheberrechtlich geschützte Texte nachnutzen - Der XSample-Workflow	
<i>Andresen, Melanie; Gaertner, Markus; Jacke, Janina; Ketschik, Nora; Pichler, Axel</i>	422
Vom Finden, Filtern und Auswerten der relevanten Daten im digitalen Nachlass von Friedrich Kittler im Deutschen Literaturarchiv Marbach	
<i>Holz, Alex; Çakir, Dilan Canan</i>	424
Was heißt eigentlich ‚offen‘? Eine korpuslinguistische Untersuchung am Beispiel des bibliothekarischen Diskurses der SLUB Dresden	
<i>Meier-Vieracker, Simon; Weigelt, Lucie; Dutschke, René; Lasch, Alexander; Scherbaum, Stefan; Seemann, Sophia; Pfeifer, Ulrike</i>	426
Weißbuch Digitale Edition	
<i>Galka, Selina; Klug, Helmut W.</i>	428

Wie die OPERAS-Projekte PRISM und TRIPLE Open Humanities unterstützen können <i>Piel, Patrick; Töpfer, Marlene; Günther, Johanna</i>	430
---	-----

Anhang

Index der Autorinnen und Autoren	433
--	-----

Workshops

Algorithmen anwenden – algorithmisch denken. „Algorithmizität“ als Brücke zwischen Geisteswissenschaften und Informatik?

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de
Universität Leipzig

Geiger, Jonathan D.

Jonathan.Geiger@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz

Horstmann, Jan

jan.horstmann@uni-muenster.de
Westfälische Wilhelms-Universität Münster

Kleymann, Rabea

kleymann@zfl-berlin.org
Leibniz Zentrum für Literatur- und Kulturforschung
Berlin

Schmitz, Jascha

smitzjak@hu-berlin.de
Humboldt-Universität Berlin

Schwandt, Silke

silke.schwandt@uni-bielefeld.de
Universität Bielefeld

Latente und explizite Algorithmizität

In den Geisteswissenschaften werden immer wieder tradierte Strukturen der Wissensproduktion hinterfragt und neu geordnet – so auch vor dem Hintergrund der Digital Humanities, die die traditionellen Geisteswissenschaften mit digitalen Forschungsmethoden teils unterstützen, teils konfrontieren. Neben Reflexionen über Medialität und Sozialität geisteswissenschaftlicher Erkenntnisprozesse werden dichotomische Setzungen wie analog/digital, qualitativ/quantitativ sowie kontinuierlich/diskret auf den Prüfstand gestellt. Gleichzeitig gewinnen auch andere „epistemische Tugenden“ (Daston

und Galison 2007), wie zum Beispiel Transparenz, Evidenz und Reproduzierbarkeit, für geisteswissenschaftliche Forschungsfelder eine neue Relevanz. Im Zuge dessen lässt sich eine umfassende Neuvermessung disziplinspezifischer Kulturen des Verstehens beobachten, die jenseits eines tradierten „Two Cultures“-Paradigmas (Snow 1959) verfährt. Kennzeichnend sind vielmehr Querverbindungen und Verschränkungen, welche die disziplinären Profile der Geisteswissenschaften sowie der Informatik neu ins Verhältnis setzen. Als eine solche mögliche disziplinäre Querverbindung möchten wir im Workshop *Algorithmizität* verhandeln.

Der von der AG „Digital Humanities Theorie“ organisierte Workshop widmet sich Voraussetzungen, Potenzialen und Implikationen von Algorithmizität in den Humanities. Dabei knüpft der Workshop an aktuelle Forschungsinteressen der AG an und setzt zugleich Diskussionen der im Juni 2022 organisierten Tagung „Algorithmizität als Kultur des Verstehens“ fort. Einen ersten Ausgangspunkt des Workshops bildet nun die Annahme, dass sich in traditionellen und aktuellen Forschungspraktiken der Geisteswissenschaften latente Formen der Algorithmizität wiederfinden lassen. Darunter sind nicht nur digitale Formen geisteswissenschaftlichen Arbeitens im Sinne einer prozessual aufgefassten Algorithmisierung zu verstehen. Vielmehr bezeichnet Algorithmizität einen Formalisierungsgrad von Handlungsanweisungen, der sich in geisteswissenschaftlichen Methoden *sui generis* manifestiert. Insbesondere Versuche einer systematischen und symbolischen Externalisierung geistiger Operationen sind seit langem Bestandteil philosophischen Denkens (vgl. Gramelsberger 2020). Im Verlauf der Tagung wurden konkrete Beispiele latenter (d.h. impliziter) bis expliziter Algorithmizität in den Geisteswissenschaften skizziert und diskutiert, die sich etwa in der Mathematisierung der Musik und algorithmischer Musikproduktion in den Musikwissenschaften (vgl. Braguinski 2018), dem systematisierten Vorgehen der *Objektiven Hermeneutik* und *Grounded Theory* in den Sozialwissenschaften (vgl. Müller et al. 2016; vorgestellt durch Dennis Möbus) oder in Binarität als Zugang zu literaturwissenschaftlichen Untersuchungen von Textmaterialität (vgl. Coch, Hahn und Pethes 2022) ausdrückt. Weitere historische Beispiele wären die Konzeption von Rede und Gegenrede in den platonischen Dialogen, das dialektische Format der mittelalterlichen *Quaestio* im Anschluss an Petrus Abaelardus, die methodologischen Überlegungen der frühneuzeitlichen Wissenschaftsphilosophie wie sie beispielsweise bei Bacon oder Descartes zu finden sind, die phänomenologische Reduktion nach Husserl oder schließlich die Hermeneutik als verbindendes Element der Geisteswissenschaften, der als methodisches und stufenweises Vorgehen der generellen Textrezeption eine besondere Rolle zukommt.

Vor diesem Hintergrund verstehen wir Algorithmizität als ein graduelles Phänomen (vgl. Abb. 1). So können latente und explizite Formen der Algorithmizität unterschieden werden, die einerseits von Prozessen der Quantifizierung und Diskretisierung bis zu maschinenlesbaren Handlungsabläufen reichen, deren prozesshafte Kontingenzreduktion aufgrund zeichenbasierter Befehle maximal ist. Andererseits werden auch disziplinspezifische

Umgangsformen mit Algorithmizität und Praktiken einer algorithmischen Gegenstandskonstruktion sichtbar.

latente Algorithmizität

Quantifizierung

Diskretisierung

maschinenlesbare Handlungsabläufe

explizite Algorithmizität

Abb. 1: Algorithmizität als graduelles Phänomen

Algorithmizität wollen wir im Workshop aus drei unterschiedlichen Perspektiven betrachten und diskutieren: (1) als Konzept in der theoretischen Begriffsarbeit, (2) als spezifisch (geisteswissenschaftliche) Erkenntnis- und Denkstruktur und (3) als (trans-)disziplinärer Kompass, d.h. als eine Art Navigationshilfe zwischen den einzelnen Disziplinen der Digital Humanities. Ziel des Workshops ist es, gemeinsam mit der DH-Community zu elaborieren, inwiefern Algorithmizität eine geeignete Beschreibungskategorie ist, um bestimmte Formen geisteswissenschaftlicher Wissensproduktionen zu charakterisieren. Die gemeinsame Diskussion von Beschreibungskategorien trägt dabei, so argumentieren wir, auch Forderungen von Open Science Rechnung, da die Dokumentation solcher Kategorien die Transparenz, Nachvollziehbarkeit und Reproduzierbarkeit von Interpretationen erhöht. Was wird eigentlich verhandelbar, wenn Strukturen der Problembehandlung und Bedeutungszuweisung in den Humanities als algorithmisch beschrieben werden? Inwiefern stellt Algorithmizität eine Bereicherung für den interdisziplinären Austausch dar? Denn Algorithmizität kann nicht nur als impliziter Teil von etablierten Kulturen des Verstehens, sondern selbst als Kulturtechnik des Verstehens begriffen werden. In welchen latenten und expliziten Ausprägungen Algorithmizität in einzelnen geisteswissenschaftlichen Erkenntnisprozessen stattfindet, wollen wir im Workshop diskutieren. Darüber hinaus fragen wir auch, inwiefern Algorithmizität in den computationellen Disziplinen einerseits und in den Geistes- und Kulturwissenschaften andererseits strukturelle Kongruenzen aufweisen, die methodisch, technisch und wissenschaftskulturell als Brücke zwischen diesen beiden Wissenschaftsbereichen dienen können.

Vermessung von Algorithmizität als Beschreibungskategorie

Aktuell wird Algorithmizität in den verschiedenen Disziplinen der (Digital) Humanities in unterschiedlicher Form diskutiert: Die Sozial- und Wirtschaftsgeschichte beispielsweise nimmt auf verschiedene Art Bezug auf algorithmische Methoden der empirischen Sozialforschung oder historischen Demographie, um ihre Interpretationen zu stützen (vgl. Schremmer 1998). In medienwissenschaftlichen Kontexten wird Algorithmizität einerseits als eine Eigenschaft der Kultur der Digitalität (vgl. Stalder 2019, 13) bezeichnet, andererseits aber auch als eine *potentia* beschrieben, die dem Algorithmus als spezifische Form vorausgeht (vgl. Rutz 2016, 41). In einer informativen Perspektive taucht Algorithmizität neben Sequenzialität und Abstraktion als Teilaspekt eines *computational thinking* auf (vgl. Denning und Tedre 2019). Für die DH eröffnet sich ein Reflexionsraum für die eigenen Verstehenspraktiken (vgl. Rehbein 2022). Für die beteiligten geisteswissenschaftlichen Disziplinen stellt sich darüber hinaus die Frage nach dem Potenzial der Auseinandersetzung mit dem Begriff der Algorithmizität für die eigene Theoriereflexion.

Um dieses interdisziplinäre Feld aus Perspektive der Algorithmizität zu sondieren, widmet sich unser Workshop folgenden zentralen Fragekomplexen:

Algorithmizität und Begriffsarbeit

- Was sind Kriterien zur Bestimmung des Algorithmischen? (Performanz, Effizienz, Binarität ...) Welche Rolle(n) spielen sie?
- Gibt es ein Spannungsverhältnis zwischen Algorithmizität und Verstehen oder ist das eine im anderen enthalten?
- Wie können Begriffe wie 'Algorithmus', 'Algorithmizität' oder 'Algorithmisierung' differenziert werden?
- Was sind die Grenzen des Algorithmizitätsbegriffs? Wie unterscheidet sich Algorithmizität von den Begriffen Kalkül (vgl. Krämer 1991), Modell (vgl. Flanders und Jannidis 2019) oder Regel (vgl. Daston 2022)?

Algorithmizität als Erkenntnis- und Denkstruktur

- Ist *computational thinking* das Gleiche wie ein algorithmischer Erkenntnisprozess und das Gleiche wie ein regelgeleitetes Vorgehen (wie etwa Dilthey oder Descartes es beschreiben)?
- Was nützt uns die Klassifizierung von (geisteswissenschaftlichen) Erkenntnisprozessen als algorithmisch oder nicht-algorithmisch?
- Wird alles algorithmisch? Was sind die Grenzen des Phänomens?

Algorithmizität als (trans-)disziplinärer Kompass

- Wie unterscheiden sich die in den einzelnen Disziplinen der Humanities angewandten Algorithmen oder regelgeleiteten Methodensettings? Was sind Gemeinsamkeiten?
- Was sind disziplinäre Unterschiede zwischen geisteswissenschaftlicher und informatischer Wissensproduktion?

- Was nützt uns das Wissen um Gemeinsamkeiten und Unterschiede der weniger oder stärker algorithmisch geprägten Art der Erkenntnisproduktion in den Disziplinen?
- Ist das „Two Cultures“-Paradigma weiterhin gültig?

Methodik und Ablauf des Workshops

Um partizipative Strukturen vor, während und nach dem Workshop zu ermöglichen, haben wir ein begleitendes Programm geplant. Zur Vorbereitung des Workshops sollen bis zur DHd-Konferenz 2023 sukzessive drei Beiträge für den Theorie-Blog¹ publiziert werden: Nach einer Zusammenfassung der Tagung von 2022 mit ihren zentralen Thesen und Diskussionspunkten planen wir, zwei kontrastierende *opinion pieces* zu veröffentlichen, die das Für und Wider einer Verwendung des Begriffs 'Algorithmizität' im geisteswissenschaftlichen Diskurs versammeln. Um schon im Vorfeld Spannweiten des Algorithmizitätsbegriffs zu erproben, wird die Forschungscommunity ausdrücklich zur Partizipation aufgerufen, indem wir Diskussionen über die blögeigene Kommentarfunktion oder über Twitter² und Mastodon³ ermöglichen. Im Workshop selbst sollen einerseits die vorläufigen Ergebnisse und Kernthesen der vorangegangenen Tagung in kondensierter Form präsentiert werden, vor allem wollen wir andererseits aber auch eine systematische Diskussion zur Verfasstheit und zum Potenzial von Algorithmizität führen. Unser Ziel für den Workshop ist es, den Teilnehmenden das Konzept bzw. Konzepte von Algorithmizität nahezubringen, anhand konkreter Beispiele aus den (digitalen) Geisteswissenschaften kritisch zu diskutieren und schließlich seine Eignung und seinen Mehrwert für den Theoriediskurs in den DH herauszuarbeiten.

Der Workshop ist für vier Stunden angesetzt und der Ablauf ist folgendermaßen strukturiert:

0:00–0:15 Kurze Vorstellung der AG Theorie und der Workshopteilnehmenden

0:15–0:30 Erste Annäherungen an den Algorithmizitätsbegriff

0:30–0:45 Zusammenfassung der wesentlichen Thesen und vorläufigen Ergebnisse aus der vorangegangenen Tagung

0:45–1:30 Kontrastierende Stellungnahmen (pointierte Provokationen) und anschließende Diskussionen

- Algorithmizität als leere Worthölse
- Algorithmizität als Projektionsfläche für Theoriediskurse in den DH

1:30–2:00 Pause

2:00–3:00 World-Café zu einzelnen Diskussionspunkten (vgl. zentrale Fragekomplexe), moderiert durch AG-Mitglieder; Ziele:

- Begriffsarbeit: Erarbeitung von interdisziplinären und disziplinspezifischen Kriterien und Definitionen von 'Algorithmizität'
- Erkenntnisstruktur: Bestimmung (möglicher) Funktionen des Begriffs im wissenschaftlichen Diskurs

- (trans-)disziplinärer Kompass: Sammlung von Beispielen des Algorithmischen in bestehenden geisteswissenschaftlichen Wissensstrukturen

3:00–3:15 Pause

3:15–3:45 Vorstellung der Gruppenarbeiten und moderierte Diskussion

3:45–4:00 Dokumentation der Ergebnisse; Ziel: Potenzialbewertung des Algorithmizitätsbegriffs für die Charakterisierung der DH allgemein und die Theoriebildung in den DH im Besonderen

Zielpublikum und Teilnehmer*innenzahl

Am Workshop können bis zu 25 Theorie-interessierte Personen aus allen Teildisziplinen der Digital Humanities teilnehmen. Erfahrungen im Einsatz oder in der Reflexion von Algorithmen und/oder epistemischen Strukturen ist hilfreich, aber keine Teilnahmevoraussetzung.

Beteiligte und Forschungsinteressen

Manuel Burghardt ist Professor für Computational Humanities im Institut für Informatik an der Universität Leipzig. Seine Forschungsinteressen umfassen computergestützte Verfahren der Annotation, Analyse und Visualisierung von geisteswissenschaftlichen Forschungsdaten.

Jonathan D. Geiger arbeitet an der Akademie der Wissenschaften und der Literatur | Mainz im Infrastrukturprojekt NFDI4Culture. Seine Interessenschwerpunkte liegen auf der philosophisch-theoretischen Reflexion digitaler Forschungsmethoden in den Geisteswissenschaften und der Digitalität insgesamt, sowie auf Fragen infrastruktureller Bedürfnisse der NFDI4Culture-Bedarfe und der Philosophie-Community.

Jan Horstmann leitet das Service Center for Digital Humanities (SCDH) an der ULB der Westfälischen Wilhelms-Universität Münster. Seine Forschungsinteressen und -schwerpunkte liegen im Bereich der digitalen Methodologie mit besonderem Fokus auf die Textannotation, -analyse und Visualisierung im Bereich der computationalen Literaturwissenschaft.

Rabea Kleymann ist Postdoktorandin am Leibniz-Zentrum für Literatur- und Kulturforschung Berlin. Dort leitet sie seit 2020 das Projekt „Diffraktive Epistemik. Wissenskulturen in den Digital Humanities“. Ihre Forschungsinteressen liegen im Bereich der Wissenschaftstheorie, Science & Technology Studies sowie den computationalen Literaturwissenschaften.

Jascha Schmitz studiert im Master Geschichtswissenschaften an der Humboldt-Universität zu Berlin mit dem Schwerpunkt Digital History und arbeitet als wissenschaftliche Hilfskraft am Max-Planck-Institut für Wissenschaftsgeschichte. Im Rahmen seiner Masterarbeit liegt sein Forschungsfokus auf Simulationsmethoden für die Geschichtswissenschaften.

Silke Schwandt ist Professorin für Digital History an der Universität Bielefeld. Sie forscht aktuell zur Veränderung von geschichtswissenschaftlichen Forschungspraktiken unter Bedingungen der Digitalität sowie zum Potential von digitalen Methoden für die Selbstreflexion der Geschichtswissenschaft.

Benötigte technische Ausstattung

Wir benötigen einen Raum mit flexibler Bestuhlung und Tischen, die zu Gruppentischen verschoben werden können. Außerdem benötigen wir einen Moderationskoffer mit Materialien für Gruppenarbeiten und einen Beamer.

Fußnoten

1. Vgl. <https://dhtheorien.hypotheses.org/> (zugegriffen: 09. Dezember 2022).
2. Vgl. die bisherige Diskussion unter #dhtheorie_Algo22 (zugegriffen: 09. Dezember 2022).
3. Vgl. <https://fedihum.org/@DHTheorie> (zugegriffen: 09. Dezember 2022).

Bibliographie

- Braguinski, Nikita. 2018. *RANDOM. Die Archäologie der elektronischen Spielzeugklänge*. Bochum: projektverlag.
- Coch, Charlotte, Torsten Hahn und Nicolas Pethes (Hg.). Im Erscheinen. *Lesen / Sehen. Literatur als wahrnehmbare Kommunikation*. Bielefeld: transcript.
- Daston, Lorraine. 2022. *Rules: A Short History of What We Live By*, Princeton: Princeton University Press. <https://doi.org/10.1515/9780691239187>.
- Daston, Lorraine und Peter Galison. 2007. *Objektivität*. 1. Aufl. Frankfurt am Main: Suhrkamp.
- Denning, Peter J. und Matti Tedre. 2019. *Computational Thinking*. Cambridge: MIT Press.
- Flanders, Julia und Fotis Jannidis (Hg.). 2019. *The Shape of Data in the Digital Humanities: Modeling Texts and Text-Based Resources. Digital research in the arts and humanities*. London, New York: Routledge.
- Gramelsberger, Gabriele. 2020. *Operative Epistemologie. (Re-)Organisation von Anschauung und Erfahrung durch die Formkraft der Mathematik*. Hamburg: Meiner. <https://doi.org/10.28937/978-3-7873-3900-6>.
- Krämer, Sybille. 1991. *Berechenbare Vernunft: Kalkül und Rationalismus im 17. Jahrhundert*, Berlin, Boston: de Gruyter. <https://doi.org/10.1515/9783110847079>.
- Muller, Michael, Shion Guha, Eric P. S. Baumer, David Mimno und N. Sadat Shami. 2016. "Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination." *Proceedings of GROUP'16*, Sanibel Island. <https://doi.org/10.1145/2957276.2957280>.
- Rehbein, Malte. 2022. "Digitizing the Humanities." In *Handbook Industry 4.0. Law, Technology, Society*, hg. von Walter Frenz, 1171-1176. Berlin: Springer.

Rutz, Hanns Holger. 2016. "Making a Space of Algorithmicity." In *xCoAx 2016: Proceedings of the Fourth Conference on Computation, Communication, Aesthetics and X*, hg. von Mario Verdicchio, Alison Clifford, André Rangel und Miguel Carvalhais, 29-42.

Schremmer, Eckart (Hg.). 1998. *Wirtschafts- und Sozialgeschichte. Gegenstand und Methode: 17. Arbeitstagung der Gesellschaft für Sozial- und Wirtschaftsgeschichte in Jena 1997* (Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte). Stuttgart: Steiner.

Snow, Charles Percy. 2012 [1959]. *The two cultures*. Cambridge, New York: Cambridge UP.

Stalder, Felix. 2019. *Kultur Der Digitalität*. 4. Auflage. Berlin: Suhrkamp.

Data Driven Storytelling zu kulturellen Objekten und Biographien

Liem, Johannes

johannes.liem@donau-uni.ac.at
Universität für Weiterbildung Krems, Österreich

Kusnick, Jakob

kusnick@imada.sdu.dk
Fluxguide GesmbH, Österreich

Jänicke, Steffan

stjaenicke@imada.sdu.dk
Fluxguide GesmbH, Österreich

Doppler, Carina

carina@fluxguide.com
University of Southern Denmark, Dänemark

Passecker, Markus

markus@fluxguide.com
University of Southern Denmark, Dänemark

Mayr, Eva

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

Windhager, Florian

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

Hintergrund

Die Omnipräsenz von Geschichten in der menschlichen Kultur - wie auch zahlreiche akademische Refle-

xionen - machen deutlich: Narrative Strukturierung von Inhalten gehört zu den wesentlichsten Gestalten und Gestaltungsstrategien für die Vermittlung neuer, relevanter oder unterhaltsamer Informationen sowohl in der heutigen Kultur als auch in der breiteren Geschichte (Bolin, 2010; Dykes, 2019). Zahlreiche Hypothesen liefern hierfür mögliche Gründe, (u. a. ein spezieller Modus der narrativen Informationsverarbeitung in der menschlichen Kognition), aber die zentrale Folgerung für modernes Informations- und Kommunikationsdesign ist verhältnismäßig simpel: Gutes Storytelling kann die Aufnahme und das Verständnis von komplexer Information entscheidend verbessern und auch in digitalen Medien zu einer vertieften Auseinandersetzung mit diversen Inhalten führen. Dies hat auch im Feld der Visualisierung dazu geführt, dass datengetriebenes, visuelles Storytelling in den letzten Jahren zu einem allgegenwärtigen Thema in der Erforschung und Entwicklung von Visualisierungen geworden (Segel & Heer, 2010; Riche et al., 2018). Auch in der visuellen Vermittlung von digitalen kulturellen Sammlungen (Windhager et al., 2018) oder von Biografiedaten kommen narrative Methoden immer öfters zum Einsatz (Kusnick et al., 2021).

Das InTaVia-Projekt

Das InTaVia-Projekt ("In/Tangible European Heritage - Visual Analysis, Curation and Communication, <https://intavia.eu>) zieht größere Bestände von materiellem und immateriellem Kulturerbe in eine transnationale Datenbasis zusammen (Windhager, Mayr, Schlögl, & Kaiser, 2022, in Druck). In den letzten Jahrzehnten wurde sowohl die Digitalisierung von materiellen Objektsammlungen vorangetrieben (Khan, Shafi, & Ahangar, 2018), wie auch die Digitalisierung von biografischem Wissen über Kulturschaffende (ter Braake et al., 2015; 2017). Diese Entwicklungen bieten eine gute Basis für eine digitale Analyse und Kommunikation des Lebens und Werks von Kulturschaffenden (Khulusi et al., 2016; Schlögl, Windhager, Mayr, & Kaiser, 2019; Windhager et al., 2018), aber fehlende Verknüpfungen, Harmonisierungen und ein Mangel an Werkzeugen erschweren die entsprechende Arbeit - besonders für Praktiker*innen im Bereich der Kulturvermittlung. Mit Blick auf etablierte Nationalbiografien und korrespondierende kulturelle Objektdaten arbeitet das InTaVia-Konsortium an der Entwicklung von Lösungen und harmonisiert zu diesem Zweck nationale Datenbestände (inkl. der Biografieprojekte von Finnland, Niederlande, Österreich und Slowenien). Darauf aufbauend entwickelt es ein prototypisches Informationsportal für die visuelle Analyse und Kommunikation dieser integrierten Kulturdaten. So werden synoptische Ansichten auf historischen Daten zu Leben und Werken aus verschiedenen Perspektiven der Datenvisualisierung (geografisch, relational, kategorial, chronologisch) möglich, sowie die narrative Gestaltung und Vermittlung dieser Information mittels Methoden des individuellen und kollektiven Storytellings.

Zielsetzung Workshop

Als "Early-Access Workshop" zielt die Veranstaltung auf die Erprobung und Diskussion von prototypischen Methoden des visuellen Storytellings mit Kulturdaten. Teilnehmende werden zur Nutzung und Erprobung der InTaVia-Plattform eingeladen, um dort mit exemplarischen Daten die Möglichkeiten der narrativen visuellen Gestaltung auszuloten und zu diskutieren. Er richtet sich sowohl an Forscher*innen wie auch Praktiker*innen im Bereich des kulturellen Erbes, der Kunst- und Kulturgeschichte, sowie angrenzender Geisteswissenschaften. Eine einführende Diskussion von State-of-the-Art-Methoden aus DH-Perspektive wird dabei verbunden mit einer kurzen Vorstellung der InTaVia-Plattform und ihrer Technologien - mit spezifischen Fokus auf Module der Datenkuratierung und des visuellen Storytellings. Teilnehmende Expert*innen können so Einblicke in aktuelle Entwicklungen der narrativen Visualisierung gewinnen, während die Veranstalter*innen des Workshops mögliche Anregungen und Wünsche für die partizipative Weiterentwicklung der Plattform dokumentieren werden.

Bei Interesse wird in diesem Kontext über die Veranstaltung hinaus auch die Entwicklung von gemeinsamen *Fallstudien* angeregt. Für Teilnehmer*innen wird in diesem Fall ein persistenter Zugang zur InTaVia-Plattform geschaffen, über den die Auswahl oder der Import von eigenen Daten mit Bezug zu individuellen Forschungs- und Vermittlungsthemen möglich ist. So wird ein extensiver Austausch zu den Möglichkeiten und Grenzen der Plattform möglich. Expert*innen können die Plattform nutzen, um neue Designs und Vermittlungsmethoden für ihre eigenen Daten und Themen zu gewinnen und um diese im Rahmen von gemeinsamen Fallstudien für eigene kommunikative Zwecke zu nutzen. Feedback zu den Möglichkeiten und Grenzen der Plattform wird wiederum dem Konsortium wertvolle Einblicke in die entscheidenden Bedürfnisse von Praktiker*innen liefern.

Ablauf Workshop

1) Projektvorstellung: Die InTaVia-Plattform verknüpft Datensammlungen verschiedenen Typs (i.e. kulturelle Objektsammlungen und biografische Textsammlungen) zu einer integrierten Graphdatenbank (Abbildung 1). In einer kurzen Vorstellung werden die wichtigsten Forschungsfragen des Projekts gemeinsam mit seinen technologischen Zielen und Modulen vorgestellt. Dies inkludiert Information über das Modul zur manuellen Kuratierung dieser Daten (Data Curation Lab) und das Modul zum visuellen Storytelling (Visual Storytelling Suite).

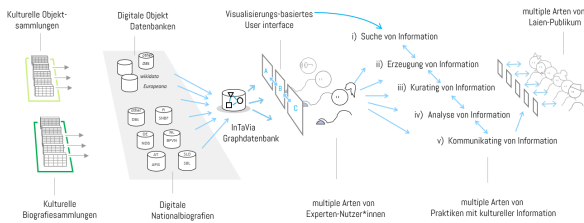


Abbildung 1: Architektur der InTaVia Plattform

2) Hands-On-Vorstellung des Storytelling-Moduls: Eine kurze Vorstellung der integrierten Graphdatenbank (IKG - InTaVia Knowledge Graph) wird zu einem Verständnis des zugrundeliegenden Datenmodells führen. Auf diese Weise werden Teilnehmende mit wichtigen Aspekten der Lebens- und Werkdaten vertraut, deren narrative Vermittlung die InTaVia-Plattform unterstützt. Dies ist von besonderer Relevanz für die Möglichkeit der manuellen Aufbereitung und Zusammenführung von Kulturdaten (sowohl Biografie- wie auch kulturelle Objektdaten), welche in einem eigenen Datenkuratierungs-Modul angesiedelt ist. Anhand einer Auswahl von Arbeitsdaten für den Workshop werden hierbei die Möglichkeiten aufgezeigt, die sich aus einer etwaigen Nutzung der Plattform für eigene Fallstudien ergeben.

Kulturelle Objektdaten und Biografiedaten haben eine Vielzahl von Facetten und Dimensionen die für Historiker*innen und Kulturwissenschaftler*innen von Interesse sein können. Zu diesen Dimensionen zählen die geografische Position von biografischen oder künstlerischen Ereignissen, diverse Kategorien von Ereignissen oder kulturellen Entitäten (Objekte oder Personen), Relationen zwischen Personen und/oder Objekten, sowie chronologische Abfolgen und Zusammenhänge. Diese Aspekte können auf verschiedenen Ebenen der Aggregation visualisiert werden - und in der Folge mit Medien, Texten und interaktiven Elementen angereichert und narrativ vermittelt werden (vgl. Abbildung 2). Der Workshop wird zu diesem Zweck das Visualisierungsmodul der InTaVia-Plattform kurz einführen, um den Schwerpunkt auf die praktische Gestaltung von Geschichten mit exemplarischen Objekt- und Akteursdaten zu legen.

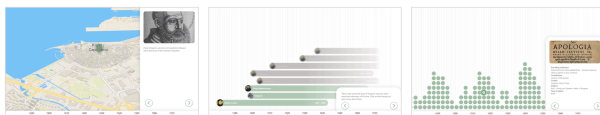


Abbildung 2: Ausschnitte aus einem Storyboard mit geographischer, temporaler und Häufigkeitsvisualisierung (v. links nach rechts)

3) Feedback: Während der explorativen Arbeit mit den Modulen der Plattform werden Fragen und Hinweise der Teilnehmer*innen notiert um im Rahmen der weiteren Arbeit am Forschungsprojekt in die nutzer*innen-zentrierte Entwicklung der Plattform einfließen zu können. Dazu werden sowohl die anonymisierten Notizen zu Aktivitäten des 'lauten Denkens' von Teilnehmer*innen dienen, wie auch die Rückmeldungen eines kompakten strukturierten Feedbackbogens.

Format:

Der Workshop ist als Halbtagesveranstaltung konzipiert mit Fokus auf die Erkundung, Erprobung und Diskussion von Methoden der narrativen Kulturdatenvisualisierung. Seine intendierte Zielgruppe reicht von interessierten Praktiker*innen aller kulturellen Institutionen bis hin zu Historiker*innen und Expert*innen der digitalen Geisteswissenschaften mit Interesse an Storytelling und Wissensvermittlung. Für die Teilnahme gibt es keine Voraussetzungen mit Blick auf inhaltliches oder technisches Vorwissen. Für die praktische Arbeit an den Daten genügt die Mitnahme eines Laptops. Die Gruppengröße ist auf 30 Teilnehmer*innen beschränkt. Für die technische Raumausstattung wird ein Beamer, ein Medienkoffer, sowie Whiteboards oder Pinnwände beantragt. Fördernachweis: Das Projekt InTaVia (<https://intavia.eu>) wird von der Europäischen Kommission im Rahmen des H2020 Research and Innovation Programme, Grant Agreement No. 101004825 gefördert.

Bibliographie

Bolin, Hans. 2010. "The re-generation of mythical messages: Rock art and storytelling in northern Fennoscandia". *Fennoscandia Archaeologica* 27: 21-34.

Dykes, Brent (2019). *Effective data storytelling: how to drive change with data, narrative and visuals*. John Wiley & Sons.

Khan, Nadim Akhtar, Shafi, S. M., and Ahangar, Humma. 2018. "Digitization of cultural heritage: Global initiatives, opportunities and challenges." *Journal of Cases on Information Technology (JCIT)* 20: 1-16.

Khulusi, Richard, Kusnick, Jakob, Focht, Josef, and Jänicke, Stefan. 2019. "An interactive chart of biography". In *2019 IEEE Pacific Visualization Symposium (PacificVis)*, 257-266. IEEE.

Kusnick, Jakob, Jänicke, Stefan, Doppler, Carina, Seirafi, Kasra, Liem, Johannes, Windhager, Florian, and Mayr, Eva. 2021. "Report on narrative visualization techniques for OPDB data". Deliverable of the H2020 project InTaVia. Online

Riche, Nathalie Henry, Hurter, Christophe, Diakopoulos, Nicolas, & Carpendale, Sheelagh. 2018. *Data-driven storytelling*. CRC Press.

Schlögl, Matthias, Windhager, Florian, Mayr, Eva, und Kaiser, Maximilian. 2019. *Biographische Informationssysteme (DPBs, Digital Knowledge Databases, Virtual Research Environments)* [Data set]. Zenodo. 10.5281/zenodo.2593761

Segel, Edward, and Heer, Jeffrey. 2010. "Narrative visualization: Telling stories with data". *IEEE transactions on visualization and computer graphics* 16: 1139-1148.

ter Braake, Serge, Fokkens, Antske S., Sluijter, Ronald, and Declerck, Thierry. 2015. *Biographical Data in a Digital World 2015: Proceedings of the First Conference on Biographical Data in a Digital World (BD2015)*. CEUR Workshop Proceedings. <http://ceur-ws.org/Vol-1399/>

ter Braake, Serge, Fokkens, Antske, Sluijter, Ronald, Arthur, Paul, and Wandl-Vogt, Eveline. 2018. *Biographical Data in a Digital World 2017: Proceedings of the Second Conference on Biographical Data in a Digital World 2017*

(BD2017) . CEUR Wokshop Proceedings, 2119 . <http://ceur-ws.org/Vol-2119/>

Windhager, Florian, Federico, Paolo, Schreder, Günther, Glinka, Katrin, Dörk, Marian, Miksch, Silvia, and Mayr, Eva. 2018. "Visualization of cultural heritage collection data: State of the art and future challenges". *IEEE transactions on visualization and computer graphics* 25: 2311-2330.

Windhager, Florian, Mayr, Eva, Schlögl, Matthias, and Kaiser, Maximilian. 2022. "Visuelle Analyse und Kuratierung von Biographiedaten". In *Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft*, ed. K. Döring et al., 137-150. Amsterdam: DeGruyter. 10.1515/9783110757101-008

Data Feminism in DH: Hackathon und Netzwerktreffen

Lang, Sarah

sarah.lang@uni-graz.at
Universität Graz, Österreich

Borek, Luise

luise.borek@tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Probst, Nora

nora.probst@uni-koeln.de
Universität zu Köln, Deutschland

Einleitung

Das Ziel von *Data Feminism* ist es, ausgehend von Positionen des intersektionalen Feminismus die vorwiegend männlich dominierten Narrative in den Digital Humanities kritisch zu reflektieren und bislang marginalisierten Stimmen insbesondere von Frauen* mehr Sichtbarkeit zukommen zu lassen. In den Geisteswissenschaften ist häufig zu beobachten, dass sich der weiß, cis-männlich und hegemonial dominierte Bias aus den Archiven und Quellen in die erhobenen Daten und digitalen Technologien überträgt. Die vom *Data Feminism* angestrebte Balance im Sinne einer gendersensiblen Repräsentation ist ein wichtiger Baustein, um der Forderung nach einer grundlegenden Offenheit geisteswissenschaftlicher Forschung nachzukommen.

Data Feminism in den DH

Das 2020 von Catherine D'Ignazio und Lauren F. Klein veröffentlichte Buch *Data Feminism* (MIT Press) hat für die Digital Humanities eine hohe Relevanz, wendet sich allerdings eher an ein Publikum aus der Data Science.

Zwar bieten Kleins Forschungen durchaus Schnittstellen zu Fragestellungen in den DH, doch lässt sich aus dem Buch keine unmittelbar anwendbare Anleitung für die Umsetzung von Data Feminism in den Digital Humanities generieren. Vielmehr handelt es sich um ein Manifest, das vor allem die Wichtig- und Dringlichkeit von intersektionalen, datenfeministischen Ansätzen deutlich macht und grundlegende Leitsätze zu deren Umsetzung formuliert. Diese Leitsätze auf konkrete Forschungsfragen oder Datenrepositorien aus den Digital Humanities zu übertragen, ist nach wie vor ein Desiderat.

Hier setzt unser Workshop an. Gemeinsam mit den Teilnehmenden soll eine Handreichung erarbeitet werden, die grundlegende Informationen und Leitlinien bezüglich folgender Fragen zusammenstellt:

1) Was lässt sich unter dem Begriff Data Feminism verstehen?

2) Wieso braucht man Data Feminism in den DH und inwiefern sollten dessen Ansätze in den bestehenden Diskursen der DH Berücksichtigung finden?

3) Wie können datenfeministische Projekte in den DH konkret aussehen? Inwiefern lassen sich Forschungsfelder definieren, die für Interessierte einen guten Ausgangspunkt bilden, um einige Ansätze des Data Feminism unmittelbar und praktisch umzusetzen?

Ein solche Handreichung sollte die Umsetzung datenfeministischer Ansätze in den DH insofern befördern, als sie die Einstiegshürden senkt und Interessierten diejenigen Informationen zur Verfügung stellt, die sie für einen effizienten, konkreten und praxisbezogenen Einstieg in Data Feminism benötigen.

Forschungsstand: Offene Fragen und Desiderate im Kontext des Data Feminism in den DH

Problemstellung 1: Data Gender Gap und das kulturelle Erbe

Beim Arbeiten mit Quellen zum kulturellen Erbe wird eine Vielzahl von Problemen offenbar, die bei vielen anderen Datensätzen (die beispielsweise auf statistischen Erhebungen und Befragungen beruhen) nicht bestehen und die sich genuin aus dem geisteswissenschaftlichen Forschungsgegenstand ergeben: So lassen sich Positionen des Data Feminism zum Teil gar nicht auf historische Daten anwenden, wenn beispielsweise Informationen über Frauen* oder andere Gruppen, die der hegemonialen Norm nicht entsprechen, gar nicht erst dokumentiert wurden (Mandell 2019, Lampe 2021, Rezai 2022). Nicht nur in den materiellen Beständen von vielen GLAM-Institutionen, sondern auch in den daraus hervorgegangenen Datenbeständen sind Frauen* unterrepräsentiert (Flanders 2018, Wiens et al. 2020). Diese Problematik eines Gender Data Gaps (Criado-Perez 2020) spiegelt sich auch in den verfügbaren Normdaten wider, da – etwa in der GND – zunächst vor allem publizierende Personen systematisch erfasst wurden.¹ Auch eine adäquate Verschlagwortung etwa in Bibliothekskatalogen

stellt in diesem Zusammenhang eine große Herausforderung dar (vgl. Juen 2021).

Bevor also eine datenfeministische Analyse überhaupt beginnen kann, ist umfangreiche Erschließungsarbeit zu leisten und es ist je nach konkretem Anwendungsfall gegebenenfalls notwendig, bestehende Korpora um Quellen von oder über Frauen* zu erweitern. Ein Beispiel für solche Arbeiten, bei denen bestehende Datensätze mit bislang unterrepräsentierten Gruppen ergänzt werden, ist das Women Film Pioneers Project (WFPP) (<https://wfpp.columbia.edu/>, vgl. Wreyford/Cobb 2017, Dang 2020, Dickel et al. 2022, Gaines et al. 2022). In Themenfeldern, in denen Daten von Frauen* zwar vorhanden, aber noch tiefergehend erschlossen oder digitalisiert werden müssen, ist dies arbeitsaufwendig und wird erfordern, dass die DH auf Dauer ausreichend Energie und Fördergelder in solche Tätigkeiten investieren. Wenn es also um Datenfeminismus und intersektionale Arbeit im Kontext der Digital Humanities geht, handelt es sich oftmals noch gar nicht um Datenanalysen, sondern vielmehr um erschließungsbezogene Tätigkeiten. Denn es gilt zunächst, die statistischen Ungleichgewichte der uns zur Verfügung stehenden Datensätze zum Thema zu machen und in diesem Sinne auf Lösungsstrategien hinzuarbeiten.

Eine weitere Herausforderung besteht hier ausgerechnet in der erstrebenswerten Offenheit von Daten, die inzwischen bei allen Förderinstitutionen und -richtlinien verlangt wird: Für marginalisierte und oftmals sensible Daten kann diese Offenheit aus Datenschutzgründen jedoch schnell zum Ausschlusskriterium werden. Hier müssen förderpolitisch Lösungen implementiert werden, die eine Erschließung und wissenschaftliche Nutzung ermöglichen, ohne den Datenschutz zu verletzen. Nicht zuletzt stellen Fälle, in denen Frauen* oder unterrepräsentierte Gruppen in Datensätzen unsichtbar, schwer sichtbar oder überhaupt nicht vorhanden sind, eine schwer zu lösende Aufgabe dar. Was ist mit Fällen, in denen es nicht möglich ist, diese Lücke durch neue Datenerhebungen zu füllen? Kann man unterrepräsentierte Daten aus bestehenden Datensätzen interpolieren?

Problemstellung 2: Modellierung, Kuratierung, Daten- und Korpuskritik

Positionen des Data Feminism halten bereits insofern Einzug in die DH, als in den vergangenen Jahren Hege- monie-kritische Forschungsthemen wie Gender-sensible Datenmodellierung, Korpuskritik, Decolonializing DH usw. stärkere Berücksichtigung finden (Risam 2015, Wernimont 2015, Koh/Stommel 2018, Losh/Wernimont 2018, Kim/Koh 2021, Guiliano/Heitmann 2019, Mandell 2019, Risam/Bordalejo 2019). Da ein feministischer und intersektional angelegter Umgang mit Daten in der Regel mit einer Analyse von Machtstrukturen beginnt, bieten die DH einen besonders geeigneten Nährboden für diesen Ansatz. Grundsätzlich gilt es nicht nur, dem Gender Data Gap durch die Erhebung von Daten zu Frauen* entgegenzuwirken, sondern es müssen auch die bestehenden Kategorien, Klassifizierungen und Datenmodelle auf Grundlage intersektionaler feministischer Perspektiven kritisch hinterfragt und ggf. komplexer gestaltet werden (Kyvernitou/Bikakis 2017) – dazu gehören Überle-

gungen zu Marginalisierungen an der Intersektion von Gender und rassistischen, klassistischen, ableistischen und weiteren Diskriminierungserfahrungen. Denn personenbezogenen Daten lassen sich in der Regel nicht durch Selbstbezeichnungen und/oder zusätzliche Befragungen ergänzen – Gender-bezogene Daten sind bei historischen Datensätzen in der Regel Zuschreibungen, die auf äußerlichen Merkmalen beruhen. In aller Regel erlauben diese Datensätze weder eine Unterscheidung zwischen biologischem Geschlecht und Gender, noch gehen sie über das binäre Geschlechtermodell hinaus. Data Feminism muss in den DH nicht nur den Data Gender Gap im Blick behalten und die Digitalisierung von bislang nicht digitalisierten Quellen forcieren, sondern auch bestehende Klassifizierungen, Datenmodelle und -kuratierungen kritisch hinterfragen.

Problemstellung 3: Anwendung außerhalb offensichtlicher Anwendungsfälle

Online finden sich eine ganze Reihe von Arbeiten, die sich des Data-Feminism-Begriffs bereits bedienen, doch handelt es sich häufig noch um Work-in-Progress, das noch nicht in Form von wissenschaftlichen Publikationen veröffentlicht ist (Bui/Gleißner/Kühn/Neeninger 2021, Keck 2021a, 2021b, Klein 2018, 2022). Zudem ergibt sich hier die Auseinandersetzung mit Data Feminism häufig unmittelbar aus einer thematischen Schwerpunktsetzung heraus. Wie aber könnte man mit Datensätzen oder Forschungsfragen umgehen, die nicht für Ansätze des Data Feminism prädestiniert scheinen, aber dennoch von diesen Ansätzen in hohem Maße profitieren würden? Um ein solches Paradigma zu etablieren, genügt es nicht, wenn sich lediglich einzelne Projekte für datenfeministische Perspektiven öffnen, bei denen der Nutzen des Data Feminism unmittelbar auf der Hand liegt. Um die Erkenntnisse der Intersektionalitäts- und Datenfeminismus-Forschung konsequent in den DH zu etablieren – wie unter anderem von Roopika Risam 2015 gefordert –, werden Leitlinien benötigt, die aufzeigen, wie diese auf eine größere Breite an Datensätzen und Forschungsfragen Anwendung finden können.

Problemstellung 4: Maschinelles Lernen als Bias-Verstärker oder Chance?

Zu fragen wäre schließlich, ob in diesem Kontext Algorithmen des maschinellen Lernens Abhilfe leisten könnten. Bisherige Publikationen zur Daten- und Korpuskritik in Bezug auf marginalisierte Gruppen halten sich diesbezüglich eher bedeckt und betonen, dass Algorithmen vor allem Bias der sie vornehmlich programmierenden (weißen cis-männlichen) Gruppen (Klinger/Svensson 2021) reproduzieren (GUILIANO/HEITMANN 2019).

Andere Publikationen verweisen allerdings darauf, dass die Aufgabe der DH in Bezug auf maschinelles Lernen gerade darin besteht, den Maschinen jenen "computer science gaze" abzutrainieren und stattdessen einen "curatorial gaze" zu etablieren (Bönisch 2021). Zu fragen wäre folglich, ob Algorithmen des maschinellen Lernens programmiert werden könnten, die trotz der in Problem-

stellung 1 und 2 skizzierten Herausforderungen einen "intersectional decolonialist gaze" auf Daten des kulturellen Erbes etablieren.

Aufbau des Workshops

Ziel dieses Workshops ist zum einen die Vernetzung aller Forschenden, die Positionen des Data Feminism bereits in ihren Forschungen berücksichtigen oder perspektivisch berücksichtigen wollen. Zum anderen geht es um die Erarbeitung konkreter Umsetzungsstrategien, wie sich die Forderungen des Data Feminism im Kontext der DH realisieren lassen. Der Workshop nennt sich Hackathon, um die kollektive Arbeitsform im Sinne einer gemeinsamen Daten-bezogenen Arbeitspraxis zu betonen. Zu Anfang werden zunächst die grundlegenden Konzepte des Data Feminism vorgestellt, um den Einstieg auch Personen ohne Vorerfahrungen zu ermöglichen. Im Anschluss können Teilnehmende eigene Arbeiten kurz vorstellen und ihren Zugriff auf Data Feminism im Kontext der DH erläutern. Aus diesen Inputs ergibt sich dann ein Austausch in Kleingruppen. Aufbauend auf den Erläuterungen und Diskussionen zu Beginn erarbeiten die Teilnehmenden in der Folge gemeinsam eine Handreichung, die auf wenigen Seiten grundlegende Informationen zum praktischen datenfeministischen Arbeiten in den Digital Humanities vermittelt und konkrete Leitlinien enthält. Auf diese Weise können sich Interessierte durch die Lektüre grundlegende Kenntnisse des Data Feminism selbstständig aneignen. Wir hoffen, dass damit die Einstiegshürde in datenfeministische DH-Forschung gesenkt und Data Feminism zukünftig in den DH einen breiteren Rückhalt finden wird. Die Handreichung soll idealerweise auf Deutsch und Englisch herausgegeben werden. Um all diese Aspekte (Einführung in grundlegende Konzepte, Kurzvorstellung eigener Arbeiten, Kleingruppendiskussion sowie Erstellung der Handreichung) zeitlich abdecken zu können, werden zwei Halbtage für den Workshop veranschlagt.

Organisationsteam

Sarah Lang, Universität Graz, sarah.lang@uni-graz.at, <https://orcid.org/0000-0002-4618-9481>. Forschungsinteressen/Hintergrund: Digital Humanities PostDoc, Wissenschaftsgeschichte (Alchemie).

Luise Borek, Technische Universität Darmstadt, luisse.borek@tu-darmstadt.de, 0000-0001-5849-374X. Forschungsinteressen/Hintergrund: PostDoc (TU Darmstadt) Vertretungsprofessur Digital Humanities (Universität Graz); germanistische Mediävistik. DFG-Netzwerk Offenes Mittelalter.

Nora Probst, Universität zu Köln, Deutschland, nora.probst@uni-koeln.de, 0000-0001-6932-0879. Forschungsinteressen/Hintergrund: Digitale Kulturwissenschaften, Vertretungsprofessorin der Professur "Kulturen der Digitalität".

Alle Einreichenden verbindet durch ihre Aktivitäten in der AG Empowerment das Interesse an Korpuskritik und Datenfeminismus.

Fußnoten

1. Criado-Perez' Buch *Invisible Women* hat in den letzten zwei Jahren viel Aufmerksamkeit für das Thema generiert. Wir distanzieren uns allerdings ausdrücklich von dem in der Studie vertretenen binären Geschlechtermodell, das Personen jenseits der Cis-Geschlechtlichkeit vollständig ignoriert, und vertreten unsererseits einen intersektionalen, queerfeministischen und inklusiven Ansatz nach D'Ignazio/Klein 2020.

Bibliographie

Aleksander, Karin. 2014. "Die Frau im Bibliothekskatalog." *LIBREAS. Library Ideas* 25 (2014). <https://libreas.eu/ausgabe25/02alexander/> (zuletzt zugegriffen 02.08.2022).

Bönisch, Dominik. 2021. "The Curator's Machine: Clustering of Museum Collection Data through Annotation of Hidden Connection Patterns between Artworks". *International Journal for Digital Art History* 5 (Mai 2021): 5.20–5.35. <https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/75953> (zuletzt zugegriffen 02.08.2022).

Bui, Magdalena, Lea Gleißner, Fey Kühn und Amelie Nenninger. 2021. "Questioning Street Names Leipzig: wie Genderbias die Straßenbenennung in Leipzig beeinflusst und wie die DH helfen können Bias sichtbar zu machen." In: *Blog Public Humanities in den Digital Humanities*. 23.08.2021. <https://publicdh.hypotheses.org/282> (zuletzt zugegriffen 02.08.2022).

Criado-Perez, Caroline. 2020. *Invisible Women: Exposing Data Bias in a World Designed for Men*. London: Chatto & Windus.

Dang, Sarah-Mai. 2020. "Unknowable Facts and Digital Databases: Reflections on the Women Film Pioneers Project and Women in Film History." *Digital Humanities Quarterly* 14 (4). <http://www.digitalhumanities.org/dhq/vol/14/4/000528/000528.html>.

Dickel, Henri, Matija Miskovic, Kharazm Noori, Christian Schmidt, Atefeh Soltanifard, Sarah-Mai Dang und Thorsten Thormählen. *Women Film Pioneers Explorer*. <https://www.online.unimarburg.de/women-film-pioneers-explorer/> (zuletzt zugegriffen 02.08.2022).

D'Ignazio, Catherine und Lauren Klein. 2020. *Data Feminism*. Cambridge/MA: MIT Press. <https://datafeminism.io/> (zuletzt zugegriffen 02.08.2022).

Flanders, Julia. 2018. "Building Otherwise." In: *Bodies of Information: Intersectional Feminism and the Digital Humanities*, herausgegeben von Elizabeth Losh und Jacqueline Wernimont, 289–304. Minneapolis/MN: University of Minnesota Press. Debates in the Digital Humanities. DOI: <https://doi.org/10.5749/j.ctv9hj9r9> (zuletzt zugegriffen 02.08.2022).

Gaines, Jane, Radha Vatsal, and Monica Dall'Asta (eds.). 2022. *Women Film Pioneers Project*. New York, NY: Columbia University Libraries. <https://wfpp.columbia.edu/> (zuletzt zugegriffen 02.08.2022).

Guiliano, Jennifer und Carolyn Heitman (2019): „Difficult Heritage and the Complexities of Indigenous Data.“ *Journal of Cultural Analytics* 4/1: 1–25. doi:10.22148/16.044.

Juen, Sara. 2021. "Feminismus, Algorithmen, Gender-Data-Gap und was das alles mit Bibliotheks- und Informationswissenschaft zu tun hat." *LIBREAS. Library Ideas* 39 (2021). <https://doi.org/10.18452/23448> (zuletzt zugegriffen 02.08.2022).

Keck, Jana. 2021a. "Text Mining America's German-Language Newspapers, 1830-1914: Processing Ger(wo)manness." Historikertag 2021, Peter Haber Preis für digitale Geschichte. <https://doi.org/10.5281/zenodo.5518019> (zuletzt zugegriffen 02.08.2022).

Keck, Jana. 2021b. "A Data Feminist Approach to Studying the C19 Social Network of German-Americans." Global Digital Humanities Symposium 2021. <https://www.youtube.com/watch?v=zrg1htGXRIA> (zuletzt zugegriffen 02.08.2022).

Klein, Lauren. 2018. "Data feminism: Community, allyship, and action in the digital humanities" (Keynote address). Digital Frontiers Annual Conference, Lawrence, KS, United States. <https://digital.library.txstate.edu/handle/10877/7839> (zuletzt zugegriffen 02.08.2022).

Klein, Lauren. 2022. "Data Feminism and Digital Humanities" (Vortrag). <https://www.youtube.com/watch?v=W0siTS8a6YY> (zuletzt zugegriffen 02.08.2022).

Klinger, Ulrike und Jakob Svensson. 2021. "The power of code: women and the making of the digital world." *Information, Communication & Society* 24 (14): 2075-2090. DOI: 10.1080/1369118X.2021.1962947.

Kyvernitou, Ionna und Antonis Bikakis. 2017. "An Ontology for Gendered Content Representation of Cultural Heritage Artefacts." *Digital Humanities Quarterly* 11/3 (2017). <http://www.digitalhumanities.org/dhq/vol/11/3/000316/000316.html> (zuletzt zugegriffen 02.08.2022).

Lampe, Moritz. 2021. *Diskriminierende Begriffe und Wissensordnungen im Bildarchiv. Eine postkoloniale Perspektive am Beispiel des 'Bildindex der Kunst und Architektur'*. Berliner Handreichung zur Bibliotheks- und Informationswissenschaft 481. Berlin. DOI: 10.18452/23766 (zuletzt zugegriffen 02.08.2022).

Losh, Elizabeth und Jacqueline Wernimont. 2018. *Bodies of Information: Intersectional Feminism and the Digital Humanities*. Minneapolis/MN: University of Minnesota Press. Debates in the Digital Humanities. DOI: <https://doi.org/10.5749/j.ctv9hj9r9> (zuletzt zugegriffen 02.08.2022).

Mandell, Laura. 2019. "Gender and Cultural Analytics: Finding or Making Stereotypes?" In: *Debates in the Digital Humanities* 2019, herausgegeben von Matthew K. Gold und Lauren F. Klein. University of Minnesota Press. DOI: <http://dx.doi.org/10.5749/j.ctvg251hk.4> (zuletzt zugegriffen 02.08.2022).

Rezaei, Yasami. 2022. "Data Stories for/from All: Why Data Feminism is for Everyone." *Digital Humanities Quarterly* 16/2 (2022). <http://www.digitalhumanities.org/dhq/vol/16/2/000618/000618.html> (zuletzt zugegriffen 02.08.2022).

Risam, Roopika (2015). "Beyond the Margins: Intersectionality and the Digital Humanities." *Digital Humanities Quarterly* 9(2). <http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html> (zuletzt zugegriffen 02.08.2022).

Risam, Roopika und Barbara Bordalejo (eds.). 2019. *Intersectionality in Digital Humanities*. Amsterdam: Amsterdam University Press. DOI: <https://doi.org/10.1017/9781641890519> (zuletzt zugegriffen 02.08.2022).

Wiens, Brianna, Stan Ruecker, Jennifer Roberts-Smith, Milena Radzikowska und Shana MacDonald. 2020. "Materializing Data: New Research Methods for Feminist Digital Humanities." *Digital Studies/le Champ Numérique* 10(1). DOI: <https://doi.org/10.16995/dscn.373> (zuletzt zugegriffen 02.08.2022).

Wernimont, Jacqueline (ed.). 2015. *Digital Humanities Quarterly* 9.2 (2015), Special Issue *Feminisms in Digital Humanities*. <http://www.digitalhumanities.org/dhq/vol/9/2/index.html> (zuletzt zugegriffen 02.08.2022).

Wreyford, Natalie, und Shelley Cobb. 2017. "Data and Responsibility: Toward a Feminist Methodology for Producing Historical Data on Women in the Contemporary UK Film Industry." *Feminist Media Histories* 3/3 (2017): 107-132. doi:10.1525/fmh.2017.3.3.107.

Die perfekte digitale Open-Access-Publikation

Baum, Constanze

constanze.baum@hu-berlin.de
Humboldt-Universität zu Berlin

Dahnke, Michael

michael.dahnke@posteo.de
Niedersächsische Staats- und Universitätsbibliothek
Göttingen

Dinger, Patrick

patrick.dinger@uni-muenster.de
WWU Münster/Universitäts- und Landesbibliothek

Fadeeva, Yuliya

yuliya.fadeeva@uni-due.de
Universität Duisburg-Essen

Horstmann, Jan

jan.horstmann@uni-muenster.de
WWU Münster/Universitäts- und Landesbibliothek

Seltmann, Melanie Elisabeth-H.

melanie.seltmann@tu-darmstadt.de
Universitäts- und Landesbibliothek Darmstadt

Steyer, Timo

t.steyer@tu-braunschweig.de
Universitätsbibliothek Braunschweig

Thematische Einordnung

Offene und frei zugängliche digitale Publikationen sind die Grundbedingung für einen globalen und fairen wissenschaftlichen Austausch und Erkenntniszuwachs. Diesen Grundsätzen von Open Science (vgl. Bartling und Friesike (Hg.) 2014) steht das an maximal ökonomischer Ausschöpfung orientierte Geschäftsmodell global agierender Verlagskonzerne in der Regel entgegen. So wird durch restriktive Zugangsmöglichkeiten zu aktuellen wissenschaftlichen Publikationen oder kommerziell orientierte Geschäftsmodelle eine wachsende Ungleichheit zwischen den wissenschaftlichen Playern *Forschende*, *Bibliotheken* und *Verlage* geschaffen (vgl. z.B. Lauer 2022). Open Access (OA) kann eine wichtige Antwort auf diese ökonomisch bedingte Schieflage im Wissenschaftssystem sein. Die verschiedenen Ausformungen, Bedingungen und Möglichkeiten des offenen digitalen Publizierens sollen im Rahmen des Workshops der AG »Digitales Publizieren« (vgl. DHd 2022) auf der DHd-Konferenz 2023 thematisiert werden. Eine solche klärende Auseinandersetzung erscheint auch angesichts der teils verwirrenden Vielfalt der Open-Access-Modelle sinnvoll. Folgende Fragen dienen als Leitfaden durch den Workshop: Was ist heute bei einer digitalen OA-Publikation zu beachten? Welche rechtlichen und technischen Mindestanforderungen sollten erfüllt werden und welche Spezifika und Standards gelten für digitale Publikationen? Welche Fallstricke sind bei Open Access-Publikationen zu beachten? Was sind Hürden für (Nachwuchs-)Wissenschaftler*innen OA zu publizieren?

Besonderer Fokus wird auf die Frage gelegt, wie eine »perfekte digitale Publikation« aussehen könnte und welche Strategien zu einer erfolgreichen digitalen Publikation in unterschiedlichen Szenarien führen. Im heterogenen Feld der digitalen Publikationen gibt es selbstverständlich nicht die eine Mustervorlage, an der sich alle messen müssen. Stattdessen sollte auf die Einhaltung bestimmter Qualitätsstandards – bezogen auf eine bestimmte OA-Publikation und deren Einsatzzweck – hingearbeitet werden. Die Gemeinsamkeiten aller digitalen OA-Publikationen herauszuarbeiten und zu diskutieren soll demnach ein wesentlicher Bestandteil des offenen Workshops sein. Die sachorientierte Diskussion wird durch eine kurze Einführung in das Thema durch Mitglieder der DHd-AG »Digitales Publizieren« begleitet, sodass für Forschende aus allen Disziplinen der (Digital) Humanities, insbesondere auch Nachwuchswissenschaftler*innen, eine gemeinsame Diskussionsgrundlage geschaffen wird. Die Ergebnisse des 2021 überarbeiteten und neu veröffentlichten Arbeitspapiers der AG sollen dabei als Rahmen der Diskussion dienen sowie community-intern kritisch hinterfragt werden (vgl. AG Digitales Publizieren 2021).

Im Arbeitspapier der AG werden im Unterkapitel 3.2 Qualitätskriterien für eine Veröffentlichung vorgestellt, die auch Grundlage der thematischen Arbeit im Workshop sein werden:

- Die Leitlinien guter wissenschaftlicher Praxis werden beachtet.

- Nutzungsbedingungen der Publikation sind geklärt (z. B. durch Creative-Commons-Lizenzen; vgl. Creative Commons 2022).
- Open-Access-Empfehlungen wissenschaftlicher Institutionen oder des Open-Access-Netzwerks (vgl. Open Access Network 2022a) werden beachtet. Hier wird besonders auf die Berliner Erklärung (vgl. Max Planck Gesellschaft 2003) und die Vor- und Nachteile des Grünen vs. Goldenen Weges eingegangen (vgl. Open Access Network 2022b).
- Peer Review als Notwendigkeit der Qualitätssicherung: Blind or non-blind peer review? (vgl. AG Digitales Publizieren 2021, Abschnitt 4).

Ziel der Diskussion ist, diese Kriterien stärker ins Bewusstsein der Teilnehmenden zu bringen und gleichzeitig deren Notwendigkeit und Sinnhaftigkeit für digitale OA-Publikationen kritisch zu diskutieren. Technische Spezifikationen digitaler OA-Publikationen sind Geisteswissenschaftler*innen häufig weniger bewusst als die technischen Eigenschaften traditioneller Publikationen, insbesondere außerhalb der digitalen Geisteswissenschaften. Dazu werden im Workshop erstens die Dateiformate diskutiert, die sich besonders für eine Online-Veröffentlichung sowie die langfristige Archivierung eignen [PDF/A-Format (ISO 19005-1:2005)]. Zweitens thematisieren wir, warum digitale OA-Publikationen wie andere digitale Objekte mit einem Persistent Identifier (DOI, URN) versehen werden sollten. Idealerweise sollten auch die Autor*innen über einen Persistent Identifier wie eine ORCID in den Werken aufgeführt werden und damit referenzierbar in Erscheinung treten. Drittens eröffnen wir die Möglichkeit, über die Notwendigkeit der Verwendung internationaler Standards wie z. B. METS/MODS, EDM oder Dublin Core für die Erschließung, Speicherung und Archivierung der digitalen Objekte zu sprechen. Viertens sollte in diesem Zusammenhang unbedingt das DINI-Zertifikat für Open-Access-Publikationsdienste, insbesondere der entwickelte Kriterienkatalog, beachtet werden (Müller et al. 2019).

Schließlich ist zu diskutieren, welche finanziellen Unterstützungsmöglichkeiten es für Verlage gibt. Was sind Beispiele, Erfahrungen oder Herausforderungen, mit einem Verlag (und einem institutionellen Repositorium) Open Access (CC BY und CC BY-SA) zu publizieren? Gehört die Wahl des Verlages und die Finanzierung mit zu den Kriterien einer »guten« OA-Publikation? Neben diesen übergreifenden Fragen soll der Workshop auch Raum für das Weiterdenken des digitalen Publizierens bieten. Ein Angebot innerhalb des Workshops besteht daher in der Diskussion von Kriterien für hybride Publikationsformen am Beispiel von Monographien und Sammelbänden, die sowohl in digitaler als auch in gedruckter Form erscheinen. Der in Kürze erscheinende Leistungskatalog des Projekts AuROA vereint sowohl Aufgaben im Bereich digital enhancement als auch verschiedene Formen der inhaltlichen und prozessualen Qualitätskontrolle, z.B. durch Angaben zur Offenheit/Geschlossenheit der Begutachtung, zum Status der Reviewenden und zur Erfüllung bestehender Kriterien (u. a. DFG-Kodex, COPE, FAIR-Prinzipien, PRISM etc.). Durch die Einhaltung solcher Kriterien kann u.a. der Gefahr von Predatory Publishing durch Transparenz begegnet werden.

Beitragende (und Forschungsinteressen)

Constanze Baum ist wissenschaftliche Mitarbeiterin am Institut für deutsche Literatur an der Humboldt-Universität zu Berlin. Sie baute von 2014–2017 die Zeitschrift für digitale Geisteswissenschaften (ZfdG) als vollwertiges Open Access Journal auf. Ihre Forschungsinteressen richten sich u.a. auf Fragen des digitalen Publizierens für die Literaturwissenschaft und die Rolle der Digital Humanities für ihre Fachdisziplin.

Michael Dahnke hat 2017/2018 als Dozent für Digitalisierungskompetenz an der Universität Würzburg gearbeitet und war 2018/2019 für das Forschungsdatenmanagement im SFB 1187 der Universität Siegen verantwortlich. Als digitaler Editionsphilologe hat er sich seit 2017 in mehreren Editionsprojekten für die technische Koordination verantwortlich gezeichnet.

Patrick Dinger arbeitet seit 2020 an der Universitäts- und Landesbibliothek Münster/Universität Münster und ist als Referent für das digitale Service- und Sammlungsmanagement zuständig. Der studierte Historiker interessiert sich u.a. für die Digitalisierung und die digitale Präsentation von Kulturerbe, Standards, Infrastrukturentwicklung sowie digitale Publikationsformen.

Yuliya Fadeeva ist seit 2020 im Bereich Open Access (in den Geisteswissenschaften) an der Universitätsbibliothek Duisburg-Essen tätig. Zurzeit ist sie wissenschaftliche Mitarbeiterin im Projekt AuROA. Ihre Open-Access-bezogenen Forschungsinteressen liegen im Bereich des Qualitätsbegriffs (wissenschaftlicher Arbeiten/wissenschaftlichen Arbeitens) im wissenschaftssoziologischen und -theoretischen Kontext und der Implikationen des digitalen Publizierens für die Wissenschaftspraxis.

Jan Horstmann leitet das Service Center for Digital Humanities an der ULB der Westfälischen Wilhelms-Universität Münster. Seine Forschungsinteressen und -schwerpunkte liegen im Bereich der digitalen Methodologie mit besonderem Fokus auf die Textannotation, -analyse und Visualisierung im Bereich der computationalen Literaturwissenschaft. Infrastrukturell setzt er sich ein für die Einhaltung der FAIR- und CARE-Prinzipien einer nachhaltigen und offenen Wissenschaft.

Melanie Seltmann arbeitet seit 2021 im Zentrum für digitale Editionen der ULB Darmstadt und ist dort für das Citizen-Science-Projekt Gruß & Kuss sowie für das NFDI-Konsortium Text+, Task Area Editions zuständig. Die studierte Linguistin interessiert sich u.a. für die Bereiche Citizen Science, Wissenschaftskommunikation, Open Science, Standards sowie insbesondere für Annotationen.

Timo Steyer leitet das Referat Informationskompetenz an der Universitätsbibliothek Braunschweig und ist Fachreferent für die Fächer Anglistik, Germanistik und Geschichte. Zu seinen Forschungsinteressen zählen neben dem digitalen Publizieren vor allem die Bereiche Datenmodellierung und Metadaten sowie aktuelle Entwicklungen in der Wissenschaftskommunikation.

Format und Zeitplan

Der Workshop soll als interaktives Diskussionsformat mit Gruppenarbeitseinheiten insgesamt 4 Stunden dauern, die folgendermaßen strukturiert sein werden:

0–0:15: Vorstellung der AG ›Digitales Publizieren‹ und der Workshopteilnehmenden

0:15–0:30: Vorstellung des AG Working Papers

0:30–0:45: Eingrenzung des Themas; Kurzvorstellung der Tische

0:45–1:30: Bearbeitung von Einzelthemen/Teilthemen in Worldcafe-Tables

1:30–2:00: Pause

2:00–3:00: Bearbeitung von Einzelthemen/Teilthemen in Worldcafe-Tables

3:00–3:45: Übertragung der Gruppenergebnisse in die Gesamtdiskussion; Rückbinden/Überprüfung der Working Paper-Überlegungen

3:45–4:00: Ausblick

Zielpublikum

Der Workshop richtet sich an Forschende aus allen Disziplinen der (Digital) Humanities, insbesondere auch Nachwuchswissenschaftler*innen, die sich aufgrund des systembedingt notwendigen Renommee-Erwerbs häufig gezwungen sehen, in teuren und nicht offen zugänglichen Publikationsorganen zu publizieren, und die alternative Publikationsmöglichkeiten wie in Bibliotheksverlagen oder scholar-led Publikationsorganen häufig nicht in Betracht ziehen. Grundvoraussetzung für die Teilnahme ist lediglich das Interesse an zeitgemäßen und zukunftsweisenden Formen des digitalen Publizierens im Sinne der Open Science. Es werden keine besonderen Kenntnisse vorausgesetzt. Möchte man sich vorab näher in die Materie einarbeiten, böte es sich an, das Workingpaper der AG ›Digitales Publizieren‹ (2021) zu konsultieren. Es können bis zu 25 Personen am Workshop teilnehmen.

Lernziele

Nach diesem Workshop sind die Teilnehmenden sensibilisiert für Möglichkeiten des Open-Access-Publizierens und Möglichkeiten, selbst das richtige Publikationsorgan für ihre zukünftigen Veröffentlichungen zu finden. Sie haben einen Überblick über verschiedene Publikationsorgane und deren Umgang mit Open-Access. Zudem werden sie darin unterstützt, selbstbewusst Verlagen gegenüberzutreten und ihre Bedingungen, Open Access zu publizieren, zu vertreten.

Benötigte technische Ausstattung

Benötigt werden ein Beamer, Moderationskoffer, offene Bestuhlung und Tische.

Bibliographie

AuROA - Autor:innen und Rechtssicherheit für Open Access. Im Erscheinen. *Leistungskatalog für wissenschaftliche Open-Access-Publikationen*. <https://projekt-auroa.de/veroeffentlichungen-veranstaltungen/#veroeffentlichungen> (zugegriffen: 1. August 2022).

AG Digitales Publizieren. 2021. *Digitales Publizieren in den Geisteswissenschaften: Begriffe, Standards, Empfehlungen* 10.17175/WP_2021_001.

Bartling, Sönke und Sascha Friesike, Hrsg. 2014. *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*. Springer Cham 10.1007/978-3-319-00026-8.

Max Planck Gesellschaft. 2003. *Berliner Erklärung*. <https://openaccess.mpg.de/Berliner-Erklärung> (zugegriffen: 26. Juli 2022).

COPE - Committee on Publication Ethics. <https://publicationethics.org/> (zugegriffen: 1. August 2022).

Creative Commons. 2022. <https://creativecommons.org/> (zugegriffen: 26. Juli 2022).

DHd. 2022. "AG Digitales Publizieren". *digital humanities im deutschsprachigen raum*. <https://dig-hum.de/ag-digitales-publizieren> (zugegriffen: 26. Juli 2022).

Lauer, Gerhard. 2022. "Datentracking in den Wissenschaften: Wissenschaftsorganisationen und die bizarre Asymmetrie im wissenschaftlichen Publikationssystem." *o-bib. Das offene Bibliotheksjournal* / Herausgeber VDB 9, Nr. 1 (31. März): 1-13 10.5282/o-bib/5796.

Müller, Uwe, Frank Scholze, Paul Vierkant, Ursula Arning, Daniel Beucke, Ute Blumtritt, Karolin Bove, u. a. 2019. *DINI-Zertifikat für Open-Access-Publikationsdienste* 2019 (Oktober) 10.18452/20545.

Open Access Network. 2022a. <https://open-access.network/startseite> (zugegriffen: 26. Juli 2022).

Open Access Network. 2022b. "Grün und Gold." <https://open-access.network/informieren/open-access-grundlagen/open-access-gruen-und-gold> (zugegriffen: 26. Juli 2022).

PRISM - Peer Review Information Service for Monographs. <https://www.doabooks.org/en/librarians/prism> (zugegriffen: 1. August 2022).

3D- und 4D-Modellierung in den Digital Humanities. Eine praktische und theoretische Einführung in Blender

Hunziker, Manuel

manuel.hunziker@lmu.de
Ludwig-Maximilians-Universität München

von Pippich, Waltraud

waltraud.v.pippich@kunstgeschichte.org
Ludwig-Maximilians-Universität München

Rensinghoff, Berenike

berenike.rensinghoff@hotmail.de
Akademie der Wissenschaften und der Literatur Mainz

Der zweitägige Workshop bietet eine Einführung in die 3D-Bearbeitungssoftware Blender (Blender 3.3 LTS). Blender ist eine freie, offen zugängliche und kostenlose 3D-Grafiksoftware, die ursprünglich vor allem in der Videospielbranche reüssieren konnte. Doch mit dem Erstarken der digitalen dreidimensionalen (3D) Modellierung und Rekonstruktion im Bereich der Denkmalwissenschaften, Kulturwissenschaften, Geschichte und ähnlicher Disziplinen, findet sich zunehmend Befürwortung auch aus dem wissenschaftlichen Umfeld. Mit der 3D-Rekonstruktion werden Forschungsgegenstände und -ergebnisse räumlich greifbar und nachvollziehbar. Das Modell wird förmlich zum digitalen Wissensspeicher.

Zielsetzung des Workshops

Der Workshop soll eine Plattform für ein interessiertes Publikum bilden, das einen umfassenden Einblick in die Funktionsweise von Blender und die daraus resultierenden Workflows erhalten möchte. Vorkenntnisse in der 3D-Modellierung oder mit der Modellierungssoftware sind nicht erforderlich. Fachwissenschaftlich sind für die Einführung keine Grenzen gesetzt, da jede Disziplin und unterschiedliche Forschungsfelder der Digital Humanities eigene Aspekte mit einbringen können und die 3D-Softwarelösung Blender äußerst flexibel einsetzbar ist. Durch Erweiterungen und Zusatzkomponenten kann diese an das eigene, spezifische Arbeitsumfeld angepasst werden.

Die Teilnehmer*innen des Workshops erwartet ein umfassender Einstieg in die Thematik der 3D-Rekonstruktion, der 4D-Komponente und den aus den digitalen Technologien resultierenden Möglichkeiten für Forschung, Lehre und Kunst- und Kulturvermittlung. Durch die Erweiterung um die vierte Dimension lassen sich Veränderungen und Prozesse, z.B. zeitliche Abfolgen, darstellen. Neben einer an Praxisbeispielen orientierten Einführung in die 3D-Software Blender wird auch eine kritische Diskussion über die Grenzen und Möglichkeiten von 3D-/4D-Rekonstruktionen und -Modellierungen mit den Teilnehmer*innen geführt.

Am Ende der Veranstaltung soll jede*r Teilnehmer*in reflektiert mit der Thematik der digitalen 3D- und 4D-Modellierung umgehen können, die Grundfunktionen der Software Blender beherrschen und eine Vorstellung von dem Spektrum möglicher Forschungsfragen haben. Zu den Ergebnissen des Workshops gehören unterschiedliche Modelle, die den Teilnehmer*innen auch im Nachhinein zur weiteren Bearbeitung zur Verfügung stehen.

3D- und 4D-Methoden für die Digital Humanities

Methoden, die eine räumliche Visualisierung, eine Rekonstruktion von Orten oder die Darstellung historischer Verläufe und sukzessive Reihenfolgen in der Zeit ermöglichen, können für sämtliche historische Fragestellungen relevant werden. Über die Fragen der Rekonstruktion und Vermittlung von vergangenen Zuständen hinaus ist auch die grundsätzliche Frage nach dem epistemologischen Stellenwert von 3D- und 4D-Methoden betroffen: Welche Aspekte und Zusammenhänge fallen rein durch Visualisierung in den Fokus der Wissenschaft? Über eine angenehme Hilfestellung für die Anschaulichkeit oder die Illustration bereits formulierter Thesen hinaus werden 3D- und 4D-Prozesse für bestimmte Konstellationen zum Mittelpunkt der Forschung selbst. Etwa, wenn historische Kausal-, Raum- und Größenverhältnisse erst durch die Anschauung ihre Prägnanz erhalten. Inwiefern wären die Drei- und Vierdimensionalität jedoch „Science Fiction“, so dass die Maximen der wissenschaftlichen Objektivität und der Quellenkritik berührt würden? Könnten die Ergebnisse von Modellierungssoftware unser Erkenntnisinteresse irritieren oder fehlleiten? Im Workshop werden an den jeweiligen Stellen im Rahmen der Lehr-Module diese, die Theorie der Wissenschaftlichkeit der digitalen Modellierung betreffenden, Themen diskutiert.

Virtuelle Modelle sind zunehmend Teil des Arbeitsprozesses für Forschungsfragen in den Digital Humanities. Es existieren unterschiedliche Methoden, die wiederum abhängig vom Gegenstand und Ziel des Modells sind. Neben Laserscans und photogrammetrischen Aufnahmen sind auch digitale Modellierungen eine Option. Im Gegensatz zum Scan und zur Photogrammetrie, bei denen nur existierende Objekte innerhalb eines Zustandes aufgenommen werden können, kann mit Hilfe der Modellierung ein Objekt gleichsam von Null auf erstellt oder ein bestehendes Modell ergänzt werden. Diese Methode kann zum Beispiel dann zum Einsatz gelangen, wenn ein nicht mehr bestehender Gebäudezustand dargestellt werden soll. Diese digitale 3D-Modellierung hilft bei der Visualisierung von Vergangenem. Wenn neben den optischen Gesichtspunkten der Modellierung auch zeitliche Komponenten wissenschaftlich betrachtet werden sollen, vermag sich die Dreidimensionalität um die vierte Dimension der Zeit zu erweitern. Im Vorfeld einer jeden Modellierung sollte ein kritischer wissenschaftlicher Diskurs stehen, um dem Modell so viele Informationen wie möglich geben zu können, jedoch nicht die Grenze zur freien Interpretation zu überschreiten.

Grundbegriffe der Theorie und Systematik der wissenschaftlichen Datenmodellierung, wie das System des „Level of Detail“ (LOD), werden ebenfalls in der Veranstaltung erklärt. Die fünf Detaillierungsgrade stufen die Genauigkeit des digitalen 3D-Modells ein: von LOD 0 mit dem sog. Geländemodell bis hin zu LOD 4, einer detaillierten Nachbildung von Innen- und Außenräumen eines Gebäudes (Münster 2019, 53). Durch diese Abstufungen wird die Datenökonomie, als Strategie im Datenlebenszyklus, gewährleistet. So kann beispielsweise eine Gebäuderekonstruktion unmittelbar in einen urbanen Kontext eingebunden werden, ohne eine um-

fassende Modellierung der Nachbarschaftsbebauung vorzunehmen. Weitere Gebäude werden entsprechend einer vereinfachten Kubatur in LOD 1 oder 2 modelliert.

Die durch die Software Blender ermöglichte Technologie kann als Baustein im Repertoire weiterer bestehender Technologien zur digitalen 3D- und 4D-Modellierung und -Rekonstruktion gelten. Die Beherrschung der im Workshop vorgestellten Technologien und die Kenntnis der vermittelten exemplarischen Workflows sollen als Erweiterung des Instrumentariums zur Forschungspraxis in den Digital Humanities zu verstehen sein.

Open Source Software ‚Blender‘

Blender entwickelte sich in den letzten Jahren immer mehr zu einer vielseitig einsetzbaren 3D-Softwarelösung. Heute führt an dieser freien und plattformunabhängigen Software für 3D-Modellierungen kaum ein Weg vorbei. Mit ihrem offenen Quellcode und der großen weltweiten Community, durch die Weiterentwicklung und Updates gesichert sind, erweist sie sich daher für den Bereich der Digital Humanities als empfehlenswert. Mit einer beständig wachsenden Community mit zahlreichen Online-Tutorials, Materialien oder gar ganzen Modellen, die vermehrt kostenlos genutzt werden können, wird die Hürde am Anfang auch für den ersten Einstieg niedrig gehalten. Auch fortgeschrittene Nutzer*innen profitieren von den zahlreichen Funktionen und Erweiterungen. Zudem wird Blender als Werkzeug für die Bearbeitung und Erstellung von 3D-Modellen von weiteren Programmen, wie Game Engines oder CAD-Software, unterstützt und erlaubt so einen vereinfachten Datenaustausch zwischen diesen Programmen. Für die wissenschaftliche Modellierung und Rekonstruktion liegen Funktionen wie der Import von Scan-Modellen oder 2D-Grundrissen die Verwendung nahe.

Ablauf des Workshops

Der Workshop findet an zwei Tagen mit jeweils vier Stunden Dauer statt. Im Workshop sollen Kenntnisse über die Grundfunktionen der 3D-Software Blender erlangt werden. Der erste Tag ist der Einführung in die Software gewidmet, am zweiten Tag folgen, aufbauend auf den vorangegangenen Modulen, weitere, fortgeschrittenere Übungen zu den Funktionen der Software.

Am ersten Workshop-Tag werden, nach einer allgemeinen, Überblickshaften Einführung in Chancen, Methoden und Forschungsfelder mit Bezug zu 3D-Software, erste Schritte der 3D-Modellierung vorgeführt und in die Grundlagen von Blender eingeleitet. Im anschließenden Modul liegt der Schwerpunkt auf den Grundprinzipien und Techniken einfacher Modellierung: Grundgeometrien, erste Kolorierung, Speichern, Darstellung und Rendering. Im anschließenden Modul werden Use Cases für 3D-Rekonstruktionen präsentiert. Der erste Workshop-Tag schließt mit einer Einführung in einen Workflow für den Forschungsprozess, Modul „Workflow I“: 2D-/3D-Pläne bzw. -Karten, Skalierung und Georeferenzierung.

Am zweiten Workshop-Tag erfolgt, nach einer kurzen Rekapitulation der vorangegangenen Module, der zweite Teil der Einführung in den Workflow innerhalb von Blender, das Modul „Workflow II“: Modellierung von Architektur, Texturierung, das Mapping von Bildern, das Thema Beleuchtungssituation sowie Rendern, Animation und Speichern/Export von Dateien. Schließlich werden die 3D-Fragestellungen um die vierte Dimension erweitert. Abschließend findet eine Schlussdiskussion statt.

Grundsätzlich sollen die Teilnehmer*innen befähigt werden, ein Verständnis für mögliche Forschungsfragen aus dem 3D-/4D-Bereich zu entwickeln. An beiden Tagen finden Diskussionen und kritische Reflektionen der Potentiale und Grenzen der erlernten Technologien für Forschungsfragen der Digital Humanities statt.

Zur Organisation

Max. Teilnehmer*innenzahl: 15

Teilnehmer*innen benötigen einen eigenen Laptop mit Internetzugang und vorab installierter Blender-Software, Version 3.3 LTS (Download: Blender 2022a, Handbuch: Blender 2022b). Die Systemanforderungen können über die Blender-Webseite abgefragt werden (Blender 2022d). Für die Bedienung der Software wird zudem eine Maus mit integriertem Mausrad benötigt. Der Workshop findet an zwei Tagen mit je vier Stunden Dauer statt. Technische Ausstattung vor Ort: Beamer und WLAN.

Die Referent*innen stellen im Vorfeld des Workshops ein digitales Dossier mit Informationsmaterial und Daten auf GitHub (<https://github.com/manuelhunziker/blendergoesdhd>) zur Verfügung.

Beitragende und Kontaktdaten

Manuel Hunziker (manuel.hunziker@lmu.de) (ORCID: 0000-0002-4684-938X) ist wissenschaftlicher Mitarbeiter und Dozent für Digitale Archäologie am Department für Kulturwissenschaften und Altertumskunde der Ludwig-Maximilians-Universität München. Als Kulturinformatiker setzt er zudem für das Museum für Abgüsse Klassischer Bildwerke in München digitale Strategien zur barrierefreien Vermittlung und zur musealen Inszenierung um. Er studierte Angewandte Informatik sowie Klassische Archäologie an der Universität Heidelberg und Denkmalpflege an der Universität Bamberg. Sein aktueller Forschungsschwerpunkt liegt in der Anwendung und Entwicklung computergestützter Verfahren zur dreidimensionalen Dokumentation, Visualisierung und virtuellen Rekonstruktion in den archäologischen Wissenschaften.

Waltraud von Pippich (waltraud.v.pippich@kunstgeschichte.org) (ORCID: 0000-0002-4555-2816) Forschungsprojekt zur 3D-Visualisierung von Farbräumen, Fraunhofer-Institut für Graphische Datenverarbeitung IGD Darmstadt, Forschungsprojekt „Rot rechnen“ zur digitalen Farbanalyse, Schwerpunkte in der Theorie der Digital Humanities, digitale Bild- und Farbanalyse, Methoden der Stilometrie von Bild und Text, Methoden der Visualisierung, Rechtsfragen in Forschung und Entwicklung, politische Ikonographie, digitale Kunstgeschichte.

Berenike Rensinghoff (berenike.rensinghoff@hotmail.de) (ORCID: 0000-0002-2717-3409) arbeitet als Trainee an der Akademie der Wissenschaften und der Literatur Mainz in der Abteilung Digitale Akademie für das Akademienvorhaben Corpus Vitrearum Medii Aevi (CVMA) Deutschland. Sie ist ebenfalls Mitorganisatorin des 3D Hackathons 2022 Creating New Dimensions. Im Rahmen der Masterarbeit im Studiengang Digitale Denkmaltechnologien an der Universität Bamberg und der Hochschule Coburg erstellte sie digitale 3D-Modelle und -Rekonstruktionen des Badezimmers des Palais Beauharnais in Paris. Aktuell modelliert sie für das CVMA Deutschland Glasmalereien mit Blender.

Bibliographie

Blender. 2022a. „Blender 3.3 LTS.“ In *blender.org*. <https://www.blender.org/download/lts/3-3/> (zugeschrieben: 14. Dezember 2022).

Blender. 2022b. „Blender 3.3 Reference Manual.“ In *Blender Documentation: User Manual*. <https://docs.blender.org/manual/en/3.3/> (zugegriffen: 14. Dezember 2022).

Blender. 2022c. „Home of the Blender project.“ In *blender.org*. <https://www.blender.org/> (zugegriffen: 14. Dezember 2022).

Blender. 2022d. „Requirements.“ In *blender.org*: Download. <https://www.blender.org/download/requirements/> (zugegriffen: 14. Dezember 2022).

Denard, Hugh (Hrsg.). 2009. London Charter. For the Computer-based Visualisation of Cultural Heritage. Draft 2.1. 7 February 2009. London. <http://www.london-charter.org/> (zugegriffen: 14. Dezember 2022).

Kohle, Hubertus. 2017. „Digitale Rekonstruktion und Simulation.“ In *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle, Malte Rehbein, 315–327. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05446-3_22.

Münster, Sander. 2019. „Die Begrifflichkeiten der 3D-Rekonstruktion.“ In *Der Modelle Tugend 2.0: Digitale 3D-Rekonstruktion als virtueller Raum der architekturhistorischen Forschung*, hg. von Piotr Kuroczyński, Mieke Pfarr-Harfst und Sander Münster, 38–57. Computing in Art and Architecture 2. Heidelberg: arthistoricum.net. <https://doi.org/10.11588/arthistoricum.515.c7444>.

Forschungssoftware rezensieren – Konzeption, Durchführung und Umsetzung

Homburg, Timo

timo.homburg@hs-mainz.de
Hochschule Mainz, Mainz, Deutschland

Klammt, Anne

aklammt@hotmail.com
Deutsches Forum für Kunstgeschichte Paris,
Frankreich

Offert, Fabian

offert@ucsb.edu
University of California, Santa Barbara, USA

Thiery, Florian

florian.thiery@rgzm.de
Römisch-Germanisches Zentralmuseum, Mainz,
Deutschland

Der Workshop adressiert eine signifikante Lücke in den Praktiken und Formaten der digitalen Geisteswissenschaften als *Open Humanities*: die bisher noch untergeordnete Rolle von Softwarerezensionen von *Forschungssoftware*.

Wissenschaftliche Rezensionen von Forschungssoftware schätzen deren Beitrag zur Lösung einer Aufgabe im Forschungsprozess, ihre Anwendbarkeit und Zielgruppe, sowie ihre handwerkliche Qualität und Nachhaltigkeit ein. Softwarerezensionen wären daher ein wesentlicher Bestandteil einer Wissenschaftspraxis, die ihre Methoden offenlegt und kritisch reflektiert. Bislang aber erscheinen noch wenige Rezensionen und das, obwohl es mittlerweile in den Geisteswissenschaften im deutschsprachigen Raum erste Zeitschriften zu ihrer Veröffentlichung gibt. Es finden sich jedoch noch kaum Autor*innen.¹ Der Workshop soll Interessierten einen Einstieg bieten und somit zur Verbreitung und Anerkennung dieses wichtigen wissenschaftlichen Formats beitragen.

Gemeinschaftlich mit den Teilnehmer*innen möchten wir am Beispiel kleiner, überschaubarer Tools alle Arbeitsschritte einer Rezension von Forschungssoftware am praktischen Beispiel durchführen. Im Mittelpunkt steht das Kennenlernen und Anwenden von Aspekten, mit denen eine Software besprochen werden kann. Die Beurteilung von Forschungssoftware muss ihre Aufgabe in der Forschung, ihre Nutzbarkeit aus Anwender*innensicht und ihre Nachhaltigkeit aus technischer Sicht umfassen. Damit fragen Rezensionen von Software ganz unterschiedliche Kompetenzen ab und sind daher besonders gut und effizient als Team zu bearbeiten. Im Idealfall entsteht während des Workshops genug Material, um unterstützt von den Workshopenbieter*innen ohne umfangreiche Nacharbeiten eine Rezension beim Journal CKIT oder den Archäologischen Informationen einzureichen.

Forschungssoftware ist ein wesentlicher Bestandteil (digitaler) geisteswissenschaftlicher Forschung. Sie ermöglicht und steuert oft den gesamten Forschungsprozess (Katerbow u. Feulner 2018; Schmidt u. Marwick 2020). Dieser Einfluss reicht von der Auswahl und Strukturierung der Daten, über die angewandten rechnerischen Verfahren bis hin zu den Ausgabeformaten. Entsprechend können fehlerhaft implementierte Algorithmen, irreführende Nutzeroberflächen und unvollständige Dokumentationen die Forschung erschweren oder sogar zum Scheitern von Forschungsprozessen führen.

Auch bei einem positiven Verlauf schreiben sich die Tools so tief in die Ergebnisse ein, dass eine digitale Quellenkritik immer wieder auch die Betrachtung der vorhergehenden Prozessierung und somit der Tools einbezieht. Dokumentierte, quelloffene Software ermöglicht schließlich, sie gezielt für spezifische Bedarfe weiterzuentwickeln oder ihre grundlegende Idee und Konzeption in Softwareprojekten mit aktuellen Bibliotheken erneut umzusetzen.

Aus unterschiedlichen Gesichtspunkten ist somit die kritische Betrachtung und Würdigung von Software, die in spezifischer Weise in der Forschung eingesetzt wird, sehr wünschenswert. Bislang wird der Bedarf, verstreut und über verschiedene Nutzergruppen verteilt, oft durch Erfahrungsberichte, Tutorials und zitierfähige Publikationen der Software etwa über Github oder Zenodo bedient. Insbesondere Erfahrungsberichte enthalten vielfach bereits Informationen, die Auswahlkriterien für die Nutzer*innen sind. Dies betrifft nicht alleine Anmerkungen zur Form und Verständlichkeit der Interaktion mit dem Programm oder die Import- und Exportfunktionen, sondern die Berichte zeigen zudem, wie die Software für die Bearbeitung einer geisteswissenschaftlichen Forschungsfrage eingesetzt wird. Allerdings finden Merkmale zur Beurteilung der Stabilität der Programme, zu ihrer Erweiterbarkeit oder auch ihrem Einsatz in anderen technischen Setups kaum Berücksichtigung. Tutorials, oft von den Entwickler*innen selbst verfasst, geben meist ebenfalls umfassend Einblick in die Funktionen, bewerten diese aber nicht. *Benchmark papers* und Softwarepublikationen richten sich schließlich nur an Entwickler*innen.

Nach der Bestimmung von Zielen und Blickwinkeln für Softwarerezensionen in der geisteswissenschaftlichen Forschung stellt sich die Frage nach den Kriterien, der Vorgehensweise und spezifischen Anforderungen. Die Zeitschrift Archäologische Informationen hat 2021 einen Vorschlag veröffentlicht (Homburg et al. 2021), der gemeinsam von Fachwissenschaftlerinnen und Informatiker*innen unterschiedlicher Schwerpunktsetzungen verfasst, und inzwischen kommentiert wurde (Carloni 2021; High-Steskal 2021). Der Text war eine wichtige Grundlage für die Ausformulierung der von den Herausgeber*innen der Zeitschrift CKIT formulierten Leitfragen, an denen sich die erwünschten Beiträge ausrichten sollen (CKIT 2021). Sie sind im Expert*innenforum der Task Area 3 "Research tools and data services" von NFDI4Culture 2022 diskutiert und angenommen worden, so dass sie heute in einer inhaltlichen und funktionalen Verbindung mit der entstehenden Software Registry stehen. Sie sind damit ein Anfang, um Gewohnheiten und *best practices* in der Forschungsgemeinschaft zu entwickeln. Entsprechend werden sie auch von den Workshopenbieter*innen als Orientierung verwendet.

Gemeinsame Ziele

- Erwerb der Kenntnis von wesentlichen Parametern zur Beurteilung einer Software im Forschungskontext
- Sammeln von Erfahrungen in der Zusammenstellung von spezifischen Nutzeranforderungen an eine Software

- Erwerb der Kenntnis von Vorgehensweisen, um Informationen zur handwerklichen Qualität von Software zu sammeln
- Sammeln von Erfahrungen im Einschätzen eigener Kompetenzen zur Beurteilung von Software
- Sammeln von Erfahrungen im (gemeinsamen) Verfassen einer Softwarerezension im Forschungskontext

Ergebnisformat

Ziel des Workshops ist es, auf Grundlage der genannten Handreichung (CKIT 2021) gemeinsam die Funktionalität und handwerklichen Qualität einer Software zu beschreiben sowie ihre Nutzbarkeit im Zuge der Bearbeitung einer geisteswissenschaftlichen Frage zu beurteilen.

Im Idealfall soll daraus eine gemeinsam verfasste Rezension entstehen, die bei der Zeitschrift CKIT oder den Archäologischen Informationen zur Veröffentlichung eingereicht wird.

Beispiele zur Rezension vorge-schlagener Forschungssoftware

1. SPARQLing Unicorn QGIS Plugin

Beschreibung: QGIS-Plugin (noch kein *stable release*, experimental), das eine einfache Integration von Geodaten aus Wikidata und anderen Linked Open Data SPARQL Endpoints ermöglicht. (Plugin: <https://plugins.qgis.org/plugins/sparqlunicorn/>; Github: <https://github.com/sparqlunicorn/sparqlunicornGoesGIS>)

2. PixPlot

Beschreibung: WebApp / eigenständige Software zum Clustern und anschließender Visualisierung von Bildern auf Grundlage von *neural network features*. (Projekt: <https://dhlabs.yale.edu/projects/pixplot/> ; Github: <https://github.com/YaleDHlab/pix-plot>)

3. Wax

Beschreibung: Eine Software-Lösung, um einfach digitale Ausstellungen mit IIIF-Technologien zu realisieren. Die Software verfolgt einen *minimal computing*-Ansatz. (Projekt: <https://minicomp.github.io/wax/> ; Github: <https://github.com/minicomp/wax>)

Workshoporganisation

Der Workshop wird in fünf Schritten durchgeführt. Dabei übernehmen die Workshopanbieter*innen die Funktion als Lotsen und geben zu jedem Schritt einen inhaltlichen Input. Anschließend wird gemeinsam von den Teilnehmer*innen die Untersuchung der Software auf die zuvor festgelegten Kriterien vorgenommen und die Ergebnisse von ihnen dokumentiert. Als Arbeitsumgebung wird ein GitHub-Repositorium genutzt. Der Fokus liegt auf der Evaluierung der Software in der Gruppe und dem Austausch über die Anwendbarkeit und Bedeutung

der verschiedenen Kriterien zur Beschreibung der ausgewählten Software. Der Workshop schließt mit einer gemeinsamen Reflexion zum Verlauf des Workshops ab.

Ablauf

Vorfeld

Schritt 1 (im Vorfeld der DHd)

- Im Vorfeld des Workshops installieren die Teilnehmer*innen die Software und benutzen sie in einem vorgegebenen *use case* mit einem Testdatensatz. Sie notieren dabei, unterstützt durch einige Leitfragen, für sich ihre Erfahrungen.
- In einer gemeinsamen Tabelle geben sie anonym Kennwerte zu ihrem technischen Setup an. Arbeitsumfang: etwa 120 min.

Workshop (0,5 Tage)

- Kennenlernen
- Erste Eindrücke zur Software sammeln

Schritt 2 - Forschungskontext der Software und grundlegende Funktionen

- Impuls 1 (5-10 min.): *Vorstellung des Anwendungsbeereichs der Software in der geisteswissenschaftlichen Forschung*
- Gemeinsame Auswahl und Anwendung von Kriterien, um die generelle Funktion der Software zu beschreiben und zu beurteilen.
- Beurteilung der Software nach diesen Kriterien (Gruppenarbeit)

Schritt 3 - Perspektive Anwender*innen

- Impuls 2 (5-10 min.): *Den Wald vor lauter Bäumen nicht sehen - Schwerpunkte in der Beschreibung aus Anwender*innensicht setzen*
- Gemeinsame Auswahl von Kriterien
- Einbeziehung subjektiver Aspekte (Kompetenzen, Forschungsinteresse, Gewohnheiten)
- Beurteilung der Software nach diesen Kriterien (Gruppenarbeit)

Workshop (0,5 Tage)

Schritt 4 - Perspektive Entwickler*innen

- Impuls 3 (5-10 min.): *Finden, was für Entwickler*innen interessant ist*
- Gemeinsame Auswahl und Anwendung von Kriterien, um die Software mit Blick auf ihre handwerkliche Qualität, Nachhaltigkeit und Entwicklungsfähigkeit zu beschreiben
- Zwischenbilanz : Einschätzung der Software

Schritt 5 - Bilanz und Ausblick

- Zusammenführung der ausgewerteten Kriterien und der Zwischenbilanzen

- Gemeinsame Verständigung, ob eine Fortführung als Publikationsprojekt möglich, sinnvoll und machbar ist
- Feedback zum Workshop

Vorkenntnisse und Kompetenzen

- Es sind keine spezifischen Vorkenntnisse erforderlich. Wesentlich ist ein allgemeines Interesse an Software und die Motivation zum offenen Arbeiten im Team.

Technisches Setup

- Die Teilnehmer*innen bringen ihren eigenen Rechner mit. Vor Ort wird ein Bildschirm benötigt.

Teilnehmer*innenzahl

Max. 15

Anbieter*innen Workshop

- Timo Homburg ist Wissenschaftlicher Mitarbeiter am i3mainz Institut für Raumbezogene Informations- und Messtechnik der Hochschule Mainz. Er erforscht Anwendungen im Bereich (Geospatial) Semantic Web, Computerlinguistik und den Digital Humanities. Seit 2019 ist er auch aktiver Wegbereiter für neue Geodatenstandards in der Standardisierungsgruppe zu GeoSPARQL, der OGC.
- Anne Klammt ist zum Zeitpunkt der Einreichung als Forschungsleiterin am Deutschen Forum für Kunstgeschichte Paris verantwortlich für die Digital Humanities. Seit 2020 publiziert sie zur Frage, wie Forschungssoftware und Datendienste wissenschaftlich rezensiert werden können.
- Fabian Offert ist Assistant Professor for the History and Theory of the Digital Humanities an der University of California in Santa Barbara. Er ist Mitherausgeber der Zeitschrift CKIT und forscht zur Nutzung von neuesten Praktiken des maschinellen Lernens in den digitalen Geisteswissenschaften, mit einem Schwerpunkt im Bereich digitale Kunstgeschichte.
- Florian Thiery ist Research Software Engineer im "Arbeitsbereich Wissenschaftliche IT, digitale Plattformen und Tools" des Römisch-Germanischen Zentralmuseums und forscht dort und entwickelt Forschungssoftware im Sonderforschungsgebiet "Explorative Forschung, Theorien- und Methodenentwicklung" im Handlungsfeld "Semantic Modelling and Knowledge Graphs". Seit 2020 ist er Mitglied des Vorstands der Gesellschaft für Forschungssoftware (de-RSE e.V.), Mitentwickler des SPARQLing Unicorn QGIS Plugin und forscht im Bereich semantischer Modellierung und Linked Open Data und Wikidata zu Irischen Ogham Steinen, was im Rahmen des Wikimedia Deutschland Fellow-Programms Freies Wissen gefördert wurde.

Fußnoten

1. Nach eigenen Erfahrungen der Workshopanbieter*innen und Rückmeldungen der Redaktion der Zeitschrift Archäologische Informationen, die seit 2019 eine eigene Rubrik für Softwarerezensionen bereitstellen.

Bibliographie

«Construction KIT: a review journal for research tools and data services in the humanities». Zugriffen 26. Juli 2022. <https://journals.ub.uni-heidelberg.de/index.php/ckit/index>.

High-Steskal, Nicole. 2021. «Contribution to the discussion: Handreichung zur Rezension von Forschungssoftware in der Archäologie und den Altertumswissenschaften: (Homburg, Klammt, Mara et al., 2020)». *Archäologische Informationen* 44: 257–59. <https://doi.org/10.11588/ai.2021.1.89200>.

Homburg, Timo, Anne Klammt, Hubert Mara, Clemens Schmid, Sophie C. Schmidt, Florian Thiery, und Martina Trognitz. 2021. «Diskussionsbeitrag: Handreichung zur Rezension von Forschungssoftware in der Archäologie und den Altertumswissenschaften». *Archäologische Informationen* 43: 357–72.

Katerbow, Mathias und Feulner, Georg. 2018. «Handreichung zum Umgang mit Forschungssoftware». Zenodo, 27.2.2018. <https://doi.org/10.5281/zenodo.1172970>.

Carlioni, Massimiliano. 2021. «Einige Anmerkungen zur ‚Handreichung zur Rezension von Forschungssoftware‘ : (Homburg, Klammt, Mara et al., 2020)». *Archäologische Informationen* 44: 253–56. <https://doi.org/10.11588/ai.2021.1.89199>.

Schmidt, Sophie. C. und Ben Marwick. 2020. «Tool-Driven Revolutions in Archaeological Science». *Journal of Computer Applications in Archaeology* 3: 18–32. <https://doi.org/10.5334/jcaa.29>.

GitMA oder CATMA für Fortgeschrittene

Schumacher, Mareike

schumacher@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Meister, Malte

meister@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Gerstorfer, Dominik

gerstorfer@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Dieser GitMA-Workshop richtet sich an fortgeschrittene CATMA User*innen mit Vorkenntnissen in digita-

ler Annotation, die im Rahmen der eigenen Arbeit oder von Forschungsprojekten mit größeren Mengen von Annotationsdaten operieren (wollen). Bei GitMA handelt es sich um ein Python-Package, mit dem Annotationsdaten, die in CATMA erstellt wurden, weiter verarbeitet werden können (Vauth et al. 2021). Wie greife ich über Git auf meine CATMA-Annotationsdaten zu? Wie visualisiere ich kollaborativ erstellte Annotationsdaten, die in mehreren Collections abgelegt sind? 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. 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 Workshops 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

Annotation ist eine zentrale Kultur- und Forschungspraxis, die bereits seit sehr langer Zeit analog betrieben wurde (vgl. Moulin 2010), bevor sie im Rahmen der Digital Humanities ins Digitale übertragen wurde. Textauszeichnung und -anreicherung, Freitextkommentare und das taxonomiebasierte Annotieren sind Formen der Annotation, die sich zum Teil überschneiden (vgl. Jacke 2018, § 9). Alle diese Formen werden von CATMA 6 digital unterstützt. Mithilfe selbst erstellter oder auf der Plattform forTEXT.net bereitgestellter Tagsets (z.B. Flüh 2020) kann einzeln oder im Team taxonomiebasiert annotiert werden.

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. Einzelne und Mehrfachannotationen, einander überlagernde oder überlappende Annotationen oder auch widersprüchliche Annotationen sind in CATMA durch die

Speicherung der Daten als Standoff-Markup 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. Die Einführung des auf Git basierenden Backends macht das Tool für die Digital-Humanities-Community aber noch interessanter. 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.

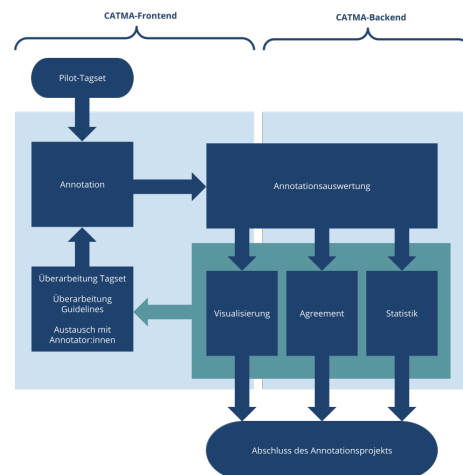


Abbildung 1: Im Workshop vermittelter Workflow zur Annotationsauswertung und -überarbeitung mit dem CATMA-Backend

Annotationen auswerten mit GitMA

Technische Niedrigschwelligkeit und Nähe zu traditionell-analogen Methoden der Geisteswissenschaften (vgl. Schumacher und Gius 2022) 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 (vgl. Piez 2010) auch die Einhaltung von Best Practices und Standards, die innerhalb der Digital-Humanities-Community

entwickelt wurden, von Bedeutung. Eine Verschmelzung von CATMA und dem direkten Datenzugriff über Git zu "GitMA" ermöglicht beides. Die im Annotationsprozess erstellten Daten können zum Beispiel nach der Übereinstimmung der Annotierenden untereinander (Artstein und Poesio 2008) 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 bzw. hervorgehoben werden können. So ergibt sich ein Workflow vom Frontend zum Backend und zurück, der auch die Erstellung von Goldannotationen (vgl. Wissler et al. 2014) unterstützt.

Format und Ablauf des Workshops

Der Workshop wird als halbtägiges hands-on Tutorial angeboten.

Ablauf:

CATMA 6 (45 Minuten)

- kurze Einführung in das CATMA-Frontend
- Struktur: Tagsets, Documents, Annotation Collections

Zugriff auf Annotationsdaten über Git (30 Minuten)

- wie clone ich ein CATMA Project?
- wie update ich ein CATMA Project, um neue Annotationen zu laden?
- Installation des Packages
- Laden eines Projects
- Zugriff auf Annotation Collections, Dokumente und Tagsets

15 Minuten Pause

Explorative Annotationsauswertungen (60 Minuten)

- Annotationsdaten visualisieren
- Netzwerkanalysen von Annotationsdaten

15 Minuten Pause

Statistische Annotationsauswertungen (45 Minuten)

- Einführung in die Begrifflichkeiten Inter-Annot-Agreement, Silver & Gold Standard
- Festlegung der Silver Annotations
- Umgang mit Annotationsspannen
- Automatische Erstellung eines Disagreement Tagsets
- Darstellung von Disagreement als Annotationen
- Manuelle Überarbeitung von automatischen Goldannotationen

Diskussion und Feedback (30 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 GitMA (sowie dessen Abhängigkeiten) können durch die Bereitstellung eines Docker-Image vermieden werden. Die Teilnehmer*innen sollten Docker Desktop auf einem eigenen Laptop installiert haben (Touch Devices werden nicht unterstützt) und diesen zum Workshop mitbringen. 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 forT-EXT-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 zur Verfügung, mit dem man während des Workshops arbeiten kann.

Benötigte Vorkenntnisse

Die Teilnehmer*innen sollten über grundlegende Kenntnisse der Kommandozeile, Git und Python sowie Jupyter verfügen.

Beitragende

Evelyn Gius, Prof. Dr.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Residenzschloss 1, 64283 Darmstadt

Evelyn Gius ist Professorin für Digitale Philologie und Neuere Deutsche Literatur an der Technischen Universität Darmstadt. Sie promovierte an der Universität Hamburg mit einer Arbeit über die narrative Struktur von Konflikterzählungen. Ihre Forschungsschwerpunkte sind manuelle Annotation, Operationalisierung, Erzähltheorie, Segmentierung und Konflikte. Sie ist PI mehrerer Digital-Humanities-Projekte (EvENT, KatKit, CATMA, forT-EXT) und ist Vorsitzende des Vereins Digital Humanities

im deutschsprachigen Raum (DHd), Mitherausgeberin des Journal of Computational Literary Studies (JCLS) und Mitherausgeberin der Buchreihe "Digitale Literaturwissenschaft".

Dominik Gerstorfer, M.A.

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Residenzschloss 1, 64283 Darmstadt

Dominik Gerstorfer promoviert über "Philosophische Fragen der Digital Humanities" an der Universität Stuttgart. Derzeit ist er im Projekt KatKit tätig, zuvor war er im DFG-Projekt forTEXT in Darmstadt und im Digital-Humanities-Projekt CRETA 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 Argumentationsanalyse. Im Rahmen von KatKit und 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, Residenzschloss 1, 64283 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, Residenzschloss 1, 64283 Darmstadt

Mareike Schumacher koordiniert das DFG-Projekt forTEXT (<https://fortext.net>), in dem neben der Dissemination von digitalen Routinen, Ressourcen und Tools in die traditionellen Fachwissenschaften auch die Weiterentwicklung von CATMA eine wesentliche Rolle spielt. Sie promovierte 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.

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>.

Flüh, Marie. 2020. „Emotionsanalyse“. In *forTEXT*. <https://fortext.net/ressourcen/tagsets/emotionsanalyse>.

Frey-Endres, Marcel, und Tobias Simon. 2021. *Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften*. Bd. 2. Digital Philology I Working Papers in Digital Philology. Darmstadt: TUprints. <https://doi.org/10.26083/tuprints-00017850>.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh und Jan Horstmann. 2021. „CATMA 6 (Version 6.3)“. *Zenodo*. DOI: 10.5281/zenodo.1470118. URL: <https://catma.de/> [letzter Zugriff 24. November 2021]

Jacke, Janina. 2018. „Manuelle Annotation“. In *forTEXT*. <https://fortext.net/routinen/methoden/manuelle-annotation>.

Moulin, Claudine. 2010. „Am Rande der Blätter. Gebrauchsspuren, Glossen und Annotationen in Handschriften und Büchern aus kulturhistorischer Perspektive“. *Autorenbibliotheken, Quarto. Zeitschrift des Schweizerischen Literaturarchivs* 30/31: 19–26.

Piez, Wendell. 2010. „Towards Hermeneutic Markup. An Architectural Outline“. In *Digital Humanities 2010. Conference Abstracts*, 202–5. London. <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-743.html>.

Rebholz-Schuhmann, Dietrich, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, u. a. 2010. „The CALBC Silver Standard Corpus for Biomedical Named Entities – A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers“. In *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/888_Paper.pdf.

Vauth, Michael, Hans Ole Hatzel und Evelyn Gius. 2021. „forTEXT/catma_gitlab:1.0.0.“ *Zenodo*. DOI: 10.5281/ZENODO.5669221.

Wissler, Lars, Mohammed Almashraee, Dagmar Monett, und Adrian Paschke. 2014. „The Gold Standard in Corpus Annotation“. In <https://doi.org/10.13140/2.1.4316.3523>.

Greening DH: individuelle Handlungsspielräume und institutionelle Perspektiven

Baillot, Anne

anne.baillot@univ-lemans.fr
Universität Le Mans

Feidicker, Charlotte

charlotte.feidicker@uni-bielefeld.de
Universität Bielefeld

Gerber, Anja

anja.gerber@bbaw.de
Berlin-Brandenburgische Akademie der
Wissenschaften

Roeder, Torsten

dh@torstenroeder.de
Universität Würzburg

Die AG “Greening DH” wurde 2021 ins Leben gerufen mit dem Ziel, das Bewusstsein der Verbandsmitglieder für ökologische Aspekte von Aktivitäten im Bereich der Digital Humanities (Forschung, Lehre, Projektmanagement, Softwareentwicklung etc.) zu schärfen. Neben konkreten Handlungsanalysen und -empfehlungen geht es der AG darum, die grundlegenden Veränderungen, die sich daraus für das Fach ergeben, epistemologisch zu begleiten. Der angedachte Workshop auf der DHd 2023 möchte beide Aspekte adressieren: Der erste Halbtage ist der Handlungsanalyse und Vermittlung von Best Practices auf individueller Ebene gewidmet, der zweite Halbtage geht institutionellen Handlungsspielräumen nach.

Herausforderungen der Klimakrise für die DH-Community

Während Forschende einerseits die aktuellen klimatologischen Entwicklungen als “Krise” begreifen und sich über den dringenden Handlungsbedarf einig sind (vgl. Leal Filho et al. 2021), hat zum anderen die Ausübung wissenschaftlicher Tätigkeit eine direkte und maßgeblich weitreichende Auswirkung auf die Umwelt. Die Diskrepanz zwischen individuellen Überzeugungen und systemischen Anforderungen in Bezug auf die Wichtigkeit ökologischer Perspektiven beeinflusst alle Arbeitsschritte eines Projektes und wird häufig als Konfliktsituation wahrgenommen. Insbesondere in den DH gilt es kritisch zu hinterfragen, ob fachspezifische Technologien, die arbeitsökonomisch nachhaltig sein sollen, auch ökologisch nachhaltig sind und ob jeder wissenschaftliche Erkenntnisgewinn den damit verbundenen Energieverbrauch aufwiegt. Dieser Fragenkomplex stand in den DH bislang nicht im Fokus fachspezifischer Reflexionen und wird in der laufenden Diskussion um Sustainability (FAIR, CARE) allenfalls marginal berücksichtigt.

In der Forschungsliteratur wird die Thematik der ökologisch nachhaltigen Digitalisierung bereits seit einigen Jahren behandelt (dazu ausführlich Lange/Santarius 2018). So haben sich im Bibliotheksbereich Akteurs-Netzwerke wie Libraries4Future und Netzwerk Grüne Bibliothek formiert (dazu Hauke et al. 2013). Ein weiterer Diskursstrang behandelt das Gebiet der ökologisch nachhaltigen Softwareentwicklung (z.B. Gröger/Köhn 2014; Eder/Gallagher 2017; Höfner/Frick 2018).

Daneben liegen Untersuchungen zum Energiebedarf von Datenzentren (exemplarisch: Schomaker et al. 2014) und Methoden zu umweltverträglicher Langzeitarchivierung (exemplarisch: Pendergrass et al. 2019) vor. Letzteres gilt es im Zusammenhang mit Minimal Computing und insbesondere Minimal Publishing (hierzu Holmes/Takeda 2019) zu betrachten, die in den DH steigende Beachtung erfahren. Für die fachübergreifende Thematik der umweltverträglichen Konferenzorganisation liegen spätestens seit dem “Zoom-Jahr” 2020 zahlreiche Untersuchungen vor (Stroud/Feeley 2015, Klöwer et al. 2020; Faber 2021; Glausiusz 2021). Ferner entstanden Frameworks zur Evaluierung der Nachhaltigkeit von Forschungseinrichtungen (Ferreboeuf et al. 2019; Mariette et al. 2021). Für den Bereich der DH sind aus dem Kontext der Digital Humanities Climate Coalition inzwischen erste Beiträge erschienen, die sich konkret mit dem ökologischen Impact der DH auseinandersetzen und konkrete Handlungsempfehlungen zusammenstellen (DHCC Information, Measurement and Practice Action Group und DHCC Toolkit Action Group 2022).

Auch bei Open Science kann man sich fragen, inwiefern das Streben nach universellem Zugang zu hochqualitativen Informationen (etwa Scans, Scripts) mit den ökologischen Verhältnissen auf Dauer kompatibel sind. Die Fragen der ökologischen Nachhaltigkeit in den DH muss auch im Kontext globaler sozial-ökonomisch ungleicher Ressourcenverteilung betrachtet werden, die insbesondere Fragen der Zugänglichkeit betreffen und neue Perspektiven auf die FAIR-Kriterien eröffnen: Sind Bereitstellungsangebote, die hohe Bandbreiten und leistungsstarke Servers und Clients erfordern, unter globalen und ökologischen Gesichtspunkten überhaupt als “fair” bewertbar? Welche Tugenden der Offenheit sind noch vertretbar, wenn die Bedingungen des Zugangs das Klima für einen wachsenden Teil der Weltbevölkerung unerträglich machen?

Ablauf

World Café: Was macht die individuelle Forschungspraxis?

Der erste Halbtage fokussiert auf die Forschungspraxis der Teilnehmenden hinsichtlich ihres ökologischen Fußabdrucks. Die Teilnehmenden werden eingeladen, ihre Forschungstätigkeit einer kritischen Analyse zu unterziehen. Sie reflektieren, an welchen Stellen in ihrem Arbeitsablauf Handlungsspielräume bestehen, die eine Reduktion des ökologischen Fußabdrucks ihrer Arbeit ermöglichen.

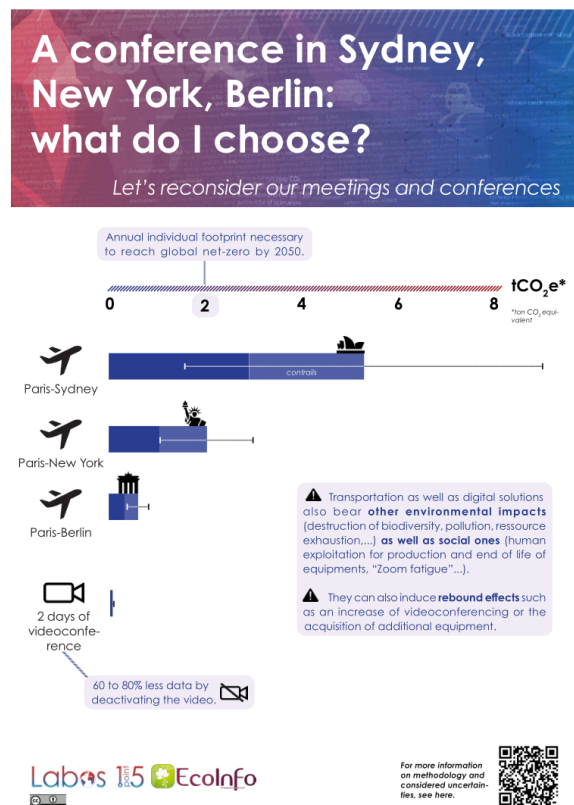
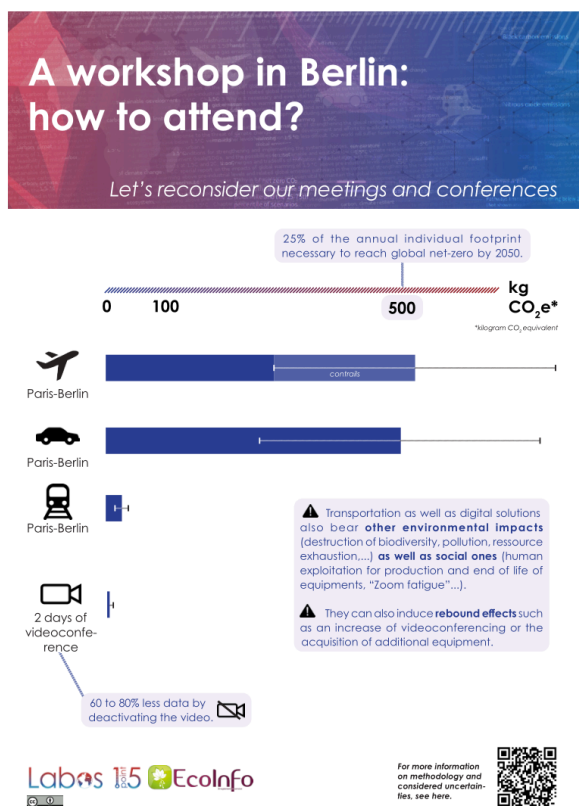
Nach diesem Einstieg werden drei einschlägige DH-Themen detaillierter betrachtet, zu denen bereits eine Handreichung der Digital Humanities Climate Coalition (s.o.) vorliegt, an welcher die Workshop-Veranstaltenden mitgearbeitet haben. Die Handreichung führt einschlägige Informationen zu ökologischen Aspekten digitaler Arbeitsprozesse zusammen.

Das erste Thema beschäftigt sich mit der Projektorganisation und fragt, welche Aspekte zur Reduktion des ökologischen Fußabdrucks im Projektmanagement berücksichtigt werden sollten.

Der zweite Themenblock nimmt die Nutzung von großer Rechenkapazität und Large Language Models in den Blick, auf denen Digital-Humanities-Forschung oft basiert.

Als drittes Thema vertiefen die Teilnehmenden die Potentiale des Minimal Computings. Dieser Themenblock thematisiert grundlegend die Verwendung von sparsamen technologischen Lösungen, die teils als konträrer Ansatz zu den erwähnten großen Rechenkapazitäten zu verstehen sind, insbesondere aber auf den Bereich der Bereitstellung abzielen.

Zu allen drei Themen werden Lösungsansätze auf Grundlage der DHCC-Handreichung durch die Teilnehmenden erarbeitet. Anschließend werden die anfangs identifizierten Handlungsspielräume unter Berücksichtigung der erarbeiteten Szenarien gemeinsam diskutiert.



Abbildungen: Labos 1point5, Infografiken zu CO₂-Belastung durch Reisen zu Workshops und Konferenzen, Verkehrsmittel und Entfernungen im Vergleich. Quelle: <https://labos1point5.org/les-infographies/poster-ecoinfo> – Lizenz: CC BY.

Panel und Rollenspiel: Institutionelle Herausforderungen

Der zweite Halbtage beginnt mit einer Podiumsdiskussion, die die DH-Community und angrenzende Felder zusammenführt. Diskutiert werden soll die Umsetzung einer grünen Agenda auf zwei Ebenen: zum einen institutionell auf Ebene der Infrastruktureinrichtungen, die für DH-Forschung unerlässlich sind, zum anderen durch Vertretende anderer geistes- und sozialwissenschaftlicher Fächer bzw. Bereiche, die in der ökologischen Diskussion weiter sind als die Digital Humanities. Die Diskussion der Panelistinnen untereinander und mit den Workshop-Teilnehmenden soll dazu beitragen, institutionelle Spielräume zu identifizieren, in denen ein Paradigmenwechsel in der Praxis angestoßen werden kann.

Es soll hier besonders der Frage nachgegangen werden, an welcher Stelle die DH-Community eigene Lenkungsmechanismen einsetzen könnte, um einen strukturellen Systemwandel in die Wege zu leiten: Was können die einschlägigen DH-Verbände (DHD, EADH, ADHO) und zukünftigen Konferenzorganisierende beitragen? Wie können Förderinstitutionen als Unterstützung erreicht werden? An welchen Stellen können Infrastrukturen ansetzen?

Eingeladene Panelistinnen:

- Dr. Magdalena Palica, Leiterin der Wissenschaftlichen Bibliothek der Stadt Trier
- Prof. Dr. Julia Affolderbach, Professorin für Nachhaltige Regional- und Standortentwicklung, Universität Trier
- Dr. Rabea Kleymann, ZfL Berlin, Vertreterin des DHd-Vorstands

Diese Diskussionsrunde möchte einen Impuls für eine ganzheitlich nachhaltige DH-Praxis geben und weitere Aktivitäten auf diesem Gebiet anstoßen. Die eingeladenen Expertinnen sind bewusst nicht ausschließlich aus den engeren Kreisen der DH-Community gewählt, sondern auch aus angrenzenden strategischen Feldern, um einen umfassenderen Austausch anzuregen sowie potentielle Synergien auszuloten. Außerdem ist vorgesehen, das Panel aufzunehmen und im Nachgang als Podcast auszustrahlen.

Im Anschluss an diese Podiumsdiskussion wird ein Rollenspiel den Teilnehmenden die Gelegenheit geben, eine Perspektive der in einem Projekt involvierten Akteurinnen und Akteure einzunehmen. Den Teilnehmenden wird eine Funktion zugeteilt (Personal aus Universitätsleitung, DH-Forschung, DH-Studium, IT, etc.) und sie sollen sich im Gespräch mit den anderen einer Herausforderung stellen. Für jede Rolle werden Profil und Ziele vorgegeben. Es sind drei Rollenspiel-Runden (jeweils ca. 20 min) geplant, wobei die Rollen jedes Mal unter den Teilnehmenden neu gemischt werden. So sind die zu lösenden Aufgaben jedes Mal neu und die Teilnehmenden werden dazu angeregt, sich die jeweiligen Argumente aus verschiedenen Perspektiven anzuschauen und sich in verschiedene Handlungsspielräume hineinzuversetzen. Die drei angedachten Szenarien sind:

- Top-Down: Das Direktorium verlangt den Entwurf einer Richtlinie, um den ökologischen Fußabdruck einer Abteilung / eines Instituts / einer Universität zu reduzieren, unter der Maßgabe des Qualitätserhalts von Forschung / Lehre / Infrastruktur.
- Bottom-Up: Mitarbeitende und Studierende engagieren sich für den Aufbau einer Nachhaltigkeitsgruppe, die sich innerhalb der Universität / Akademie für die nachhaltige Durchführung von Projekten und Bildungsveranstaltungen zu ökologischen Themen einsetzen soll und suchen Unterstützung in Verwaltung / Leitung.
- Worst Case (for now): Eine Hitzewelle von bis zu 45 Grad macht es unmöglich, den Betrieb von Forschung / Lehre / Rechenzentrum aufrecht zu erhalten: Welche Maßnahmen sollen getroffen und umgesetzt werden, um mit dieser Situation umzugehen?

Abschließend werden die gewonnenen Perspektiven, Fragen und Lösungsideen in einer gemeinsamen, moderierten Diskussion zusammengetragen. Individuelle und institutionelle Handlungsspielräume werden gegenübergestellt, passende Optionen für die Nachverwertung der Workshop-Resultate und zukünftige Perspektiven der "Green DH" erörtert.

Bibliographie

Bender, Emily M.; Geburu, Timnit; McMillan-Major, Angelina; Shmitchell, Shmargaret. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" FAccT 21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 10.1145/3442188.3445922.

DHCC Information, Measurement and Practice Action Group. 2022. "A Researcher Guide to Writing a Climate Justice Oriented Data Management Plan." Digital Humanities Climate Coalition 10.5281/zenodo.6451499.

DHCC Toolkit Action Group. 2022. Toolkit <https://sas-dhrh.github.io/dhcc-toolkit/> (zugegriffen: 15. Dezember 2022).

Eder, Kerstin; Gallagher, John P. 2017. "Energy-Aware Software Engineering." ICT – Energy Concepts for Energy Efficiency and Sustainability 10.5772/62522.

Faber, Grant. 2021. "A Framework to Estimate Emissions from Virtual Conferences." In International Journal of Environmental Studies 78: 608-623 10.1080/00207233.2020.1864190.

Ferreboeuf, Hugues; Efoui-Hess, Maxime; Kahraman, Zeynep. 2019. Lean ICT: Towards Digital Sobriety, The Shift Project https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report-The-Shift-Project_2019.pdf (zugegriffen: 15. Dezember 2022).

Glausiusz, Josie. 2021. "Rethinking travel in a post-pandemic world." In Nature 15.01. <https://www.nature.com/articles/d41586-020-03649-8> (zugegriffen: 14. Juli 2021).

Gröger, Jens; Köhn, Marina. 2014. Dokumentation des Fachgesprächs "Nachhaltige Software" am 28.11.2014, Umweltbundesamt <https://www.umweltbundesamt.de/publikationen/nachhaltige-software> (zugegriffen: 15. Dezember 2022).

Hauke, Petra; Latimer, Karen; Werner, Klaus Ulrich (eds.). 2013. The Green Library – Die grüne Bibliothek, De Gruyter / Saur.

Höfner, Anja / Frick, Vivian (eds.). 2018. Was Bits und Bäume verbindet. Digitalisierung nachhaltig gestalten, Oekom <https://www.oekom.de/buch/was-bits-und-baume-verbindet-9783962381493> (zugegriffen: 15. Dezember 2022).

Holmes, Martin; Takeda, Joseph. 2019. "The Prefabricated Website: Who Needs a Server Anyway?" TEI 2019. What is text, really? TEI and beyond gams.uni-graz.at/o:tei2019.116.

Klöwer, Milan; Hopkins, Debbie; Allen, Myles; Higham, James. 2020. "An analysis of ways to decarbonize conference travel after COVID-19." In Nature 15.07. <https://www.nature.com/articles/d41586-020-02057-2> (zugegriffen: 14. Juli 2021).

Lange, Steffen; Santarius, Tilman. 2018. Smarte Grüne Welt? Digitalisierung zwischen Überwachung, Konsum und Nachhaltigkeit, Oekom <https://www.oekom.de/buch/smart-gruene-welt-9783962380205> (zugegriffen: 15. Dezember 2022).

Leal Filho, Walter; Sima, Mihaela; Sharifi, Ayyoob et al. 2021. "Handling climate change education at universities: an overview." In Environmental Sciences Europe 33,109 10.1186/s12302-021-00552-5.

Mariette, Jérôme; Blanchard, Odile; Berné, Olivier; Ben-Ari, Tamara. 2021. "An open-source tool to assess the carbon footprint of research." In bioRxiv 16.01.10.1101/2021.01.14.426384.

Pendergrass, Keith L.; Sampson, Walker; Walsh, Tim; Alagna, Laura. 2019. "Toward Environmentally Sustainable Digital Preservation." In The American Archivist 82:165-206 10.17723/0360-9081-82.1.165.

Schomaker, Gunnar; Janacek, Stefan; Schlitt, Daniel. 2014. "The Energy Demand of Data Centers." ICT Innovations for Sustainability, 113-124 10.1007/978-3-319-09228-7_6.

Stroud, James T.; Feeley, Kenneth J. 2015. "Responsible academia: optimizing conference locations to minimize greenhouse gas emissions." In Ecography 38:402-404 10.1111/ecog.01366.

Hands-on-Workshop Datendokumentation

Lemaire, Marina

marina.lemaire@uni-trier.de

Universität Trier, Servicezentrum eSciences

Moeller, Katrin

katrin.moeller@geschichte.uni-halle.de

Martin-Luther-Universität Halle-Wittenberg,
Historisches Datenzentrum Sachsen-Anhalt

Schulz, Julian

Schulz@MaxWeberStiftung.de

Max Weber Stiftung, Geschäftsstelle, Digital Humanities
und Forschungsdatenmanagement

Söring, Sibylle

sibylle.soering@fu-berlin.de

Freie Universität Berlin, Universitätsbibliothek, Leitung
Forschungsdatenmanagement

Wettlaufer, Jörg

jwettla@gwdg.de

Akademie der Wissenschaften zu Göttingen,
Koordination Digitalisierung und Datenkuration

Einführung

Format: Workshop, ganztags

Gruppengröße: max. 30 Teilnehmende

Techn. Ausstattung: Beamer, bevorzugt digitales Whiteboard oder Pinnwände & Medienkoffer, evtl. einen weiteren Raum für Gruppenarbeit, ausreichend Steckdosen für Laptops

Bei den Einreichenden handelt es sich um Vertreter*innen von Datenzentren und universitären Infrastruk-

tureinrichtungen, die Mitglied in der DHd AG Datenzentren sind. Ihre Aufgabe ist es u. a. Forschende bei der Entwicklung und Umsetzung des Forschungsdatenmanagements (FDM) in den Geistes-, Sozial- und Kulturwissenschaften zu unterstützen sowie Forschungsinfrastrukturen und Daten für diese Disziplinen bereitzustellen. Dabei fallen häufig Beratungs- und Kompetenzvermittlungsaufgaben an, die tief in die Forschungsprozesse der Wissenschaftler*innen hineinreichen und Fragen nach Art und Umfang der Dokumentation der Forschungsdaten aufwerfen. Während die Einreichenden im Rahmen des Workshops ihre disziplinäre und infrastrukturelle Expertise und Erfahrung aus der Projektbegleitung und -durchführung einbringen, werden Forschende der Geistes- und Kulturwissenschaften aus ihren Erfahrungen bei der Erstellung und / oder Nachnutzung von Forschungsdaten berichten und im Datathon Datensätze bereitstellen, die sie selbst erstellt oder in eigenen Projekten nachgenutzt haben. Hieraus sollen perspektivisch Anforderungen auch an die Infrastrukturangebote der Einreichenden abgeleitet werden.

Workshopkonzept

Die Veröffentlichung von Forschungsdaten, d. h. von Daten, die im Rahmen der Planung, Durchführung und Dokumentation wissenschaftlicher Projekte entstehen, erlebt eine Konjunktur. Diese liegt einerseits in wachsenden Anforderungen seitens der Fördermittelgebenden begründet, die in zunehmendem Maße von den durch sie finanzierten Forschungsvorhaben eine Bereitstellung der Datenbasis als Fundament wissenschaftlicher Arbeit erwarten. (Vgl. DFG 2015; ERC 2017) Andererseits kann die steigende Zahl durch eine sich erweiternde Methodologie in den Geisteswissenschaften, d. h. in einer Hinwendung zu daten- und rechnergestützten Forschungsmethoden erklärt werden: Digitale Editionen, Text Mining oder Bildähnlichkeitsanalysen stellen heute zwar eher noch Ausnahmefälle dar, rücken aber zunehmend in das Methodenrepertoire geisteswissenschaftlicher Forschung vor. Somit steigt die Zahl an Datensätzen, die im Rahmen derart gelagerter Projekte entstehen und für eine begleitende Publikation in Frage kommen. In der Folge ergibt sich immer häufiger die Möglichkeit, eben diese Daten als Grundlage für neue Forschungsprojekte zu verwenden.

Vor dem Hintergrund einer zunehmenden Zahl an potenziell verwendbaren Forschungsdaten mag es verwundern, dass bislang eher selten auch eine Nachnutzung dieser Daten in neu gelagerten Forschungskontexten erfolgt. Es entsteht der Eindruck, dass „Success Stories“ im Bereich der Nachnutzung insbesondere bei geisteswissenschaftlichen Forschungsdaten ein Desiderat darstellen. Ein Grund hierfür mag darin liegen, dass eine strukturierte und detaillierte Form der Datendokumentation bislang wenig im Fokus stand. (Vgl. Daudrich 2018, 13) Entsprechend fehlen fach- bzw. methodenspezifische Best Practice-Modelle, wie sie zunehmend im Kontext generischer Ansätze formuliert werden. (Vgl. dazu z. B. CESSDA 2020, Kap. 2. Organise & Document; Dierkes 2021) Eine strukturierte, standardisierte Dokumentation ist jedoch zwingend erforderlich, um Daten in neuen Projekten nachnutzen zu können. Insbesondere die Grund-

lagen der Datenerhebung (Auswahl, Begrenzungen, Ursprung, Datenqualität, Prozessierungen usw.) müssen nachvollziehbar sein, um eine spätere Verwendung überhaupt erst zu ermöglichen. Dabei wird deutlich, dass selbst hinsichtlich der Ziele, der Definition und der Grundelemente einer Datendokumentation keine einheitliche Auffassung besteht.

Ein Kernelement im Bereich der Dokumentation stellt die Beantwortung und sukzessive Anpassung eines Datenmanagementplans (DMP) dar. Verstanden als „Living Document“, kann ein DMP dazu beitragen, die in einem Projekt verwendeten Daten, Software und Methoden detailliert darzustellen und die aus dem Projekt resultierenden Forschungsdaten damit zu kontextualisieren. Während die Erstellung eines DMP inzwischen vermehrt seitens der Fördermittelgebenden als obligatorisch betrachtet wird, besteht weitgehend noch keine Pflicht, diesen zusammen mit den Forschungsdaten zu veröffentlichen. Im Sinne der Nachvollziehbarkeit aller im Projekt unternommenen Schritte wäre die Veröffentlichung des DMP als Beitrag zur Dokumentation jedoch anzuraten.

Einen weiteren Baustein hinsichtlich der Dokumentation von Forschungsdaten stellt ihre umfassende Beschreibung mit Metadaten dar. Der Grad der Nutzbarkeit ist dabei in hohem Maße davon abhängig, welches Metadatenchema verwendet wird und in welcher Detailtiefe es befüllt wird – gerade abseits der (häufig) geringen Zahl an Pflichtfeldern. Aber auch ein vergleichsweise umfangreich angelegtes Metadatenchema wie das weltweit und disziplinübergreifend verbreitete DataCite (Vgl. Brase u. a. 2015) bietet nur begrenzte Möglichkeiten, Angaben zu verwendeter Software, Modellen und Methoden im Feld „Description“ in Freitextform und damit nicht strukturiert zu tätigen. (Vgl. DataCite 2021) Ein tiefergehendes Verständnis des Datensatzes und die Nachvollziehbarkeit seines Entstehungsprozesses wird damit zwar angedeutet, jedoch nicht in Gänze ermöglicht. Es offenbart sich in diesem Kontext eine Kluft zwischen den Anforderungen von Datenzentren auf der einen (Metadatenqualität) und Forschenden, die die Daten nachnutzen möchten, auf der anderen Seite (ausführliche Dokumentation des Entstehungsprozesses).

Neben Datenmanagementplänen und beschreibenden Metadaten bedarf es für die Nachnutzung von Forschungsdaten jedoch weiterer Hilfsmittel. Hier lohnt ein Blick in andere Fachbereiche, in denen die komplementäre Bereitstellung von Materialien wie Codebüchern, elektronischen Laborbüchern oder Data-Curation-Profiles bereits gängige Praxis ist. (Vgl. Heuer u. a. 2020; Hermann u. a. 2018; Jensen 2012)

Im Rahmen des Workshops soll dieses Desiderat rund um das Thema Datendokumentation aufgegriffen und mit den Teilnehmenden diskutiert werden. In einem ersten Schritt wird das Ziel verfolgt, eine Arbeitsdefinition herzustellen, um eine gemeinsame Vorstellung davon zu erhalten, was unter „Datendokumentation“ zu verstehen ist, welche Komponenten (z. B. DMP, Metadaten, Codebook) zwingend erforderlich sind und welche dagegen eher optionalen Charakter besitzen. Darauf aufbauend soll praxisnah ergründet werden, welche Formen der Dokumentation benötigt werden, um nicht nur die Auffindbarkeit von Forschungsdaten, sondern auch ihre Nachnutzung zu vereinfachen bzw. überhaupt zu ermög-

lichen. In diesem Kontext wird auch zu diskutieren sein, wer – d. h. Forschende oder Kuratierende – für die Dokumentation der Daten verantwortlich zeichnet. Schließlich wird als weiteres Ziel des Workshops vorgegeben, ein besseres Verständnis davon zu erlangen, welche Informationen zwingend Teil einer Datendokumentation sein sollten (z. B. Kontext der Erhebung, Erhebungsmethode, Struktur der Daten und deren Beziehung zueinander). Der Workshop bezieht die Perspektive der Infrastruktureinrichtungen ein (z. B. Repositoriumsbetreibende, Datenzentren) und kann dazu dienen, einen Überblick zu bereits bestehenden Formen der Datendokumentation zu erhalten.

Die DHd-AG Datenzentren hat sich in bisher zwei verschiedenen Veranstaltungen¹ mit der Dokumentation von Forschungsdaten beschäftigt. Dabei wurden vor allem die Herausforderungen der datenhaltenden Institutionen diskutiert, die besonders in der Standardisierung und effizienten Ausgestaltung von Workflows zur Dokumentation von Forschungsdaten liegen. Ziel des hier eingereichten Workshops ist dagegen die Perspektive der Nutzer*innen selbst. Gezielt soll für unterschiedliche, aber typische geisteswissenschaftliche Daten die Dokumentation von Forschungsdaten hinsichtlich ihres Informationswertes, der Verständlichkeit, der Vollständigkeit und ihres tatsächlichen Gebrauchswertes zur Nachnutzbarkeit geprüft werden.

Längerfristiges Ziel ist die Entwicklung von Standards und Guidelines, die Nutzer*innen und Forschungsdatenzentren in die Lage versetzen, aussagekräftige Dokumentationen von Forschungsdaten zu erstellen. Im Mittelpunkt des Workshops stehen daher die Analyse von Use Cases zur Dokumentation von Forschungsdaten, die aus der Ersteller- wie aus der Nachnutzungsperspektive diskutiert werden sollen, und die Erarbeitung eines Dokumentationsschemas für die einzelnen Datentypen.

Workshop-Programm

Der eintägige Workshop der DHd-AG Datenzentren gliedert sich in zwei Teile. Am Vormittag werden nach einem einführenden Vortrag durch die Organisator*innen zu den Zielen und zentralen Fragen des Workshops vier Praxisbeispiele präsentiert, die erläutern, welche Daten sie mit welchen Zielen und Methoden nachgenutzt bzw. weiterverarbeitet haben und welche Probleme sich ihnen aufgrund mangelnder oder gar fehlender Dokumentation gestellt haben. Nach jeder Präsentation soll in einer Diskussion gemeinsam mit dem/der Referent*in auf einem digitalen Whiteboard zusammengetragen werden, welche Aspekte in der Datenver- und -aufbereitung in diesem konkreten Fall hätten dokumentiert werden sollen, um die geschilderten Probleme zu vermeiden. Bewusst wurde eine spezifische Vielfalt an Fallbeispielen ausgewählt, um eine hinreichende Breite für geisteswissenschaftliche Dokumentationstypen zu analysieren. Dabei sind für jeden Vortrag 20 Minuten Referat und 20 Minuten Diskussion vorgesehen.

– Für den Bereich der quantitativen Daten wird Paul Beckus (Historiker an der Martin-Luther-Universität Halle/Wittenberg) aus der datenerstellenden Perspektive berichten, welche Fragen und Probleme sich ihm bei der Dokumentation von Datensätzen ergaben und welche

Unterstützungsangebote sich ein historisch arbeitender Wissenschaftler in diesem Prozess erhofft. (Vgl. Beckus 2021)

– Aline Deicke (Professorin für Digital Humanities, Philipps-Universität Marburg / Digitale Akademie, Akademie der Wissenschaften und der Literatur Mainz) wird für den Bereich Netzwerkanalyse aus ihrer Arbeit zur Analyse der "Streitkultur" in innerprotestantischen Auseinandersetzungen anhand polemischer Flugschriften (Vgl. Deicke 2017) berichten.

– Für die Bildwissenschaften wird Stefanie Schneider (Wissenschaftliche Assistentin für Digitale Kunstgeschichte an der Ludwig-Maximilians-Universität München) am Beispiel von ARTigo – Das Kunstgeschichtsspiel (<https://www.artigo.org>) nicht nur aus einer Außen-, sondern ebenso aus einer Innenperspektive heraus berichten und skizzieren, wie sich Datenbereitstellung und -dokumentation im Laufe der Versionen verändert haben.

– Abschließend wird Yvonne Rommelfanger (Datenkuratorin am Servicezentrum eSciences der Universität Trier) für den Bereich der qualitativen Daten am Beispiel der (Re-)Retrodigitalisierung der Edition der Kabinettsprotokolle des Landes Nordrhein-Westfalen (<http://protokolle.archive.nrw.de/>), von der Datenaufbereitung für die online-Publikation berichten.

Nach den Berichten sollen in einer halbstündigen Gruppenarbeit alle gesammelten Aspekte zur Datendokumentation gesichtet und versucht werden, eine erste Kategorisierung auf einem gemeinsam zu bearbeitenden Whiteboard vorzunehmen. Die Ergebnisse werden im Plenum diskutiert und zusammengeführt. Hierzu wird ein Schema² verwendet, das die Sicht der Datenzentren repräsentiert. Beides dient als Grundlage für die nachfolgenden Gruppenarbeiten am Nachmittag während des Datathons. Beim Datathon stellt jeweils eine Person einen Datensatz vor und macht einen Vorschlag für ein Nachnutzungsszenario, anhand dessen die Gruppen gemeinsam versuchen, den Datensatz zu verstehen und das Dokumentationsschema weiterzuentwickeln. In der Gruppenarbeit sollen sie einerseits feststellen, welche Informationen fehlen und hierfür Anforderungen formulieren, und andererseits überlegen, was sie dokumentieren müssten, damit ihre Ergebnisse wiederum verstehbar und nachnutzbar werden. Auf diese Weise sollen sie die Kategorien und Aspekte der Datendokumentation auf dem Whiteboard überarbeiten und weiterentwickeln. Für die Bereitstellung eines Datensatzes haben sich bislang folgende Personen bereit erklärt:

– Tinghui Duan vom DFG-Graduiertenkolleg "Modell Romantik" an der Universität Jena – deutsche literarische Prosatexte des langen 19. Jahrhunderts (<https://github.com/t-duan/dissertation/tree/main/data>)

– Svenja Guhr vom Institut für Sprach- und Literaturwissenschaft der Technischen Universität Darmstadt (fortext lab) – deutsche literarische Prosatexte d-Prose 1870 1920 (Vgl. Gius u. a 2021)

– Mareike König vom Deutschen Historischen Institut Paris – Adressbuch Deutscher in Paris von 1854 (<https://perspectivia.net/publikationen/quellen/adressbuch>) & Inventar der Korrespondenz der Constance de Salm (1767 1845) (<https://constance-de-salm.de/>)

– Katrin Moeller vom Historischen Datenzentrum Sachsen-Anhalt an der Martin-Luther-Universität Halle/

Wittenberg – Interviewdaten der BOLSA-Längsschnittstudie (<https://bolsa.uni-halle.de/suche/>)

– Julia Röttgermann vom Trier Center for Digital Humanities an der Universität Trier – Französische Romane des 18. Jahrhunderts (<https://github.com/MiMoText/roman18/tree/master/XML-TEI>)

Neben den hier genannten können auch andere Teilnehmende des Workshops Datensets für den Datathon mitbringen.

Nach Abschluss der Gruppenarbeiten werden die Ergebnisse im Plenum präsentiert. Zum Abschluss wird die Workshop-Gruppe diskutieren, welche weiteren Schritte notwendig sind, um erste Empfehlungen für die Dokumentation von geisteswissenschaftlichen Forschungsdaten zu erarbeiten.

Programmablauf

Uhrzeit	Programmpunkt
09:00	Begrüßung und Einführungsvortrag
09:20	Use Case 1: Quantitative Daten (20+20 Min.)
10:00	Use Case 2: Netzwerkanalyse (20+20 Min.)
10:40	Kaffeepause
11:10	Use Case 3: Bildannotationsdaten (20+20 Min.)
11:50	Use Case 4: Qualitative Daten/Re-Retrodigitalisierung (20+20 Min.)
12:30	Mittagspause
13:30	Kategorisierung der Aspekte der Datendokumentation auf der Basis der Berichte
14:00	Vorstellung, Diskussion und Zusammenführung der Gruppenergebnisse
14:30	Datathon
16:00	Kaffeepause
16:15	Vorstellung, Diskussion und Zusammenführung der Gruppenergebnisse
16:45	Next Steps
17:00	Ende

Fußnoten

1. Hands on Research Data, Workshopreihe der AG Datenzentren auf der vDHd2021, <https://vdhd2021.hypotheses.org/178>, Zugriffen 9. Dezember 2022; Dokumentation von Forschungsdaten – Erfahrungen und Aufgaben aus der Praxis, Workshop der AG Datenzentren auf der FORGE21, <https://forge2021.uni-koeln.de/programm/workshop-ag-datenzentren.html>, Zugriffen 9. Dezember 2022.

2. Das Schema wurde für den oben genannten Workshop auf der FORGE 2021 entwickelt und erprobt. https://docs.google.com/spreadsheets/d/19IBDW-w_iBnWU8oy8R92UWKCFuKOZBaONITGQUZRjLw, Zugriffen 9. Dezember 2022.

Bibliographie

ADW Mainz, Akademie der Wissenschaften und der Literatur, Mainz. o. J. "C&C digital. Datenbank zur Bekenntnisbildung und Konfessionalisierung (1548-1580)". <http://www.controversia-et-confessio.de/cc-digital.html> (zugegriffen: 9. Dezember 2022).

Beckus, Paul. 2021. Der Fürst im Kabinett: Supplikations- und Herrschaftspraxis unter Franz von An-

halt-Dessau (1758-1817). Quellen und Forschungen zur Geschichte Sachsen-Anhalts 24. Halle: Mitteldeutscher Verlag.

Brase, Jan, Michael Lautenschlager, und Irina Sens. 2015. "The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite". In: *D-Lib Magazine* 21 (1/2) 10.1045/january2015-brase.

CESSDA, Training Team. 2020. CESSDA Data Management Expert Guide. Bergen. 10.5281/ZENODO.3820473.

Data Cite. 2021. "DataCite Metadata Schema v4.4 Properties Overview". <https://support.datacite.org/docs/datacite-metadata-schema-v44-properties-overview> (zugegriffen: 9. Dezember 2022).

Daudrich, Anna. 2018. "Umgang mit Forschungsdaten in den Geistes- und Sozialwissenschaften. Bericht zur Bedarfserhebung an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)". <https://www.fdm-bayern.org/files/2018/11/forschungsdatenmanagement-in-den-geisteswissenschaften-an-der-fau-umfrage.pdf> (zugegriffen: 9. Dezember 2022).

Deicke, Aline. 2017. "Networks of Conflict: Analyzing the 'Culture of Controversy' of Polemical Pamphlets of Intra-Protestant Disputes (1548-1580)". In *Journal of Historical Network Research* 1 (Oktober): 71-105. <http://jhnr.uni.lu/index.php/jhnr/article/view/8> (zugegriffen: 9. Dezember 2022).

DFG, Deutsche Forschungsgemeinschaft. 2015. "Leitlinien zum Umgang mit Forschungsdaten". http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf (zugegriffen: 9. Dezember 2022).

Dierkes, Jens. 2021. "4.1 Planung, Beschreibung und Dokumentation von Forschungsdaten". In *Praxishandbuch Forschungsdatenmanagement*, herausgegeben von Markus Putnigs, Heike Neuroth, und Janna Neumann, 1st Aufl., 303-26. Boston: De Gruyter Saur 10.1515/9783110657807.

ERC, European Research Council. 2017. "Guidelines on Implementation of Open Access to Scientific Publications and Research Data". https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf (zugegriffen: 9. Dezember 2022).

Gius, Evelyn, Svenja Guhr, und Benedikt Adelman. 2021. "d-Prose 1870-1920". 10.5281/ZENODO.4315208.

Hermann, Sybille, Uli Hahn, Markus Gärtner, und Florian Fritze. 2018. "Nachträglich ist nicht gleich nachnutzbar: Ansätze für integrierte Prozessdokumentation im Forschungsalltag". In: *o-bib. Das offene Bibliotheksjournal* 5 (3): 32-45 10.5282/o-bib/2018H3S32-45.

Heuer, Jan-Ocko, Susanne Kretzer, Kati Mozygemba, Elisabeth Huber, und Betina Hollstein. 2020. "Kontextualisierung qualitativer Forschungsdaten für die Nachnutzung: eine Handreichung für Forschende zur Erstellung eines Studienreports". Herausgegeben von Forschungsdatenzentrum Qualiservice. Qualiservice Working Papers. Bremen: Universität Bremen. 10.26092/elib/166.

Jensen, Uwe. 2012. "Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten". Herausgegeben von Leibniz Institut für Sozialwissenschaften GESIS. Technical Reports. Köln. [\[berichte/2012/TechnicalReport_2012-07.pdf\]\(berichte/2012/TechnicalReport_2012-07.pdf\) \(zugegriffen: 9. Dezember 2022\).](http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methoden-</p>
</div>
<div data-bbox=)

Hands-on-Workshop Wissenschaftsbloggen mit de.hypotheses Halbtägiger Workshop

König, Mareike

mkoenig@dhi-paris.fr

Deutsches Historisches Institut Paris, Frankreich

Die leichte Zugänglichkeit von Publikationsmedien wie Blogs ermöglicht es Forschenden seit der Jahrtausendwende, selbst zu entscheiden, wann, wo und was sie veröffentlichen. Diese selbstbestimmte Aneignung eines wissenschaftlichen Publikationsraums war ein durchaus spektakulärer Schritt (König 2015, 58). Durch die zunehmende Verbreitung von Wissenschaftsblogs und Angebote von qualitätsgestützten Plattformen speziell für die Geisteswissenschaften wie *hypotheses*¹, die persistente URLs genauso bieten wie eine Archivierung der Inhalte und einen Beratungsservice für Bloggende, hat diese Form der Kommunikations- und Publikationspraktik in der Wissenschaft zwischenzeitlich viel von ihrem revolutionären Charakter verloren. Vielmehr ist sie selbst in den Geisteswissenschaften in der Mitte der Disziplinen angekommen, wie beispielhaft der stetig wachsende Katalog der Blogplattform *hypotheses* zeigt, der gegenwärtig 4326 Wissenschaftsblogs umfasst². Wissenschaftsblogs – so aktuelle Resümeees – sind aus der wissenschaftlichen Publikations- und Kommunikationslandschaft nicht mehr wegzudenken (Baillot 2022, 105; Wuttke und Gebert 2021, 428-431; Gebert und van Beek 2019, 274-276).

Damit erübrigt sich aber das Thema Wissenschaftsblogs oder ein Workshop dazu keineswegs, im Gegenteil: Vielmehr hat Wissenschaftskommunikation seit einigen Jahren Konjunktur angesichts einer spürbaren Polarisierung der Gesellschaft, einem Glaubwürdigkeitsverlust der Wissenschaft eingeheizt in der Pandemiekrise und einem Bedürfnis von Forschenden, die Transformation der verlagszentrierten wissenschaftlichen Publikationslandschaft mitzugestalten. Aktiv Wissenschaftskommunikation zu betreiben, über Projekte früh im Forschungsprozess zu kommunizieren, sich in hochschulpolitische sowie gesellschaftliche Debatten einzumischen und Open Access zu publizieren, sind die Gebote der Stunde, wie sich auch in Leitlinien und Debatten der Wissenschaftsinstitutionen zeigt³. Die Nachfrage bei der deutschsprachigen Blogplattform für die Geisteswissenschaften *de.hypotheses* nach einführenden Workshops hat zudem seit Beginn der Pandemie stark zugenommen. Dies gibt zusammen mit der wahrnehmbaren Professionalisierung Anlass genug, im geplanten halbtägigen Workshop inhaltliche, technische und gestalterische Unterstützung rund um das Thema Wissenschafts-

blogs speziell in den Geisteswissenschaften zu geben und aufzuzeigen, wie sich das Wissenschaftsbloggen als Teil des eigenen Publikations- und Lehrportfolios und als Vernetzungsbaustein kompetent und strategisch verwenden lässt.

Wissenschaftsblogs: Einblicke in die laufende Forschung

Wissenschaftliche Blogs sind ein Ort für die Veröffentlichung und Diskussion laufender Forschungsarbeiten, ein Kanal für die Selbstdarstellung und Vernetzung von Forschenden. Sie zeigen Wissenschaft im Entstehen und sorgen für Reichweite und Sichtbarkeit der eigenen Forschung. Blogs wirken im Vergleich zu statischen Websites sehr viel dynamischer, zumal es über die Kommentarfunktion einfach ist, Diskussionen anzustoßen (Wuttke und Gebert 2021, 429). Der wissenschaftliche Austausch über Blogbeiträge, Kommentare und Links ist interaktiv, schnell und direkt, auch wenn sich Bloggende gerade in den Geisteswissenschaften häufig eine höhere Anzahl an Kommentaren wünschen (vgl. die Umfrage bei de.hypotheses, König 2019, 12-14). Studien über die Nutzung von sozialen Medien in den Wissenschaften weisen ein breites Spektrum unterschiedlicher Verwendungszwecke, Ziele und Motive wissenschaftlicher Bloggenden auf (Mahrt und Puschmann 2014), die von der Plattform, dem akademischen Rang und dem Status der Forschenden, Alter und Geschlecht ebenso abhängig sind wie von der jeweiligen Disziplin und vom Herkunftsort der oder des Bloggenden (Sugimoto et al. 2017, 2039, 2046). Die meisten Plattformen erlauben eine sehr breite Palette an Nutzungsszenarien, was sich insbesondere beim Wissenschaftsbloggen zeigt: Es gibt keine Vorgaben, Richtlinien oder Beschränkungen im Hinblick auf Umfang und Länge der Beiträge, auf die Häufigkeit der Publikation, auf den verwendeten Stil. Neben Text können ohne Kosten genauso Abbildungen, Podcasts, Videos, animierte Karten etc. eingebunden werden. Wissenschaftsbloggen erleben viele Forschende daher als Befreiung, auch wenn die völlige Offenheit im Blog ein Phantasma sein mag (König 2015, 66) und nicht zuletzt die Wahl der Sprache Auswirkungen auf die Ausdrucksmöglichkeiten hat (Baillot 2022). Zugleich mag diese Freiheit gerade beginnende Bloggerinnen und Blogger einschüchtern und in der kreativen Nutzung ihres Blogs begrenzen. Durch das Aufzeigen von Beispielen sollen im Workshop genau dieses Spannungsverhältnis adressiert und Lösungsmöglichkeiten aufgezeigt werden.

Gleichzeitig haben sich Blogs als ein Ort erwiesen, an dem in traditionellen Zeitschriften veröffentlichte Artikel kritisiert und korrigiert werden (Sugimoto et al. 2017, 2045-2046). Neuere Projekte zeigen außerdem Wege auf, um Wissenschaftsblogs an Fachzeitschriften heranzuführen, indem etwa wie beim Mittelalterblog⁴ eine Redaktion Auswahl und Lektorat von Beiträgen übernimmt, diese in Repositories hinterlegt und in Kooperation mit Bibliotheken für die Katalogisierung sorgt (Döring & Gebert 2022, 93-97). Andere Nutzungen zielen in Richtung „kleine Editionen“ für das Edieren von Texten im Blog und mit Wikisource (Bemme 2022).

Wissenschaftliches Bloggen gehört in praktischer wie in theoretischer Hinsicht zum Gegenstandsbereich der DH: Es ist ein zentrales Mittel der Community Building, hat Berührungspunkte zum weiten Feld des digitalen Publizierens und ist ein Baustein im Medien-Portfolio von Open Science (Wuttke und Gebert 2021). In der Lehre finden Blogs Einsatz als Schreib- und Übungsblog für Studierende wie als Gegenstand, um vielfältige Themenbereiche wie Publikation, Open Science, ethische und rechtliche Fragen etc. zu diskutieren (Tantner 2015).

Didaktischer Zugang und Aufbau des Workshops

Im Workshop sollen grundlegende inhaltliche und technische Kenntnisse des Bloggens mit Wordpress vermittelt und Ansätze für das Entwickeln einer Blogstrategie diskutiert werden. Workshopteilnehmende sollen hinterher in der Lage sein, ein Blog einzurichten und zu führen (ob bei hypotheses oder anderswo), Inhalte dafür zu produzieren, sichtbar zu machen und zu lizenzieren sowie verschiedene Blogpraktiken zu kennen.

Der halbtägige Workshop (4h) ist didaktisch als Hands-On-Workshop aufgebaut und gliedert sich in drei Teile: In einem ersten Teil (1,5h) werden grundlegende Fragen zum Wissenschaftsbloggen thematisiert. Die Vermittlung von Inhalten erfolgt nach einem kurzen Input in Form eines Gesprächs dreier erfahrener Wissenschaftsbloggerinnen (Anne Baillot, Mareike König, Ulrike Stockhausen) über Best Practice-Beispiele entlang der Fragen: Wie sieht ein guter wissenschaftlicher Blogbeitrag aus? Wie funktioniert die Interaktion mit den Leser:innen? Welche Sprache, welcher Stil ist bei wissenschaftlichen Blogs angemessen? Wann ist ein Blog erfolgreich? Wie geht man mit unsachlichen Kommentaren oder gar mit Shitstorms um? Wie lassen sich Blogs in der Lehre fruchtbar einsetzen?

Der zweite und dritte Teil (je 1h) sind als praktische Übungseinheiten mit inhaltlichen Vertiefungen konzipiert, bei denen die Teilnehmenden die Schritte in Schulungsblogs von de.hypotheses sofort umsetzen können. Geübt werden grundlegende technische und grafische Einstellungen eines Blogs, das Anlegen eines Artikels, das Verschlagworten und Zuordnen von Kategorien. Die Übungen werden ergänzt durch eine Vertiefung der Diskussionen über Themenfindung, Aufbau und Gliederung von Beiträgen sowie sachliche Erschließung von Webinhalten.

Im dritten Teil wird das Einbinden von multimedialen Elementen geübt, kombiniert mit Exkursen zu Bildrechten, zu Anbietern von OA-Inhalten (Audio, Video), zu CC-Lizenzen, Impressum und zur DSGVO. Ebenso werden Hinweise zur Suchmaschinenoptimierung und zur Erhöhung der Sichtbarkeit von Bloginhalten diskutiert. Abschließend werden Tipps für die Anfangsphase eines Wissenschaftsblogs sowie zu Diskussionen über Kommentare gegeben. Die Schulungsblogs stehen den Teilnehmenden auch nach dem Workshop zur Verfügung, so dass sie Inhalte daraus in ihr zukünftiges Blog übernehmen können.

Max. 20 Teilnehmende

Beitragende

Dr. Mareike König ist stellvertretende Direktorin am Deutschen Historischen Institut und leitet das Blogportal *de.hypotheses*. Zu ihren Forschungsinteressen gehören: Digitale Geschichtswissenschaft, Wissenschaftskommunikation mit sozialen Medien, Wissenschaftliches Bloggen, Open Access.

Deutsches Historisches Institut Paris, 8, rue du Parc Royal, 75003 Paris, Frankreich, mkoenig@dhi-paris.fr

Prof. Dr. Anne Baillot ist Professorin für Germanistik an der Universität Le Mans. Zu ihren Forschungsinteressen gehören Digitale Editionen, Open Access, Archive, Wissenschaftsbloggen, Greening DH.

Faculté des Lettres, Langues & Sciences Humaines. Avenue Olivier Messiaen, 72085 LE MANS Cedex 09, France, anne.baillot@univ-lemans.fr.

Dr. Ulrike Stockhausen ist Community Managerin bei *de.hypotheses*.

Max Weber Stiftung, Rheinallee 6, 53173 Bonn, stockhausen@maxweberstiftung.de.

Fußnoten

1. Blogplattform für die Geistes- und Sozialwissenschaften, *hypotheses*: <https://hypotheses.org>, zugegriffen am 30.7.2022).

2. Katalog von *hypotheses*, <https://www.openedition.org/catalogue-notebooks?page=catalogue&pubtype=carnet&lang=en>, zugegriffen am 29.7.2022.

3. Im DFG-Kodex zur Sicherung guter wissenschaftlicher Praxis werden Blogs als akzeptiertes Format dezidiert aufgeführt, <https://zenodo.org/record/6472827#YueOaITP2Uk>. Zur Wissenschaftskommunikation generell siehe die Diskussion rund um das WÖM 2-Papier der Akademien der Wissenschaften von 2017 unter: <https://www.wissenschaftskommunikation.de/themen/woem/>, zugegriffen am 30.7.2022.

4. Mittelalterblog, <https://mittelalter.hypotheses.org>, zugegriffen am 29.7.2022.

Bibliographie

Baillot, Anne. 2022. "Chacun.e ses langues. Retour sur une expérience de blogging scientifique en anglais dans un contexte de recherche franco-allemand." *Traverse* 1: 104-107.

Bemme, Jens. 2022. "Kleine Editionen für Digital Humanities." *Public Humanities* 15. Juli. <https://publicdh.hypotheses.org/476>.

Döring, Karoline und Björn Gebert. 2022. "Digital, offen, dynamisch: Erfahrungen und Perspektiven aus 10 Jahren Mittelalterblog." *Traverse* 1: 93-99.

Gebert, Björn und Lena van Beek. 2019. "Wissenschaftsblogs als zeitgemäße Publikationsmedien: Das Beispiel Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte." *Mitteilungen des Deutschen Germanistenverbandes* 66/2: 273-281.

König, Mareike. 2015. "Herausforderung für unsere Wissenschaftskultur: Weblogs in den Geisteswissenschaften."

In *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität*, hg. v. Wolfgang Schmale, 57-74. Stuttgart: Franz Steiner Verlag.

König, Mareike. 2019. *Geisteswissenschaftliches Bloggen bei de.hypotheses. Erste Ergebnisse der Umfrage zu Motivationen, Praktiken und Routinen. Datenreport*. Paris: Deutsches Historisches Institut. <https://halshs.archives-ouvertes.fr/halshs-02150327v2>.

Mahrt, Merja und Cornelius Puschmann. 2014. "Science Blogging: an Exploratory Study of Motives, Styles, and Audience Reactions." *Journal of Science Communication* 13/3. http://jcom.sissa.it/archive/13/03/JCOM_1303_2014_A05/.

Sugimoto, Cassidy R. et al. 2017. "Scholarly Use of Social Media and Altmetrics: A Review of the Literature." *Journal of the Association for Information Science and Technology* 68/9: 2037-2062. <https://doi.org/10.1002/asi.23833>.

Tantner, Anton. "Wikipedia und Weblogs in der universitären Lehre." In *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität*, hg. v. Wolfgang Schmale, 45-56. Stuttgart: Franz Steiner Verlag.

Wuttke, Ulrike und Björn Gebert. 2021. "How to Make Your Medieval Research More Visible with Open Scholarship Methods and Tools." *Imago temporis: medium Aevum* 15: 415-450. <https://doi.org/10.21001/itma.2021.15.14>.

Offen für Professionalisierung? Wie Software und Entwickler*innen in den Digital Humanities gestärkt werden können

Czmiel, Alexander

czmiel@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften

Henny-Krahmer, Ulrike

ulrike.henny-krahmer@uni-rostock.de

Universität Rostock

Jettka, Daniel

daniel.jettka@uni-paderborn.de

Universität Paderborn

Hintergrund

Forschungssoftware zu entwickeln ist neben der Anwendung digitaler Methoden ein wichtiger Teil der Forschungstätigkeiten in den Digital Humanities, findet jedoch meist immer noch in einer wenig professionalisierten Form statt (siehe schon Schrade et al. 2018). Zwar ist es in manchen Bereichen der DH möglich, für bestimmte Aufgaben auf bestehende, außerhalb der DH selbst entwickelte und teilweise kommerzielle Softwarelösungen zurückzugreifen, z. B. auf den Oxygen XML-Editor (SyncRO Soft SRL 2022) für die Bearbeitung von XML-Dateien oder auf die Software Gephi (Bastian et al. 2009) zur Visualisierung von Netzwerkdaten. Häufig sind aber solche allgemeinen Tools für die geisteswissenschaftlichen Forschungsfragen und -projekte nicht geeignet und speziellere Werkzeuge nicht vorhanden. Das führt dazu, dass die digitalen Geisteswissenschaftler*innen selbst als Software-Entwickler*innen tätig werden oder eng mit Research Software Engineers (RSE) zusammenarbeiten, um neue Forschungssoftware zu entwickeln.

Ein typischer Fall sind digitale Editionen, die einerseits auf einem zugrunde liegenden Datenmodell und einer Repräsentation in Daten – den Forschungsdaten selbst – basieren. Andererseits nehmen sie in den meisten Fällen erst durch eine digitale Webpräsentation Gestalt an, die oft je nach Gegenstand der Edition maßgeschneidert entwickelt wird. Auch für Analysen von geisteswissenschaftlichen Daten kann zwar häufig auf bestehende Softwarekomponenten und -module zurückgegriffen werden, diese werden aber oft erst dadurch sinnvoll für die Forschung nutzbar, dass sie programmtechnisch für einen spezifischen Workflow eingesetzt und miteinander verknüpft werden, wie z. B. einzelne Python-Module für einen Textanalyseworkflow. Oder es werden ganz neue Analysetools für geisteswissenschaftliche Anwendungsfälle entwickelt, wie etwa im Fall des Tools Stylo (Eder et al. 2016) für stilometrische Analysen. Schließlich beinhaltet geisteswissenschaftliche Softwareentwicklung auch das Entwickeln von Skripten zur Aufbereitung und Umwandlung von Daten, damit diese wiederum in anderen Tools verarbeitet werden können.

Art und Umfang der Softwareentwicklung in den Geisteswissenschaften können also stark variieren. Genauso kann auch der Anteil und Stellenwert, den die Tätigkeit „Softwareentwicklung“ für Forscher*innen in den DH hat, sehr unterschiedlich sein. Das Entwickeln von Forschungssoftware kann einen Teil der gesamten Tätigkeit ausmachen, z. B. wenn ein Projekt individuell von Einzelnen umgesetzt wird, von einer geisteswissenschaftlichen Fragestellung über die Operationalisierung und technische Implementierung bis hin zur Interpretation und Aufbereitung der Ergebnisse. Softwareentwicklung kann aber auch für einzelne Forschende zur Haupttätigkeit werden, insbesondere in größeren Teams, die arbeitsteilig tätig sind, oder wenn die Entwicklung von Forschungssoftware im Mittelpunkt eines Vorhabens steht. Für diejenigen, die hauptsächlich mit der Entwicklung von Forschungssoftware beschäftigt sind, hat sich der Begriff *Research Software Engineer* etabliert.

Gerade dadurch, dass die Softwareentwicklung im geisteswissenschaftlichen Forschungskontext und in

den DH in den wenigsten Fällen von hierfür ausgebildeten Softwareentwickler*innen oder Informatiker*innen durchgeführt wird, gibt es in diesem Bereich noch viel Raum und viele Möglichkeiten für eine stärkere Professionalisierung. Häufig entsteht geisteswissenschaftliche Forschungssoftware in den ersten Stadien aus der jeweiligen wissenschaftlichen Domäne heraus. Die Forschenden selbst eignen sich hierfür durch Kurse oder Zusatzausbildungen die Grundlagen einzelner Programmiersprachen an. Für diese zusätzliche Kompetenz und auch die zusätzliche Arbeit, die durch die Softwareentwicklung entsteht, gibt es allerdings aus der Community nur selten die verdiente wissenschaftliche Anerkennung. Da weder das Wissen um eine stabile Softwarearchitektur noch ausreichend Zeit für Tests oder gar eine umfangreiche Dokumentation vorhanden ist, ist auch die Software selbst oft in einem verbesserungswürdigen Zustand und nicht für einen nachhaltigen Einsatz vorbereitet. Hier setzt dieser Workshop an.

Ziele des Workshops

Der Workshop wird von der AG „Research Software Engineering in den Digital Humanities“ des DHd-Verbands ausgerichtet und richtet sich an alle Wissenschaftler*innen, die im Rahmen ihrer Forschung oder in Forschungsprojekten Software entwickeln oder an der Entwicklung von Software beteiligt sind. Wir argumentieren, dass eine Offenheit für stärkere Professionalisierung seitens aller an geisteswissenschaftlicher Softwareentwicklung Beteiligten helfen kann, sowohl die Software selbst zu verbessern als auch die Position ihrer Entwickler*innen in der Wissenschaft zu stärken. Auch wenn Professionalisierung mit steigender Bedeutung von Softwareentwicklung im jeweiligen Kontext im Forschungsprozess und als wissenschaftliche Tätigkeit generell wichtiger wird, können doch alle Bereiche der DH, in denen Softwareentwicklung betrieben wird, von einer stärkeren Professionalisierung profitieren, die grob auf drei Ebenen beschrieben werden kann:

1. technische Professionalisierung

Die technische Professionalisierung betrifft in erster Linie die Bereiche, die in der täglichen Arbeit in der Softwareentwicklung auf der Ebene von Code und Architektur von Bedeutung sind, z. B. *Clean Code* (Martin 2009, Clean Code Developer 2022), Versionierung, Dokumentation, Tests, *DevOps* (Halstenberg et al. 2020) usw.

2. organisatorische Professionalisierung

Die organisatorische Professionalisierung bezieht sich auf alle Aspekte der Planung, des Projektmanagements und der (Selbst-)Organisation von Softwareentwickler*innen. Dazu gehören u. a. die Bildung von Entwickler*innen-Teams, die Verwendung von Versionsverwaltungs- und Ticketsystemen, Methoden der agilen Softwareentwicklung (Beck et al. 2001), aber auch Softwaremanagementpläne (SMPs).

3. institutionelle Professionalisierung

Die institutionelle Professionalisierung schafft die Rahmenbedingungen für die beiden anderen Bereiche, indem Softwareentwicklung Teil von DH-Curricula und die Ausbildung für RSEs in den DH verbessert wird, aber auch Karrierewege ermöglicht werden. Es geht um die Bereitstellung von Mitteln und Infrastruktur für die Ent-

wicklung von Software und die Verankerung der Tätigkeit einer/s RSE im wissenschaftlichen Bereich.

Ausgehend von den drei beschriebenen Ebenen soll der Workshop über mögliche Bereiche der Professionalisierung informieren und einen Erfahrungsaustausch und Diskussionen zum Thema ermöglichen. Konkretes Ziel des Workshops ist es, durch die Beteiligung aller Teilnehmerinnen und Teilnehmer in Gruppenarbeit ein White Paper zu erarbeiten, in dem Professionalisierungsoptionen für die Forschungssoftwareentwicklung in den DH festgehalten werden, so wie sich z. B. auch der Verein de-RSE fächerübergreifend für nachhaltige Softwareentwicklung ausgesprochen hat (Anzt et al. 2020).

Der Workshop soll so dazu beizutragen, dass in den DH mehr nachhaltige Forschungssoftware von hoher Qualität entsteht. Auf allen drei Ebenen - technisch, organisatorisch und institutionell - spielen hierbei auch Aspekte der Offenheit (Open Source, Open Access, Open Science) eine große Rolle, um eine vollständige Integrität und Transparenz aller Prozesse und Werkzeuge im Forschungszyklus zu gewährleisten und eine „gute wissenschaftliche Praxis“ zu ermöglichen. Mehr Professionalität in der Softwareentwicklung sorgt damit für eine bessere Forschungsunterstützung und bessere Forschung in den Geisteswissenschaften, gemäß dem Motto „Better Software, Better Research“ (Goble 2014). Zugleich trägt sie dazu bei, dass Softwareentwicklung als Forschungstätigkeit stärker sichtbar und anerkannt wird, was die Position von Entwickler*innen in den DH verbessert.

Ablauf und Inhalte des Workshops

Der Workshop ist für zwei halbe Tage geplant und setzt auf eine starke Mitarbeit aller Teilnehmenden. Als Ergebnis des Workshops ist angestrebt, ein White Paper mit Best Practices und Professionalisierungsoptionen und -schritten zu publizieren. Das Papier soll als Argumentationshilfe für DH-RSEs dienen, die sich eine professionellere Arbeitsumgebung innerhalb ihrer Institution wünschen. Zudem soll es die Institutionen selbst dabei unterstützen, die Professionalisierung der Forschungssoftwareentwicklung in den DH voranzubringen.

Im ersten Teil des Workshops werden von den Convenern der AG zur inhaltlichen Einführung in das Thema Kurzvorträge zu den Ebenen technischer, organisatorischer und institutioneller Professionalisierung gehalten. Daran schließt sich eine offene Gesprächsrunde mit allen Teilnehmenden an, in der jede und jeder Gelegenheit bekommen soll, auf folgende Fragen einzugehen:

- Wie sehr sehen Sie sich als RSE?
- Wie genau sieht Ihre Rolle als RSE aus?
- Wie schätzen Sie Ihren zukünftigen Karriereweg ein?
- Wie würden Sie Ihren eigenen Professionalisierungsgrad und den Ihrer Institution in Bezug auf Softwareentwicklung einschätzen?
- Halten Sie eine stärkere Professionalisierung in der Softwareentwicklung in den DH für sinnvoll und warum/warum nicht?

- Welche Aspekte halten Sie für eine professionelle Softwareentwicklung für besonders wichtig?

Nach der offenen Gesprächsrunde folgt im zweiten Teil eine gemeinsame Arbeit in Teilgruppen, wobei sich jede Gruppe auf einen der drei Hauptbereiche für Professionalisierung konzentrieren wird. Die Arbeit in den Teilgruppen wird gemeinsam geplant, indem definiert wird, woran jede Gruppe arbeiten und was als Ergebnis der Gruppenarbeit erwartet wird. Für die Dokumentation der Ergebnisse werden Online-Dokumente vorbereitet, in denen kollaborativ gearbeitet werden kann. Jede Teilgruppe wird von einer der einreichenden Personen geleitet. Auch kann jede/r Teilnehmende entscheiden, in welcher Gruppe er oder sie mitarbeiten möchte. Um eine produktive Arbeit sowohl in der Gesamtgruppe als auch in den Teilgruppen zu ermöglichen, ist die maximale Teilnehmerzahl auf 30 Personen begrenzt. Im Folgenden wird ein Überblick über das Programm gegeben:

Teil 1:

- Begrüßung (10min)
- Input-Statements von Seiten der AG zu Teilaspekten von Professionalisierung (30min)
- Offene Gesprächsrunde zu Professionalisierung - Teil I (80min)
- Pause (30min)
- Offene Gesprächsrunde zu Professionalisierung - Teil II (60min)
- Planung und Abstimmung zu Gruppenarbeit (30min)

Teil 2:

- Gruppenarbeit zu Teilaspekten I (90min)
- Pause (30min)
- Gruppenarbeit zu Teilaspekten II (60min)
- Ergebniszusammenführung (30 min)
- Diskussion und Abschluss (30 min)

Zur Durchführung des Workshops werden ein Beamer, ausreichend Steckdosen und WLAN benötigt.

Kontakt Daten der Beitragenden

Alexander Czmiel ist Leiter von TELOTA - IT/DH, der IT und Digital Humanities-Abteilung der Berlin-Brandenburgischen Akademie der Wissenschaften. Er ist Mitglied im Institut für Dokumentologie und Editorik, in der Gesellschaft für Forschungssoftware - de-RSE und Co-Convenor der AG „Research Software Engineering in den Digital Humanities“. Sein Forschungsinteresse liegt in der erfolgreichen Durchführung von Digital Humanities-Projekten und hier insbesondere in der nachhaltigen Softwareentwicklung, damit alle digitalen Forschungsergebnisse möglichst lange für die Nutzung erhalten bleiben. Kontakt: czmiel@bbaw.de

Ulrike Henny-Krahmer ist Juniorprofessorin für Digital Humanities an der Universität Rostock, Mitglied des Instituts für Dokumentologie und Editorik und Co-Convenorin der DHd-AG „Research Software Engineering in den Digital Humanities“. Ihre Forschung konzentriert sich auf Digitale Editionen und Textsammlungen, quantitative Textanalyse und die Evaluation und Nachhaltig-

keit von digitalen Forschungsergebnissen. Kontakt: ulrike.henny-krahmer@uni-rostock.de

Daniel Jettka ist Wissenschaftlicher Mitarbeiter im Projekt NFDI4Culture, wo er im Arbeitsbereich „Forschungswerkzeuge und Datendienste“ tätig ist. Insbesondere arbeitet er an der Koordination der technischen Infrastruktur in NFDI4Culture mit und ist Teil der Beratungsagentur für nachhaltige Softwareentwicklung. Er ist Co-Convener der DHd-AG „Research Software Engineering in den Digital Humanities“. Kontakt: daniel.jettka@uni-paderborn.de

Bibliographie

Anzt, Hartwig, Felix Bach, Stephan Druskat, Frank Löffler, Axel Loewe, Bernhard Y. Renard, Gunnar Seemann, et al. 2020. "An environment for sustainable research software in Germany and beyond: current state, open challenges, and call for action [version 2; peer review: 2 approved]." *F1000Research* 2021, 9:295. <https://doi.org/10.12688/f1000research.23224.2>.

Bastian, Mathieu, Sebastien Heymann und Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." In *International AAAI Conference on Web and Social Media*. <https://gephi.org/publications/gephi-bastian-feb09.pdf> (zugegriffen: 26. Juli 2022).

Beck, Kent, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland und Dave Thomas. 2001. "Manifest für Agile Softwareentwicklung." <http://agilemanifesto.org/iso/de/manifesto.html> (zugegriffen: 27. Juli 2022).

Clean Code Developer. 2022. „Clean Code Developer.“ <https://clean-code-developer.de/> (zugegriffen: 27. Juli 2022).

Eder, Maciej, Jan Rybicki und Mike Kestemont. 2016. "Stylometry with R: a package for computational text analysis." *R Journal* 8(1): 107-21. <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html> (zugegriffen: 26. Juli 2022).

Goble, Carole. 2014. "Better Software, Better Research." Software Sustainability Institute. <https://www.software.ac.uk/resources/publications/better-software-better-research> (zugegriffen: 27. Juli 2022).

Halstenberg, Jürgen, Bernd Pfizinger und Thomas Jestädt. 2020. *DevOps: Ein Überblick*. Wiesbaden: Springer.

Hettrick, Simon. 2020. „The growth and professionalisation of Research Software Engineering“. <https://slides.com/simonhettrick/rse-professionalisation-cw20> (zugegriffen: 28. Juli 2022).

Katerbow, Matthias und Georg Feulner. 2018. „Handreichung zum Umgang mit Forschungssoftware“. Zenodo. <https://doi.org/10.5281/zenodo.1172970>.

Martin, Robert C. 2009. *Clean Code. Refactoring, Patterns, Testen und Techniken für sauberen Code*. Frechen: mitp-Verlag.

Schrade, Torsten, Alexander Czymiel und Stephan Druskat. 2018. "Research Software Engineering und

Digital Humanities. Reflexion, Kartierung, Organisation." In *DHd 2018. Book of Abstracts*. <https://doi.org/10.5281/zenodo.4622564>.

SyncRO Soft SRL. 2022. „Oxygen XML Editor.“ https://www.oxygenxml.com/xml_editor.html (zugegriffen: 26. Juli 2022).

Pipelines für Natural Language Processing und digitale Literaturanalyse in spaCy

Varachkina, Hanna

hanna.varachkina@stud.uni-goettingen.de
Seminar für Deutsche Philologie, Georg-August-Universität Göttingen

Barth, Florian

florian.barth@uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Dönicke, Tillmann

tillmann.doenicke@uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Biermann, Johannes

johannes.biermann@gwdg.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Altmann, Friederike

friederike.altmann@stud.uni-goettingen.de
Seminar für Deutsche Philologie, Georg-August-Universität Göttingen

Neitzke, Thorben

thorben.neitzke@stud.uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Sporleder, Caroline

caroline.sporleder@cs.uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Kontext und Bedarf

In diesem halbtägigen Workshop stellen wir ein auf spaCy basierendes Pipeline-System für das Natural Language Processing (NLP) narrativer Texte vor und erproben mit den Teilnehmer*innen dessen praktische Anwendung, besonders im Hinblick auf Untersuchungsgegenstände der digitalen Literaturanalyse.

Die Analyse von literarischen Texten ist eine besondere Herausforderung für die automatische Sprachverarbeitung, da sie oft komplexe Interaktionen linguistischer Strukturen auf der syntaktischen, semantischen und pragmatischen Ebene betrifft. Für die Interpretation solcher Texte ist es zum Beispiel wichtig, neben traditionellen NLP-Verarbeitungsschritten wie Eigennamenerkennung, Sentiment-Analyse etc., auch komplexere Analysen durchzuführen, um z. B. die Sprechinstanzen im Text zu identifizieren, Bezüge zur realen Welt zu erkennen oder zeitliche Strukturen im Text zu analysieren. Auf der praktischen Ebene bedeutet dies, dass automatische Analysen in der digitalen Literaturwissenschaft in der Regel die (oft komplexe) Kombination mehrerer basaler Sprachverarbeitungswerkzeuge auf Token-, Teilsatz-, Satz- und Passagen-/Diskursebene erfordert. Dies ist in der Praxis nicht immer trivial, z. B. weil Ein- und Ausgabeformate verschiedener Werkzeuge nicht kompatibel sind.

Bibliotheken wie spaCy stellen sehr umfassende Sammlungen von Sprachverarbeitungswerkzeugen zur Verfügung, können potenzielle Nutzer*innen durch ihre Fülle und Heterogenität aber auch überfordern. Das vorgestellte Pipeline-System MONAPipe (Döncke u. a. 2022) soll hier Abhilfe schaffen, indem es Werkzeuge für linguistische und literaturwissenschaftliche Analysen komfortabel bündelt und flexibel erweiterbar ist. Der Fokus liegt dabei auf narrativen Texten und auf typischen Anwendungsszenarien der digitalen Literaturanalyse.

Der Workshop vermittelt (i) die Grundlagen von spaCy und dessen Kernkomponenten (Tokenisierung, Lemmatisierung, Erkennung von Satz- und Teilsatzgrenzen, Dependency Parsing), (ii) demonstriert, wie MONAPipe an die eigenen Zwecke durch Custom-Komponenten angepasst werden kann, und versetzt (iii) die Teilnehmer*innen mit hands-on Praxisbeispielen in die Lage, die in MONAPipe integrierten Komponenten zur Erschließung der linguistischen und narrativen Struktur eines Textes im Rahmen eigener Projekte kompetent auszuwählen, anzuwenden, zu erweitern und die Ergebnisse zu beurteilen. Unter anderem behandeln wir die Erkennung von Named Entities (sowie das Linking zu Normdaten; vgl. Barth u. a. 2022), von Zeitformen (Döncke 2020), Eventtypen (Vauth u. a. 2021) und Redeformen (direkte, indirekte, erlebte Rede; vgl. Brunner u. a. 2020), Animatheit (Tugener u. Klenner 2014) sowie Sentiment- (Remus u. a. 2010) und Emotionsanalyse (Mohammad u. Turney 2013). Schließlich erproben wir mit den Teilnehmer*innen, wie die Wechselwirkung einzelner Komponenten von MONAPipe Muster in Erzähltexten aufdecken kann, die zur Modellierung komplexer linguistischer und narrativer Phänomene geeignet sind (z. B. Generalisierungen (Gödeke u. a. 2022) oder narrative Kommentare (Weimer u. a. 2022)).

Technische Voraussetzungen

Wir stellen Jupyter-Notebooks bereit, in denen MONAPipe und alle benötigten Dependencies vorinstalliert sind. Die Teilnehmer*innen benötigen Kenntnisse in Python; Erfahrung im Umgang mit Jupyter-Notebooks und der Unix-Kommandozeile ist hilfreich.

Zielpublikum

Der Workshop ist als Tutorial geplant und richtet sich an Literaturwissenschaftler*innen, Linguist*innen, DH-Forschende, und andere Personen, die an Textanalyse interessiert sind. Die Teilnehmer*innen bekommen die Möglichkeit, die Funktionalitäten von MONAPipe auszuprobieren und in vorbereiteten Texten eine Reihe von Phänomenen automatisch zu identifizieren. Die Teilnehmerzahl ist auf 30 beschränkt.

Lernziele und Methodik

Der Workshop verfolgt mehrere Ziele: (1) Er soll die Teilnehmer*innen mit spaCy und dessen Kernkomponenten vertraut machen und Ihnen praktische Erfahrung in der Nutzung von MONAPipe für typische Textanalysekomponenten auf Token-, Satz-/Teilsatz- und Passagenebene vermitteln. (2) Darüber hinaus erproben die Teilnehmer*innen die Einbindung neuer Komponenten, um damit wie sie MONAPipe für eigene Zwecke anpassen können. Aufbauend auf diesen Grundlagen lernen die Teilnehmer*innen an einem konkreten Beispiel, (3) wie sie MONAPipe konkret für Forschungsprojekte insbesondere in der digitalen Literaturanalyse nutzen können. Dies umfasst die Auswahl geeigneter Komponenten für die Forschungsfrage sowie die Reflektion der Ergebnisse. Am Ende des Workshops haben die Teilnehmer*innen zum einen (i) ein besseres theoretisches Verständnis für die verschiedenen Sprachanalyseschritte, können komplexe Analysen durch Kombination mehrerer basaler Werkzeuge durchführen und die Qualität der automatischen Analyse beurteilen; Zum anderen (ii) haben die Teilnehmer*innen praktische Erfahrung im Umgang mit spaCy und verschiedenen Sprachverarbeitungswerkzeugen erworben und Problemlösungsstrategien für den Umgang mit NLP-Werkzeugen gelernt.

Methodisch kombiniert der Workshop Theorie und Praxis, wobei der Praxisanteil überwiegt. Um das Gelernte zu festigen und zu vertiefen, bekommen die Teilnehmer*innen zunächst kurze Arbeitsaufträge (zu den Sprachverarbeitungskomponenten) und später komplexere Aufgaben (zur Analyse narrativer Texte), deren Lösungen im Anschluss diskutiert werden. Der Praxisteil im zweiten Teil des Workshops bietet außerdem die Möglichkeit, MONAPipe für ein eigenes Forschungsproblem anzuwenden und dazu Feedback von den Organisator*innen des Workshops zu bekommen.

Auf technischer Ebene arbeiten wir mit der interaktiven Programmierumgebung Jupyter-Notebook und stellen vorbereitete und ausführlich dokumentierte Notebooks zur Verfügung, um einen möglichst reibungslosen Ablauf zu ermöglichen und den Teilnehmer*innen zu helfen, sich auf die Workshopinhalte zu konzentrieren.

Organisation und Ablauf

Wir planen einen vierstündigen Workshop bestehend aus zwei Blöcken. Der erste Block (1:45 h) beinhaltet aus einem einführenden Vortrag sowie einem Zeitslot zur Einrichtung der Jupyter-Notebooks, wobei die Organisator*innen nach Bedarf Hilfestellung bei der Einrichtung leisten. Anschließend erfolgt eine 45-minütige Session mit vorbereiteten Notebooks, bei der zunächst kürzere textuelle Phänomene auf Token-Ebene (wie Named Entities), Phänomene auf Teilsatz-Ebene (z. B. Zeitformen) sowie Phänomene, die längere Textpassagen umfassen (z. B. Redeformen), behandelt werden.

Im zweiten Block des Workshops (1:45 h) erstellen die Teilnehmer*innen eine eigene Komponente in spaCy. Anschließend erhalten die Teilnehmer*innen die Möglichkeit durch Lektüre narrative Strukturen in exemplarischen Textpassagen qualitativ zu bestimmen. Anhand der zur Verfügung stehenden spaCy-Komponenten soll evaluiert werden, welche Features sich zur Identifikation komplexer narrativer Strukturen eignen. Alternativ können die Teilnehmer*innen an eigenen Texten und Fragenstellungen arbeiten und hierfür Unterstützung durch die Workshoporganisator*innen erhalten.

Alle Teilnehmer*innen erhalten einen Gastaccount bei der GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen), um die Jupyter-Notebooks auf den GWDG-Jupyter-HPC-Servern nutzen zu können. Eine vorherige lokale Installation von Python oder zugehörigen Paketen ist nicht notwendig, dies wird im Vorfeld von der GWDG erledigt. Die HPC (High Performance Computing) Umgebung bietet die Möglichkeit, auch rechenaufwendige Pipelines zu testen. Im Rahmen der text- und sprachbasierten Forschungsdateninfrastruktur Text+ stellt die SUB Göttingen Schnittstellen bereit, mit denen literarische Texte aus der Digitalen Bibliothek in TextGrid direkt verwendet werden können.

Tabelle 1

Phase	Inhalt(e)	Zeit in Minuten
1. Einführung (Vortrag)	Grundkonzepte der maschinellen Sprachverarbeitung, narrativen Konzepten und der Programmiersprache spaCy	20
2. Einrichtung Jupyter-Notebooks	Technische Einrichtung und Kurzüberblick zur Funktionsweise von Jupyter-Notebooks	20
3. Textuelle Phänomene (Vortrag + hands-on)	Vorbereitete Jupyter-Notebooks mit Aufgaben zu textuellen Phänomenen mit unterschiedlichen Spans: <ul style="list-style-type: none"> • Token-Ebene (Named Entities, Zeitmarker) • Teilsatzebene (Zeitformen) • Passagen (Redeformen) 	1:05
Pause		25
4. Einbindung einer Custom-Komponente (hands-on)	Teilnehmer*innen integrieren eine eigene spaCy-Custom-Komponente (z. B. Fremdworterkennung)	45
5. Narrative Strukturen (hands-on + Diskussion)	Arbeitsaufgabe: narrative Strukturen in Texten erkennen	45
6. Abschluss		5

Nach dem Workshop

Wir tragen der Nachhaltigkeit der Forschung bei und stellen MONAPipe in einem Git-Repository zur Verfügung. Jupyter-Notebooks, die im Workshop benutzt wer-

den, werden in einem separaten Git-Repository zur Verfügung gestellt.

Forschungsinteressen der Beitragenden

Hanna Varachkina, M. A., ist wissenschaftliche Mitarbeiterin und Doktorandin am Seminar für Deutsche Philologie der Universität Göttingen. Ihre Forschungsinteressen liegen in computergestützter Textanalyse: Modellierung und Erkennung von Textstrukturen und Diskurs-Phänomenen.

Florian Barth, M. A., ist wissenschaftlicher Mitarbeiter und Doktorand am Göttingen Centre for Digital Humanities und Mitarbeiter der Abteilung Forschung und Entwicklung der SUB Göttingen. Seine Forschungsinteressen liegen im Bereich der computationellen Textanalyse mit besonderem Fokus auf narrativen und fiktionstheoretischen Phänomenen sowie in der konkreten Anwendung dieser Forschung im Bereich der Infrastrukturen für die Digital Humanities.

Tillmann Dönicke, M. Sc., ist wissenschaftlicher Mitarbeiter und Doktorand am Göttingen Centre for Digital Humanities der Universität Göttingen. Seine Forschungsinteressen liegen in der strukturellen Textanalyse, insbesondere im Zusammenhang mit Narration und Fiktion, sowie der automatischen Erkennung narrativer Phänomene.

Johannes Biermann, M. A., ist wissenschaftlicher Mitarbeiter bei der Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG) im Bereich High Performance Computing (HPC). Die GWDG erfüllt u.a. die Funktion eines Rechen- und IT-Kompetenzzentrums für die Universität Göttingen. Im Zuge des Verbund für Nationales Hochleistungsrechnen (NHR-Verbund) ist er Berater für Anwendungen aus dem Bereich Digital Humanities. Sein Forschungsinteresse ist es, DH Fragestellungen auf High-Performance-Computing-Cluster zu adaptieren und dort zu rechnen.

Caroline Sporleder, ist Professorin für Digital Humanities am Institut für Informatik der Universität Göttingen und Leiterin des Göttingen Centre for Digital Humanities. Ihre Forschungsinteressen liegen im Bereich der computationellen Semantik und Diskursanalyse, besonders für Anwendungen der Geistes- und Kulturwissenschaften.

Bibliographie

Barth, Florian, Hanna Varachkina, Tillmann Dönicke, und Luisa Gödeke. 2022. "Levels of Non-Fictionality in Fictional Texts." In Proceedings of The Eighteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation. 27-32.

Brunner, Annalen, Ngoc Duyen Tanja Tu, Lukas Weimer, und Fotis Jannidis. 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 5th SwissText & 16th KONVENS Joint Conference 2020.

Dönicke, Tillmann. 2020. "Clause-level tense, mood, voice and modality tagging for German." In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, 1-17.

Dönicke, Tillmann, Luisa Gödeke, und Hanna Varachkina. 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.

Dönicke, Tillmann, Florian Barth, Hanna Varachkina und andere. 2022. *MONAPipe: Modes of Narration and Attribution Pipeline*. (Softwarepublikation) URL: <https://gitlab.gwdg.de/mona/pipy-public>.

Gödeke, Luisa, Florian Barth, Tillmann Dönicke, Anna Mareike Weimer, Hanna Varachkina, Benjamin Gittel, Anke Holler und Caroline Sporleder. 2022 (zur Publikation angenommen). "Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung." In *Zeitschrift für digitale Geisteswissenschaften*.

Mohammad, Saif und Peter Turney. 2013. "Crowdsourcing a Word-Emotion Association Lexicon." In *Computational Intelligence*, 29 (3): 436-465.

Remus, Robert, Uwe Quasthoff, und Gerhard Heyer. 2010. "SentiWS - a publicly available German language resource for sentiment analysis." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). 1168-1171.

Tuggeger, Don und Manfred Klenner. 2014. "A hybrid entity-mention pronoun resolution model for German using Markov logic networks." In *Proceedings of the 12th edition of the KONVENS conference*, 21-31.

Vauth, Michael, Hans Ole Hatzel, Evelyn Gius and Chris Biemann. 2021. "Automated Event Annotation in Literary Texts." In *CHR 2021: Computational Humanities Research Conference*, November 17-19, 2021, Amsterdam, The Netherlands, 333-345.

Weimer, Anna Mareike, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Hanna Varachkina, Anke Holler, Caroline Sporleder und Benjamin Gittel. 2022 (zur Publikation angenommen). "The (In-)Consistency of Literary Concepts Operationalising, Annotating and Detecting Literary Comment." In *Journal of Computational Literary Studies*.

Semantic Web und Linked Open Data in den Geschichtswissenschaften

Kröger, Bärbel

bkroeger@gwdg.de
Akademie der Wissenschaften zu Göttingen,
Deutschland

Störiko, Johanna

johanna.danielzik@stud.uni-goettingen.de
Akademie der Wissenschaften zu Göttingen,
Deutschland

Wettlaufer, Jörg

jwettla@gwdg.de
Akademie der Wissenschaften zu Göttingen,
Deutschland

Seit der Veröffentlichung des ersten Konzepts eines „Semantic Web“ als Erweiterung des World Wide Web (Berners-Lee und Lassila Hendler 2001) haben sich GeisteswissenschaftlerInnen mit den Möglichkeiten und Grenzen der maschinenlesbaren Modellierung ihrer Daten im Rahmen dieses Entwurfs beschäftigt. Das Datenmodell des Resource Description Framework (RDF) und die Serialisierung in Turtle oder N-Triples hat sich zum Standard in der Modellierung von maschinenlesbaren semantischen Aussagen entwickelt. Obwohl sich eine Reihe von Erwartungen aus der Entstehungszeit des Semantic Web nicht erfüllt haben (umfassende Erweiterung des WWW mit semantischen Daten, Stabilität der Uniform Resource Identifier etc.), bildet das RDF-Datenmodell heute die Grundlage verschiedener Wissensbasen (DBpedia, Wikidata) und weiterer Wissensgraphen (knowledge graphs), die zurzeit in verschiedenen Zusammenhängen entstehen. Aus diesem Grund sind das Semantic Web und die Verlinkung von offen zugänglichen Daten (Linked Open Data) für die digitalen Geisteswissenschaften weiterhin und sogar verstärkt von Interesse (Beretta 2021, Beretta & Alameracy 2020, Hiltmann & Riechert 2020, Meroño-Peñuela 2017, Meroño-Peñuela et al. 2014, Pollin 2017, Wettlaufer 2018, Wettlaufer et al. 2015).

Der ganztägige Workshop bietet eine Einführung in die Thematik „Semantic Web und Linked Open Data“ mit einem Schwerpunkt auf den Geschichtswissenschaften. Das Angebot richtet sich an Teilnehmende ohne Vorkenntnisse im Bereich Semantic Web/Linked Open Data und eignet sich für Forschende aller Fachbereiche. Didaktisch teilt sich der Workshop in vier Teile, wobei die praktische Übung etwa zwei Drittel der Zeit beansprucht.

Zu Beginn des Workshops werden die Grundlagen des Semantic Web, des Resource Description Frameworks sowie damit verbundener www-Standards im Rahmen einer einführenden Darstellung behandelt. Besondere Aufmerksamkeit kommt dabei den Themen Wikidata¹ und SPARQL² zu, die in den anschließenden Übungen eine wesentliche Rolle spielen. Folgende Themenblöcke sind für diesen ersten, einführenden Teil vorgesehen:

Die Idee des Semantic Web: kurze Vorstellung der Grundidee, entwickelt aus dem Grundproblem der maschinellen Verarbeitung natürlicher Sprache. Das Resource Description Framework (RDF) als Grundlage für formalisierte Aussagen. Die Bedeutung stabiler URIs für die Idee des Semantic Web. Die Turtle Serialisierung von RDF als Grundlage für die Abfragesprache SPARQL. Namespaces und ihre Bedeutung, auch im Semantic Web. RDF-Schema und Ontologien zur Formulierung komplexer Aussagen. Linked Open Data, Knowledge Graphs und die LOD Cloud. Wikidata (und DBpedia) als zentrale

Knoten der LOD Cloud. Kurzer Exkurs zu alternativen Ansätzen zur Verlinkung von Normdaten: Beacon-Dateien. Übersicht zu Ressourcen im Semantic Web und in der LOD Cloud für die Geschichtswissenschaften.

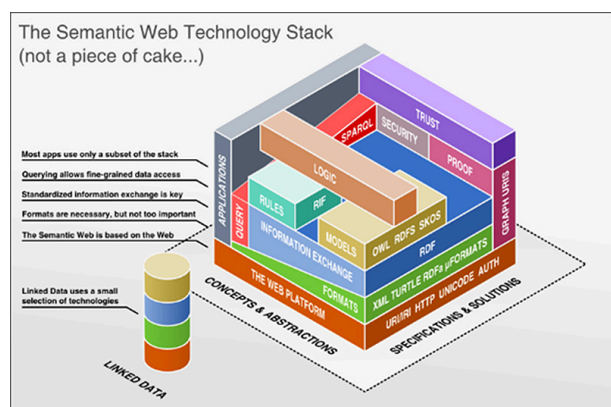


Abbildung 1. The Semantic Web Technology Stack. http://bnode.org/media/2009/07/08/semantic_web_technology_stack.png

Im zweiten Teil des Workshops sollen die Teilnehmenden in die Benutzung der Linked Open Data Plattform Wikidata eingeführt und Grundlagen für die Verwendung der Abfragesprache SPARQL gelegt werden.

Wikidata ist nicht nur der momentan größte frei verfügbare Wissensgraph, sondern bietet Daten unter freien Lizenzen und erlaubt genau wie die Wikipedia die freie Mitarbeit beim Aufbau der Wissensbasis. In der Übung lernen die Teilnehmenden somit eine der relevantesten Datenquellen für das Semantic Web kennen (Jacobsen et al. 2018). Außerdem ermöglicht Wikidata mit dem QueryService³ einen niederschweligen Einstieg in SPARQL, für den keine lokalen Installationen oder technischen Kenntnisse notwendig sind. Für die Übungen wird lediglich ein internetfähiges Gerät (vorzugsweise Laptop) sowie ein Wikidata-Account benötigt. Auch die graphische Benutzeroberfläche der Wikidata eignet sich gut für die Vermittlung der theoretischen Konzepte des Semantic Web, ohne dabei Kenntnisse der Informatik voraussetzen zu müssen.

Der erste Übungsblock erläutert zunächst die Datenstrukturen der Wikidata. Anhand eines Beispieles (Item) werden die theoretischen Konzepte aus dem ersten Teil des Workshops in ihrer Anwendung innerhalb der Wikidata gezeigt. Der Beispieles eintrag repräsentiert ein Item, das über Properties mit weiteren Items verbunden werden kann. Durch eine solche Verknüpfung entsteht ein Statement. Dieses kann wiederum durch sogenannte Qualifier näher beschrieben werden. Qualifier sind für die Nutzung der Wikidata in der Geschichtswissenschaft besonders relevant. Sie ermöglichen unter anderem die Modellierung der Herkunft einer Information. Als „Referenz“ können so Internetressourcen verlinkt werden, aus denen die Information entnommen wurde. Qualifier wie „Startzeitpunkt“ und „Endzeitpunkt“ erlauben die Spezifikation eines Zeitraumes, innerhalb dessen eine Information gültig ist. Neben dem Aufbau der Wikidata-Items behandelt dieser Teil des Workshops auch eine Einführung in das Benennungssystem der Wikidata und die verschiedenen RDF-Namespace, die dort verwendet wer-

den. Schließlich kann ein Item der Wikidata mit mehreren Labels ausgezeichnet werden, die eine Benennung und Beschreibung in unterschiedlichen Sprachen ermöglichen.

Eingeübt wird der Umgang mit Wikidata anschließend anhand von Datensätzen aus dem Forschungsprojekt Germania Sacra⁴, welches sich mit der Erforschung kirchlicher Institutionen und Personen des Mittelalters und der Frühen Neuzeit beschäftigt. Für den Workshop werden aufbereitete Forschungsdaten zur Verfügung gestellt, welche die Teilnehmenden selbstständig in die Wikidata einpflegen können. So ergänzen sie bestehende Datensätze zu Bischöfen des Alten Reiches um weitere Informationen wie ihren Begräbnisort. Die händische Eingabe der Daten vertieft das Verständnis für die Datenstrukturen, ist mit größeren Datenmengen aber nicht praktikabel. Als Ausblick auf den Einsatz von Wikidata im Forschungsalltag werden daher mit den Tools „Quickstatements“⁵ und „OpenRefine“⁶ Möglichkeiten zur seriellen Eingabe von größeren Datenmengen vorgestellt.

Der nächste Block der Übung behandelt die Grundlagen der Abfragesprache SPARQL. Ziel ist es, dass die Teilnehmenden ein Verständnis dafür entwickeln, wie geisteswissenschaftliche Fragestellungen als Abfrage formuliert und mit SPARQL auf Wikidata umgesetzt werden können. Dafür muss zunächst recherchiert werden, wie die in der Fragestellung enthaltenen geisteswissenschaftlichen Konzepte in der Wikidata modelliert sind. Anschließend wird eine passende Abfrage formuliert. Diese folgt mit SELECT, WHERE und gegebenenfalls OPTIONAL immer derselben Grundstruktur, die um weitere, komplexere Befehle ergänzt werden kann (siehe Abbildung 2). Diese Grundbausteine von SPARQL werden zunächst mit einfachen Abfragen wie „Finden Sie alle Datensätze zu Bischöfen mit einer WIAG-Kennung“ geübt. Vertiefend behandelt die Übung dann das Verketten von Abfragemustern und das Abfragen von Labels aus der Wikidata. Diese Grundlagen der Abfragesprache SPARQL werden durch Rückbezüge zum theoretischen Teil des Workshops auch mit den formalen Grundlagen des Semantic Web verknüpft. Während der Übung wechseln sich Demonstrationen neuer Konzepte und die Bearbeitung von aufeinander aufbauenden Übungsaufgaben ab. Während der Übungen sind die Teilnehmenden dazu eingeladen, sich mit anderen auszutauschen, Ergebnisse zu vergleichen und Fragen zu stellen. So erarbeiten sie sich Schritt für Schritt die Abfrage der im ersten Teil der Übung eingepflegten Datensätze.



Abbildung 2. SPARQL Abfrage im Wikidata Query-Service. Abrufbar unter <https://www.wiki/5FuN>

Das technische Framework, in das die Wikidata eingebettet ist, bietet den Nutzenden frei verfügbare Tools, mit denen abgefragte Daten graphisch dargestellt werden können. Dazu zählen ein interaktiver Graph, ein Zeitstrahl, eine Karte, eine Bildergalerie und viele andere Visualisierungsmöglichkeiten. Diese Tools werden zum Abschluss des praktischen Teiles vorgestellt und ausprobiert. Als Ergebnis dieses Hands on Übungsteils entsteht eine Visualisierung der Daten, die die Teilnehmer zu Beginn des Workshops in die Wikidata eingepflegt und mit den erworbenen SPARQL-Kenntnissen abgefragt haben.

Im vierten und letzten Teil des Workshops soll den Teilnehmenden ein Einblick in den Datenbestand der Wikidata zu geschichtlichen Themen und in die plattform-spezifische Modellierung dieser Daten vermittelt werden. Daran anknüpfend soll ein Blick auf die Potenziale von wikibase-basierten Wissensgraphen für die Geschichtswissenschaften geworfen werden. Die Vor- und Nachteile einer Datensammlung, die in einem kollaborativen Prozess entsteht, werden diskutiert. Dabei gilt ein kritischer Blick den Fragen der Datenqualität, der Datenmodellierung und der Vollständigkeit der Daten.

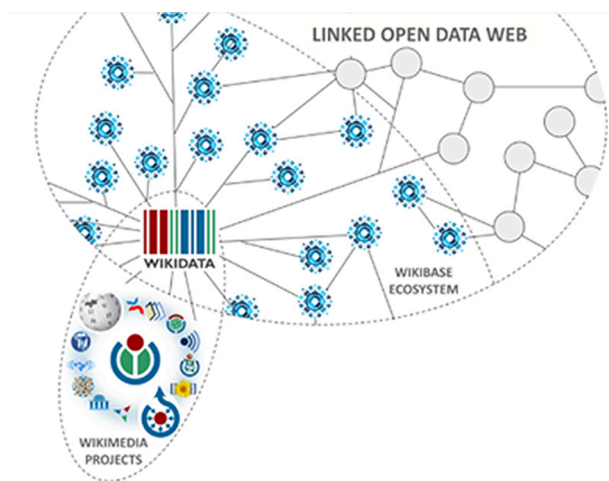


Abbildung 3. Darstellung des Linked Open Data Webs der Wikimedia (Quelle: https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Joint_Vision)

Auch Alternativen zur Wikidata werden in diesem Teil des Workshops thematisiert. Wikibase⁷, das technische Framework, das der Wikidata zugrundeliegt, kann auch als eigenständige, von Wikidata unabhängige Instanz verwendet werden. Diese Lösung hat das Potenzial, die Vorteile des Systems zur Verlinkung von Daten zu nutzen und gleichzeitig eine unabhängige und durch die Forschenden selbst kuratierte Datensammlung aufzubauen. Hierfür gibt es in den Digital Humanities einige Anwendungsbeispiele, die kurz vorgestellt werden.

Als Grundlage für eine anschließende praxisorientierte Betrachtung eigenständiger Wikibase-Instanzen dienen kleinere Pilotprojekte, die unter anderem im Rahmen von Lehrveranstaltungen an der Universität Göttingen realisiert wurden. Es wurde nicht nur die Wikidata, sondern auch die vom Forschungszentrum Gotha der Universität Erfurt betriebene Wikibase-Instanz FactGrid⁸ mit Daten angereichert. Im Fokus standen dabei die Bischöfe

des Alten Reiches, die ihre Servitienzahlungen an die päpstliche Kurie mit Hilfe des Florentiner Bankhauses der Familie Medici abwickelten. Mit den im Workshop erworbenen Kenntnissen können die Teilnehmenden diese Daten abfragen und sich einen Einblick in die konkrete Nutzung von Linked Open Data in den Geschichtswissenschaften – und auch für ihre eigene Forschungsfragen – verschaffen.

Zur Nachbereitung des Workshops werden den Teilnehmenden die Übungsaufgaben inklusive einer Musterlösung zur Verfügung gestellt. Sie erhalten außerdem in Form eines „Cheat Sheet“ eine Übersicht über alle in der Übung verwendeten Befehle.

Zielgruppe: HistorikerInnen und GeisteswissenschaftlerInnen ohne Vorkenntnisse in SWT und LOD.

Didaktisches Konzept: Einführende Vermittlung von Grundlagenwissen, Interaktive Übungen, Gruppenarbeit/Hands on Beispiele.

Erwartete Teilnehmerzahl: 5-25

Technische Ausstattung: Seminarraum mit HD Beamer
Vortragende:

Bärbel Kröger, M.A.

Akademie der Wissenschaften zu Göttingen

Geiststraße 10

37073 Göttingen

bkroege@gwdg.de

Bärbel Kröger arbeitet im Akademievorhaben Germania Sacra und forscht zum Einsatz von Linked Open Data im Bereich der mittelalterlichen Kirchengeschichte. Sie leitet ebenfalls das Linked Data Projekt WIAG (Wissensaggregator Mittelalter und Frühe Neuzeit).

Johanna Störiko, M. Sc.

Georg-August-Universität Göttingen

Institut für Digital Humanities

Nikolausberger Weg 23

37073 Göttingen

johanna.stoeriko@uni-goettingen.de

Johanna Störiko (geb. Danielzik) untersuchte in ihrer Masterarbeit historische Werbeanzeigen in Kulturzeitschriften der Jahrhundertwende mit digitalen Methoden. Sie interessiert sich für den Einsatz von Technologien des Semantic Web und Linked Open Data in den digitalen Geschichtswissenschaften.

Dr. Jörg Wettlaufer

Koordination Digitalisierung und Datenkuration | Digitale Akademie

Akademie der Wissenschaften zu Göttingen

Theaterstraße 7

37073 Göttingen

Germany

jwettla@gwdg.de

Jörg Wettlaufer leitet die Digitale Akademie der Wissenschaften zu Göttingen und forscht zu Themen der Digitalen Geschichtswissenschaft, insbesondere dem Einsatz Semantic Web Technologien in den Digitalen Geisteswissenschaften sowie zur Rechts- und Sozialgeschichte.

Geplanter Ablauf des Workshops:

9:00	Vorstellungsrunde und Einführung in die Veranstaltung (30 min.)
9:30	Teil 1: Einführung Grundlagen des Semantic Web und LOD (60 min.)
10:30	Kaffeepause
11:00	Teil 2: Übung mit Wikidata anhand von Beispielen (90 min.)
12:30	Mittagspause
13:30	Teil 3: Übung SPARQL auf Wikidata (90 min.)
15:00	Pause
15:30	Teil 4: Beispiele für den Einsatz von Wikidata und LOD in den Geschichtswissenschaften (90 min.)
17:00	Ende

Fußnoten

1. <https://www.wikidata.org/>
2. <https://www.w3.org/TR/sparql11-query/>
3. <https://query.wikidata.org/>
4. <http://www.germania-sacra.de>
5. <https://quickstatements.toolforge.org>
6. <https://openrefine.org/>
7. <https://www.wikimedia.de/projects/wikibase>
8. <https://database.factgrid.de>

Bibliographie

Beretta, Francesco. 2021. "A challenge for historical research: making data FAIR using a collaborative ontology management environment (OntoME)", *Semantic Web 12: 2*, Special issue on Semantic Web for Cultural Heritage. <https://doi.org/10.3233/SW-200416>

Beretta, Francesco and Vincent Alamertery. 2020. "Du projet symogih.org au consortium Data for History - La modélisation collaborative de l'information au service de la production de données géo-historiques et de l'interopérabilité dans le web sémantique." *Revue ouverte d'ingénierie des systèmes d'information* 1(3):1-15. <https://doi.org/10.21494/ISTE.OP.2020.0532>

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* 284(5), 34-43.

Hiltmann, Torsten and Thomas Riechert. 2020. "Digital Heraldry. The State of the Art and New Approaches Based on Semantic Web Technologies." In *L'édition en ligne de documents d'archives médiévaux*, ed. by Christelle Balouzat-Loubet, Turnhout, 102-125.

Jacobsen, Annika et al. 2018. "Wikidata as an intuitive resource towards semantic data modeling in data FAIRification." In *Semantic Web Applications and Tools for Health Care and Life Sciences. Proceedings of the 11th International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4HCLS 2018)*. Ed. by Christopher J.O. Baker, CEUR workshop proceedings Vol. 2275. <http://ceur-ws.org/Vol-2275/>

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach and Frank van Harmelen. 2015. "Semantic Technologies for Historical Research: A Survey." *Semantic Web Journal* 6: 539-564.

Meroño-Peñuela, Alberto. 2017. "Digital Humanities on the Semantic Web: Accessing Historical and Musical Lin-

ked Data," *Journal of Catalan Intellectual History (JOCIH)* 1(11): 144-149. DOI: 10.1515/jocih-2016-0013

Pollin, Christopher and Georg Vogeler. 2017. "Semantically Enriched Historical Data. Drawing on the Example of the Digital Edition of the 'Urfahdebücher der Stadt Basel'", In 2nd Workshop on Humanities in the Semantic Web (WHiSe), ed. by A. Adamou, E. Daga and L. Isaksen, 27-32.

Wettlaufer, Jörg. 2018. "Der nächste Schritt? Digitale Editionen und Semantic Web." In *Zeitschrift für Digitale Geisteswissenschaften*, Sonderheft "Digitale Metamorphosen", Hg. von Roland S. Kamzelak und Timo Steyer (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 2). DOI: 10.17175/sb002_007

Wettlaufer, Jörg, Christopher Johnson, Martin Scholz, Mark Fichtner, Sree Ganesh Thotempudi. 2015. "Semantic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science," *Digital Scholarship in the Humanities (DSH)*, Special Issue 'Digital Humanities 2014'; ed. by Melissa Terras, Claire Clivaz, Deb Verhoeven and Frederic Kaplan, 30 Supplement 1: i187-i198 https://academic.oup.com/dsh/article/30/suppl_1/i187/364720/

Skalierbare Blicke auf Leben und Werk: Visuelle Analyse und Kuratierung von kulturellen Objekten und Künstler*innen-Biographien

Windhager, Florian

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

Liem, Johannes

johannes.liem@donau-uni.ac.at
Universität für Weiterbildung Krems

Mayr, Eva

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

Schlögl, Matthias

matthias.schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Ebel, Carla

carla.ebel@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Probst, Stefan

stefan.probst@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Beck, Samuel

samuel.beck@vis.uni-stuttgart.de
Universität Stuttgart

Koch, Steffen

steffen.koch@vis.uni-stuttgart.de
Universität Stuttgart

Hintergrund

In den letzten Jahrzehnten wurde die Digitalisierung der materiellen Objektsammlungen von zahlreichen Kulturerbe-Institutionen vorangetrieben (Khan, Shafi, & Ahangar, 2018; Münster et al., 2019). Gleichzeitig wurde immaterielles Kulturerbe – wie biografisches Wissen über KünstlerInnen – digital erfasst und in biografischen Datenbanken verfügbar gemacht (Hyvönen, 2018; ter Braake et al., 2015; 2017). Diese Entwicklungen bieten eine gute Basis für eine digitale Analyse und Kommunikation des Lebens und Werks von Kulturschaffenden (Ruecker, Radzikowska, & Sinclair, 2016; Khulusi et al., 2016; Schlögl, Windhager, Mayr, & Kaiser, 2019; Windhager et al., 2018), jedoch verhindern mangelnde Verknüpfungen von lokalen Datensammlungen sowie fehlende Werkzeuge oft eine optimale Nutzung durch interessierte Forscher*innen und Praktiker*innen.

Die InTaVia-Plattform

Das InTaVia-Projekt (<https://intavia.eu>) arbeitet an der Reduktion solcher Barrieren und führt erstmalig materielles und immaterielles Kulturerbe mehrerer Länder in eine integrierte Datenbasis zusammen (Windhager, Mayr, Schlögl, & Kaiser, 2022, in Druck). Das Konsortium harmonisiert zu diesem Zweck nationale Kulturdatenbestände (inkl. Finnland, Niederlande, Österreich und Slowenien) und entwickelt ein prototypisches Informationsportal für die visuelle Analyse und Kommunikation dieser integrierten Kulturdaten. So wird eine synoptische Visualisierung und Betrachtung von historischen Daten zu Leben und Werken aus verschiedenen Perspektiven (geografisch, relational, kategorial, chronologisch) und auf verschiedenen Ebenen der Aggregation (von close bis distant reading) möglich.

Zielsetzung Workshop

Der Workshop ist als “Early-Access Workshop” für die InTaVia-Plattform konzipiert und zielt auf die Erprobung und Diskussion von prototypischen Visualisierungsmethoden, sowie auf den Austausch mit interessierten Forscher*innen in Feldern des digitalen kulturellen Erbes, der

digitalen (Kunst-)Geschichte und angrenzender Geisteswissenschaften. Zu diesem Zweck wird eine Diskussion der Thematik aus DH-Perspektive verbunden mit einer Vorstellung der InTaVia-Plattform und ihrer Technologien – mit spezifischen Fokus auf Module der Datenkuratierung und auf Methoden der visuellen Analyse. So können teilnehmende Expert*innen Einblicke in synoptische Methoden der Datenvisualisierung und -kuratierung gewinnen, während die Veranstalter*innen des Workshops mögliche Anregungen und Wünsche für die partizipative Weiterentwicklung der Plattform dokumentieren werden.

Bei Interesse soll der Workshop auch der Initiierung von gemeinsamen *Fallstudien* dienen. Für teilnehmende Expert*innen wird in diesem Fall im Nachfeld des Workshops ein Zugang zur prototypischen InTaVia-Plattform geschaffen, über den die Auswahl oder der Import von eigenen Daten mit Bezug zu individuellen Forschungsthemen möglich ist. In der Folge ist ein ausführlicherer Austausch zu den sich entwickelnden Möglichkeiten und Grenzen der Plattform angestrebt: Inhaltliche Expert*innen können die Tools der Plattform nutzen, um neue Einsichten in ihre jeweiligen Daten und Themen zu gewinnen und um Visualisierungen im Rahmen von gemeinsamen Fallstudien für eigene analytische oder kommunikative Zwecke zu nutzen. Feedback zu den sich entwickelnden Möglichkeiten und Grenzen der Plattform kann wiederum den Entwickler*innen der Plattform wertvolle Einblicke in die entscheidenden Bedürfnisse von Praktiker*innen liefern.

Ablauf Workshop:

1) Projektvorstellung: Die InTaVia-Plattform verknüpft Datensammlungen verschiedenen Typs (i.e. kulturelle Objektsammlungen und biografische Textsammlungen) zu einer integrierten Graphdatenbank (Abbildung 1). In einer kurzen Vorstellung werden die wichtigsten Forschungsfragen des Projekts gemeinsam mit seinen technologischen Zielen und Modulen vorgestellt. Dies inkludiert Information über das Datenmodell IDM (InTaVia Data Model), das Modul zur manuellen Kuratierung dieser Daten (Data Curation Lab) und das Modul zur visuellen Analyse von ausgewählten Daten und Themen (Visual Analytics Studio).

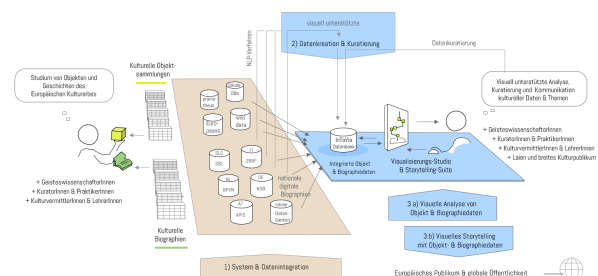


Abbildung 1: Architektur der InTaVia Plattform

2) Hands-On-Vorstellung von Datenmodell und Kuratierungsmodul: Mit Blick auf die transnationale Graphdatenbank (IKG - InTaVia Knowledge Graph) wird eine

kursorische Vorstellung der vier nationalen Biografiedatenprojekte zur Darstellung des integrierten Datenmodells führen. Auf diese Weise werden Teilnehmende mit den Facetten der Lebens- und Werkdaten vertraut, deren Analyse und Aufbereitung die InTaVia-Plattform unterstützt. Dies ist von besonderer Relevanz für die potentielle manuelle Aufbereitung und Zusammenführung von Daten (sowohl Biografie- wie auch kulturelle Objektdaten), welche in einem eigenen Datenkuratierungs-Modul angesiedelt ist. Anhand einer Auswahl von Arbeitsdaten für den Workshop werden hierbei die Möglichkeiten aufgezeigt, die sich aus einer etwaigen Nutzung der Plattform für eigene Fallstudien ergeben.

3) Hands-On-Vorstellung von Visualisierungswerkzeugen: Kulturelle Objektdaten und Biografiedaten haben eine Vielzahl von Facetten und Dimensionen die für Historiker*innen und Kulturwissenschaftler*innen von Interesse sein können. Zu diesen Dimensionen zählen die geografische Position von biografischen oder künstlerischen Ereignissen, diverse Kategorien von Ereignissen oder kulturellen Entitäten (Objekte oder Personen), Relationen zwischen Personen und/oder Objekten sowie chronologische Abfolgen und Zusammenhänge. Diese Aspekte können auf verschiedenen Ebenen der Aggregation - von historischen Individuen bis hin zu diversen Gruppierungen - für verschiedene Fragestellungen von Relevanz sein. Der Workshop wird zu diesem Zweck die Arbeit mit dem multiperspektivischen Visualisierungsmodul der InTaVia-Plattform ins Zentrum stellen und mit den Teilnehmer*innen skalierbare Blicke (inkl. close & distant reading) auf exemplarische Objekt- und Akteursdaten entwickeln (vgl. Abbildung 2).

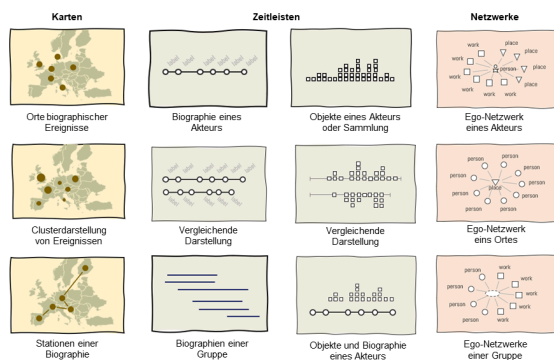


Abbildung 2: Überblick über relevante Visualisierungstechniken in InTaVia

4) Feedback: Während der explorativen Arbeit mit den Modulen der Plattform werden Fragen und Hinweise der Teilnehmer*innen notiert, die im Rahmen der weiteren Arbeit am Forschungsprojekt in die nutzer*innen-zentrierte Entwicklung der Plattform einfließen werden. Dazu werden sowohl die anonymisierten Notizen zu Aktivitäten des 'lauten Denkens' von Teilnehmer*innen dienen, wie auch die Rückmeldungen im Rahmen einer kurzen abschließenden Feedbackrunde.

Zielgruppe und Voraussetzungen

Der Workshop ist als Halbtagesveranstaltung geplant mit Fokus auf abwechslungsreiche Inputs zur Praxis, Erprobung und Evaluierung von aktuellen State-of-the-Art-Methoden der kulturellen Sammlungs- und Biografiedatenanalyse. Seine intendierte Zielgruppe reicht von interessierten Historiker*innen und Praktiker*innen bis hin zu Expert*innen der digitalen Geisteswissenschaften mit einem Schwerpunkt der Datenmodellierung, Kuratierung oder Visualisierung. Für die Teilnahme gibt es keine Voraussetzungen mit Blick auf inhaltliches oder technisches Vorwissen. Für die praktische Arbeit an den Daten genügt die Mitnahme eines Laptops. Die Gruppengröße ist auf 30 Teilnehmer*innen beschränkt. Mit Blick auf die technische Raumausstattung wird ein Beamer, ein Medienkoffer, sowie Whiteboards oder Pinnwände beantragt.

Fördernachweis: Das Projekt InTaVia (<https://intavia.eu>) wird von der Europäischen Kommission im Rahmen des H2020 Research and Innovation Programme, Grant Agreement No. 101004825 gefördert.

Bibliographie

Hyvönen, Eero, Leskinen, Petri, Tamper, Minna, and Tuominen, Jouni. 2018. "Semantic national biography of Finland." In Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018) . CEUR Workshop Proceedings.

Khan, Nadim Akhtar, Shafi, S. M., and Ahangar, Humma. 2018. "Digitization of cultural heritage: Global initiatives, opportunities and challenges." Journal of Cases on Information Technology (JCIT) 20 (4): 1-16.

Khulusi, Richard, Kusnick, Jakob, Focht, Josef, and Jänicke, Stefan (2019). "An interactive chart of biography". In 2019 IEEE Pacific Visualization Symposium (PacificVis) , 257-266. IEEE.

Münster, Sander, Apollonio, F. I., Bell, Peter, Kuroczynski, P., Di Lenardo, I., Rinaudo, F., and Tamborrino, R. 2019. "Digital cultural heritage meets digital humanities". In 27Th Cipa International Symposium: Documenting The Past For A Better Future , 812-820. ISPRS.

Ruecker, Stan, Radzikowska, Milena, and Sinclair, Stefan. 2016. Visual interface design for digital cultural heritage: A guide to rich-prospect browsing . Routledge.

Schlögl, Matthias, Windhager, Florian, Mayr, Eva, und Kaiser, Maximilian. 2019. *Biographische Informationssysteme (DPBs, Digital Knowledge Databases, Virtual Research Environments)* [Data set]. Zenodo. 10.5281/zenodo.2593761

ter Braake, Serge, Fokkens, Antske S., Sluijter, Ronald, and Declerck, Thierry. 2015. Biographical Data in a Digital World 2015: Proceedings of the First Conference on Biographical Data in a Digital World (BD2015) . CEUR Workshop Proceedings. <http://ceur-ws.org/Vol-1399/>

ter Braake, Serge, Fokkens, Antske, Sluijter, Ronald, Arthur, Paul, and Wandl-Vogt, Eveline. 2018. Biographical Data in a Digital World 2017: Proceedings of the Second Conference on Biographical Data in a Digital World 2017

(BD2017) . CEUR Wokshop Proceedings, 2119 . <http://ceur-ws.org/Vol-2119/>

Windhager, Florian, Mayr, Eva, Schlögl, Matthias, und Kaiser, Maximilian. 2022. "Visuelle Analyse und Kuratierung von Biographiedaten". In *Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft*, ed. K. Döring et al., 137-150. Amsterdam: DeGruyter. 10.1515/9783110757101-008

SPARQL für (digitale) Geisteswissen-schaftler:innen – Querying Wikidata und die MiMoTextBase

Hinzmann, Maria

hinzmannm@uni-trier.de

Trier Center for Digital Humanities, Universität Trier, Deutschland

Klee, Anne

klee@uni-trier.de

Trier Center for Digital Humanities, Universität Trier, Deutschland

Konstanciak, Johanna

konstanciak@uni-trier.de

Trier Center for Digital Humanities, Universität Trier, Deutschland

Röttgermann, Julia

roettger@uni-trier.de

Trier Center for Digital Humanities, Universität Trier, Deutschland

Schöch, Christof

schoech@uni-trier.de

Trier Center for Digital Humanities, Universität Trier, Deutschland

Steffes, Moritz

steffesm@uni-trier.de

Trier Center for Digital Humanities, Universität Trier, Deutschland

Einleitung

Nicht nur in Kultur- und Gedächtnisinstitutionen, auch in DH-Projekten ist derzeit eine Zunahme des Linked Open Data-Paradigmas sichtbar. Wie können Daten im

Sinne von „Open Data, Open Cultures“ offen, gut zugänglich, interoperabel vernetzt, maschinenlesbar und langfristig verfügbar dargeboten werden? Im Projekt „Mining and Modeling Text“ haben wir uns für die offene und kostenlose Software Wikibase entschieden, die einen eigenen SPARQL-Endpoint beinhaltet und neben Wikidata von einer wachsenden Anzahl an Forschungsprojekten verwendet wird.¹

Der Workshop setzt es sich zum Ziel, theoretisches und praktisches Wissen zur Modellierung geisteswissenschaftlichen und speziell literaturgeschichtlichen Wissens in Form von Linked Open Data (LOD) zu vermitteln, Einblick in die Syntax der Abfragesprache SPARQL zu geben und den Mehrwert der Aufbereitung von Daten als Wissensgraphen in Anwendungsszenarien aufzuzeigen. Dabei liegt der Schwerpunkt auf der Vermittlung von SPARQL in theoretischen und praktischen Sessions. Teilnehmende sollen die Kompetenz erlangen, die Struktur von SPARQL zu verstehen und eigenständig Queries zu schreiben.

```
1 #defaultView:BubbleChart
2 prefix wd:<http://data.mimotext.uni-trier.de/entity/>
3 prefix wdt:<http://data.mimotext.uni-trier.de/prop/direct/>
4 SELECT ?topLabel (count(*) as ?count)
5 WHERE {
6   ?item wdt:P36 ?top .
7   ?top rdfs:label ?topLabel .
8   filter(lang(?topLabel) = "de")
9 }
10 GROUP BY ?topLabel
11 ORDER BY desc(?count)
```

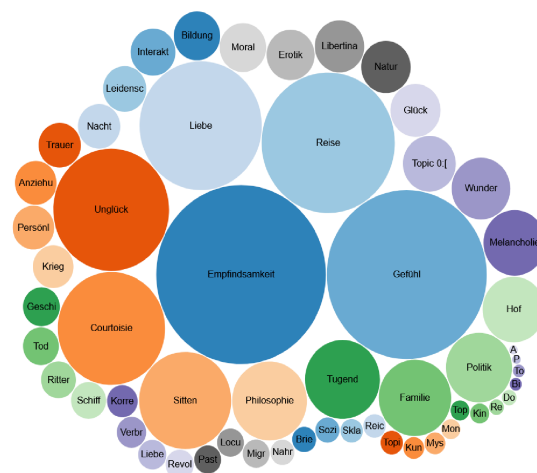


Abb. 1: Query und Ergebnisvisualisierung: Welche thematischen Konzepte sind im französischen Roman 1751-1800 vertreten? Beispiel: <https://tinyurl.com/2m8l3u99>.

Linked Open Data für die Geisteswissenschaften

Es ist zu beobachten, dass es ein zunehmendes Interesse in der DH-Community gibt, die eigenen Daten in Form von LOD zu veröffentlichen und mit dem Se-

semantic Web zu vernetzen oder die aktuellen Entwicklungen zu reflektieren (Hogan et al. 2021; Ikonik Ne# et al. 2021; Thornton et al. 2021; Alves 2022; Dörpinghaus 2022; Ohmukai / Yamada 2022; Zhao 2022). Auch das Projekt „Mining and Modeling Text“ hat es sich zum Ziel gesetzt, Daten aus unterschiedlichen Informationsquellen zu aggregieren und im Sinne des LOD-Paradigmas mit weiteren Ressourcen zu verknüpfen (Schöch et al. 2022). Der Mehrwert der aufwändigen Erschließung und Modellierung der Daten wird erst in den vielfältigen und flexiblen Abfragemöglichkeiten deutlich und ist demnach nicht loszulösen von SPARQL.

SPARQL (SPARQL Protocol and RDF Query Language) ist eine 2008 vom W3C veröffentlichte, graphenbasierte Abfragesprache für RDF (Resource Description Framework). RDF ist ein Datenmodell, mit dem sich Ressourcen im World Wide Web darstellen lassen. Es ist der zentrale Standard des W3C, der semantische Daten in der charakteristischen Tripel-Struktur bestehend aus ‚Subjekt – Prädikat – Objekt‘ repräsentiert. Ausgehend von einem einzelnen solchen Tripel wird die Struktur eines Knowledge Graphen im Workshop entfaltet und die „Übersetzung“ von Forschungsfragen in natürlicher Sprache in die SPARQL-Syntax erläutert.

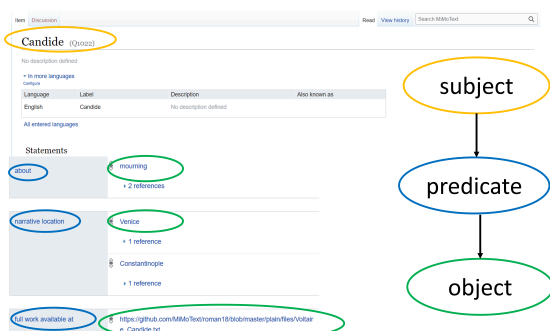


Abb. 2: Tripel-Struktur bestehend aus ‚Subjekt – Prädikat – Objekt‘, hier ein Beispiel zu einem literarischen Werk (Candide) aus der MiMoTextBase: <http://data.mimotext.uni-trier.de/wiki/Item:Q1022>.

Die Abfragesprache SPARQL setzt sich aus mehreren Bausteinen zusammen: *pattern matching* (Filtern des Datenbestands), *solution modifier* (Bearbeitung der Zwischenergebnisse) & *output* (Ausgabe als Tabelle oder Graph; Arenas et al. 2010). SPARQL ermöglicht es User:innen, durch Rekombination von Datensätzen neue Muster in den Daten zu erkennen und hypothesengeleitete Abfragen zu formulieren. Die Stärken dieser Abfragesprache und der Strukturierung von Wissen als RDF Triplestore sollen im Workshop in Anwendungsbeispielen gezeigt werden.

SPARQL-Abfragen werden häufig innerhalb eines einzelnen Knowledge Graphen gestellt. Es besteht jedoch auch die Möglichkeit, über mehrere Knowledge Graphen hinweg Abfragen zu stellen, sogenannte *federated queries* (Prud'hommeaux / Buil-Aranda 2013). Hier kommt das volle Potential von LOD zum Vorschein, denn so lässt sich Erkenntnisgewinn aus der Kombination mehrerer Graphen ziehen, ohne durch Replikationen von Datensätzen unnötige Redundanzen zu erzeugen. Im Workshop werden *federated queries* mit Wikidata erlernt und

es wird gezeigt, inwieweit Nutzen aus diesen gezogen werden kann.

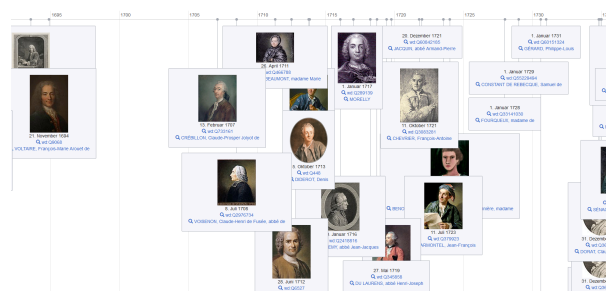


Abb. 3: *Federated queries* in SPARQL erlauben es, eigene Daten und externe Datenbestände (z. B. Wikidata) gleichzeitig abzufragen. Beispiel: <https://tinyurl.com/2ke42c3f>.

Workshop-Konzept

Der Workshop vermittelt Grundlagenwissen und Möglichkeiten, die das LOD-Paradigma bietet. Der im Projekt erstellte multilinguale Wissensgraph MiMoTextBase zur Domäne der französischen Literatur des 18. Jahrhunderts soll dabei als Anschauungsbeispiel dienen.

Lernziele

Der Workshop möchte praktisches Wissen vermitteln: Wie schreibt man SPARQL-Queries? Welchen Mehrwert kann ein Knowledge Graph für literaturgeschichtliche Fragen im Besonderen und die Geisteswissenschaften im Allgemeinen bieten?

In dem halbtägigen Workshop wird der Wissensgraph MiMoTextBase (Hinzmann et al. 2022a) des Projekts „Mining and Modeling Text“ vorgestellt und es werden Anwendungsszenarien formuliert und visualisiert. Dazu werden Grundlagen der Abfragesprache SPARQL erlernt und eigene Queries formuliert. Wir arbeiten beispielhaft auf dem projektinternen Knowledge Graphen sowie auf Wikidata und werden über *federated queries* die Verbindung mehrerer Wissensgraphen demonstrieren.

Konkrete Lernziele sind: Erwerb von Grundlagenwissen zu Semantic Web und RDF, LOD, Wikidata Graph; vertiefte Kenntnisse zu SPARQL und die praktische Fähigkeit, eigene SPARQL-Queries zu formulieren; Kennenlernen der Software Wikibase und Exploration der Visualisierungsmöglichkeiten des SPARQL-Endpoints.

Zielpublikum und Anforderungen

Der Workshop wendet sich an digitale Geisteswissenschaftler:innen mit Interesse an LOD und SPARQL. Spezielle Vorkenntnisse sind nicht notwendig. Teilnehmende benötigen einen Laptop.

Struktur / Ablauf

Der Workshop setzt sich aus aufeinander aufbauenden Sessions zusammen, die jeweils Input-Phasen und Übungsphasen verbinden. Es wird vorab eine ausführliche Tutorial-Seite (inklusive Verlinkung auf weitere hilfreiche Ressourcen zum SPARQL-Lernen) zur Verfügung gestellt, die den Teilnehmenden (und allen weiteren Interessierten) in der Vorbereitung sowie zur Vertiefung nützlich sein kann (Hinzmann et al. 2022b).

Im Zentrum des Workshops stehen drei Blöcke mit jeweils unterschiedlichem Schwerpunkt, in denen das Formulieren von SPARQL-Queries geübt wird (vgl. für Details den Ablauf im Appendix). Auch Teilnehmende ohne Vorkenntnisse werden schrittweise an zunehmend komplexere Queries herangeführt. Der Schwierigkeitsgrad wächst innerhalb der einzelnen Blöcke, wobei der Fokus auf dem eigenständigen Formulieren sowie Anpassen von Beispiel-Queries und dem Klären aller dabei auftretenden Fragen liegen wird.

1. Im ersten Teil liegt der Fokus auf Abfragen zu literarischen Werken. Im Hinblick auf SPARQL geht es hier zunächst um die zentralen Grundlagen wie das Schreiben einfacher *triple patterns* und Möglichkeiten der Kombination mehrerer *triple patterns* zu zunehmend komplexeren Queries. Der Mehrwert, der sich aus solchen Kombinationsmöglichkeiten ergibt, wird mit dem durch die MiMoTextBase gegebenen Fokus auf den französischen Aufklärungsroman besonders deutlich.

2. Im zweiten Teil widmen wir uns Wikidata als größtem öffentlichen Wissensgraphen, der sich zugleich als ‚Hub‘ begreifen lässt (Neubert 2017), und fokussieren Autor:innen als Entitäten. Autor:innen sind in allen geisteswissenschaftlichen Disziplinen relevant und ein wichtiges Scharnier zwischen verschiedenen Wissensgraphen. Bezogen auf die SPARQL-Syntax gehen wir einen Schritt weiter und integrieren Funktionen wie `OPTIONAL` und `FILTER`, um das Spektrum der Abfragemöglichkeiten zu erweitern. Ein Einstieg wird hier mit Queries zu Literatur:innen der MiMoText-Domäne gemacht. Im nächsten Schritt können die Teilnehmenden die Daten von Autor:innen in ihrer jeweiligen Domäne in Wikidata explorieren.

3. Der dritte Teil verknüpft die beiden vorigen Teile auf mehreren Ebenen. Der Schwerpunkt liegt auf *federated queries*, wobei wir uns auf Abfragen, die sich über die MiMoTextBase und Wikidata erstrecken, konzentrieren werden. Die genauere Betrachtung von Autor:innen des 2. Teils wird hier fortgesetzt und vertieft. In diesem abschließenden Teil wird der Mehrwert von Standards und geteilten Datenmodellen (Ontologien bzw. *entity schemata*) sowie die Verknüpfung von Ressourcen besonders deutlich.² Alle Autor:innen der MiMoTextBase, für die es auch Wikidata-Items gibt, können mit diesen über die Property *exact match* verknüpft werden, wodurch zusätzliche Informationen bereitstehen und diverse Abfragemöglichkeiten eröffnet werden.³ Die Wikibase-Infrastruktur bietet außerdem vielfältige Explorationsmöglichkeiten, die beispielhaft eingeführt werden (*marker cluster* für Geo-Daten, Timelines für Geburtsdatum u. ä.).

Es soll in der abschließenden Diskussion auch Raum sein, einen kritischen Blick auf Entwicklungen im Bereich des Semantic Web zu werfen, beispielsweise die Frage, welche Monopolisierungskräfte und Marktkräfte Einfluss nehmen (van Hooland / Verborgh 2014, 247–48; Singhal 2012). Zum Abschluss werden die wichtigsten An-

wendungsmöglichkeiten und Fragen zusammengetragen und weiterführende Ressourcen (DuCharme 2013; van Hooland / Verborgh 2014; Lincoln 2015; Blaney 2017) sowie bei Interesse Möglichkeiten der Kooperation thematisiert.

Appendix

Ablauf (4 Stunden)

10 Min.	Begrüßung (Vorstellung evtl. über Mentimeter)
20 Min.	Einleitung: Input zu Semantic Web, RDF, LOD für die Literaturgeschichte am Beispiel des Projekts Mining and Modeling Text. Wikidata & Wikibase Ecosystem, Mehrsprachigkeit des Graphen.
	SPARQL Teil 1 (MiMoTextBase)
20 Min.	(a) Input zu SPARQL-Grundlagen (interaktive Phase), SPARQL-Syntax, Möglichkeiten der Datenvisualisierung in Wikibase, Debugging & Help.
35 Min.	(b) Praxis-Teil: Anpassen vorhandener und Formulieren einfacher, eigener SPARQL-Queries auf der MiMoTextBase (Breakout Session bzw. Gruppenarbeit).
15 Min.	Pause
	SPARQL Teil 2 (Wikidata)
20 Min.	(a) Input: Erweiterte Elemente der SPARQL-Syntax wie <code>OPTIONAL</code> und <code>FILTER</code> ; Datenmodell für Autor:innen auf Wikidata.
35 Min.	(b) Praxis: Formulieren etwas komplexerer Queries auf Wikidata.
15 Min.	Pause
	SPARQL Teil 3 (Federated queries)
20 Min.	(a) Input: Fortgeschrittene SPARQL-Queries: <i>Federated queries</i> , <i>prefixes</i> definieren, <i>marker cluster</i> etc.
35 Min.	(b) Praxis: <i>Federated queries</i> etc. anwenden.
15 Min.	Abschließende Diskussion, Empfehlung weiterführender Ressourcen zur Vertiefung des Gelernten.

Organisatorisches

Maximale Zahl der Teilnehmenden: 25. Wir benötigen einen Raum mit WLAN und Beamer und bieten gern ein Hybrid-Szenario an.

Fördernachweis

„Mining and Modeling Text“ (Universität Trier, Trier Center for Digital Humanities) wird von der Forschungsinitiative des Landes Rheinland-Pfalz 2019-2023 gefördert.

Beitragende

Der Workshop wird von Mitarbeiter:innen des LOD-Projekts „Mining and Modeling Text“ durchgeführt. Das interdisziplinäre Projekt verfügt über einen eigenen SPARQL-Endpoint und wurde in Wikibase implementiert.

Maria Hinzmann; hinzmannm@uni-trier.de; Trier Center for Digital Humanities, Universität Trier | Historisches Seminar: Digital Humanities, Bergische Universität Wuppertal; Forschungsinteressen: Datenmodellierung, LOD, Textanalyseverfahren.

Anne Klee; klee@uni-trier.de; Trier Center for Digital Humanities; Forschungsinteressen: Digitale Textverarbeitung; Digitale Lexikographie.

Johanna Konstanciak; konstanciak@uni-trier.de; Trier Center for Digital Humanities; Forschungsinteressen: Digitale Textverarbeitung; XML/Web-Technologien.

Julia Röttgermann; roettger@uni-trier.de; Trier Center for Digital Humanities; Forschungsinteressen: LOD, Textmining-Verfahren wie Topic Modeling, NER und Sentiment Analysis.

Christof Schöch; schoech@uni-trier.de; Trier Center for Digital Humanities; Forschungsinteressen: Computational Literary Studies.

Moritz Steffes; steffesm@uni-trier.de; Trier Center for Digital Humanities; Forschungsinteressen: Softwaresysteme, Semantic Web Technologien, Forschungsinfrastrukturen.

Fußnoten

1. Einige aktuelle Beispiele für Forschungsprojekte, die Wikibase verwenden: Enslaved (Zhou et al. 2020), Rhizome Artbase (Rhizome 2021), FactGrid (Simons 2022; Brunner 2022).
2. Diese Verknüpfung entspricht dem 5. Stern im Linked Open Data-Modell von Berners-Lee (2006).
3. Vgl. dazu den entsprechenden Query auf der MiMo-TextBase: .

Bibliographie

Alves, Daniel, Hrsg. 2022. „IJHAC: A Journal of Digital Humanities. Special Issue: Linked Open Data in the Arts and Humanities“ 16 (1). <https://www.eupublishing.com/doi/epdf/10.3366/ijhac.2022.0271>.

Arenas, Marcelo, Claudio Gutierrez und Jorge Pérez. 2010. „On the Semantics of SPARQL“. In *Semantic Web Information Management: A Model-Based Perspective*, herausgegeben von Roberto de Virgilio, Fausto Giunchiglia, und Letizia Tanca, 281–307. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-04329-1_13.

Berners-Lee, Tim. 2006. „Linked Data – Design Issues“. 27. Juli 2006. <https://www.w3.org/DesignIssues/LinkedData.html>.

Blaney, Jonathan. 2017. „Introduction to the Principles of Linked Open Data“. *Programming Historian*. <https://programminghistorian.org/en/lessons/intro-to-linked-data>.

Brunner, Katharina. 2022. „FactGrid wants to become part of the Wikidata federation ecosystem“. 30. Mai 2022. <https://blog.factgrid.de/archives/2922>.

Dörpinghaus, Jens. 2022. „Wissensgraphen: Interdisziplinäre Perspektiven für Linked Data in den Geistes- und Sozialwissenschaften.“ *Zeitschrift für digitale Geisteswissenschaften* 07. https://doi.org/10.17175/2022_011.

DuCharme, Bob. 2013. Learning SPARQL. Sebastopol, UNITED STATES: O'Reilly Media.

Hinzmann, Maria, Anne Klee, Johanna Konstanciak, Julia Röttgermann, Christof Schöch und Moritz Steffes. 2022a. „MiMoTextBase“, Trier Center for Digital Humanities, <https://data.mimotext.uni-trier.de, 11/2022>.

Hinzmann, Maria, Anne Klee, Johanna Konstanciak, Julia Röttgermann, Christof Schöch und Moritz Steffes.

2022b. „MiMoTextBase Tutorial“. Juli 2022. https://mimotext.github.io/MiMoTextBase_Tutorial/.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, et al. 2021. „Knowledge Graphs“. *Synthesis Lectures on Data, Semantics, and Knowledge* 12 (2): 1–257. <https://doi.org/10.2200/S01125ED1V01Y202109DSK022>.

Hooland, Seth van und Ruben Verborgh. 2014. *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. London: Facet Publishing.

Ikonić Neić, Milica, Ranka Stanković und Biljana Rujević. 2021. „Serbian ELTeC Sub-Collection in Wikidata“. *Infotheca* 21 (2): 60–86. <https://doi.org/10.18485/infotheca.2021.21.2.4>.

Lincoln, Matthew. 2015. „Using SPARQL to access Linked Open Data“. Herausgegeben von Fred Gibbs. *The Programming Historian*, Nr. 4 (November). <https://doi.org/10.46430/phen0047>.

Neubert, Joachim. 2017. „Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings“. In *Proceedings of the 17th European NKOS workshop*. <http://ceur-ws.org/Vol-1937/paper2.pdf>.

Ohmukai, Ikki und Taizo Yamada, Hrsg. 2022. *Digital Humanities 2022. Conference Abstracts. Responding to Asian Diversity*. Tokyo: ADHO. <https://dh2022.dhii.asia/dh2022bookofabsts.pdf>.

Prud'hommeaux, Eric und Carlos Buil-Aranda. 2013. „SPARQL 1.1 Federated Query“. W3C Recommendation. 21. März 2013. <https://www.w3.org/TR/sparql11-federated-query/>.

Rhizome. 2021. „The ArtBase Relaunches: Welcome to Linked Open Data. Rhizome“. <http://rhizome.org/editorial/2021/apr/26/the-artbase-relaunches-welcome-to-linked-open-data>.

Sack, Harald und Mehwish Alam. 2020. „Knowledge Graphs“. Potsdam. <https://open.hpi.de/courses/knowledgegraphs2020>.

Schöch, Christof, Maria Hinzmann, Röttgermann Julia, Anne Klee und Katharina Dietz. 2022. „Smart Modelling for Literary History“. *IJHAC: International Journal of Humanities and Arts Computing [Special issue on Linked Open Data]* 16 (1): 78–93. <https://doi.org/10.3366/ijhac.2022.0278>.

Simons, Olaf. 2022. „FactGrid“, Forschungszentrum Gotha der Universität Erfurt, database.factgrid.de, 11/2022. <http://doi.org/10.17616/R31NJMQR>.

Singhal, Amit. 2012. „Introducing the Knowledge Graph: Things, Not Strings“. *Google (blog)*. 16. Mai 2012. <https://blog.google/products/search/introducing-knowledge-graph-things-not/>.

Thornton, Katherine, Kenneth Seals-Nutt, Marianne Van Renmoortel, Julie M. Birkholz und Pieterjan De Potter. 2021. „Linking Women Editors of Periodicals to the Wikidata Knowledge Graph“. *Semantic Web journal Special Issue Cultural Heritage* 2021. <http://www.semantic-web-journal.net/content/linking-women-editors-periodicals-wikidata-knowledge-graph>.

Zhao, Fudie. 2022. „How to Critically Utilise Wikidata - A Systematic Review of Wikidata in DH Projects“. In *Digital Humanities 2022 - Conference Abstracts*, herausgegeben von Ikki Ohmukai und Taizo Yamada, 608–10. The Univer-

sity of Tokyo, Japan: DH2022 Local Organizing Committee. <https://dh2022.dhii.asia/dh2022bookofabsts.pdf>.

Zhou, Lu, Cogan Shimizu, Pascal Hitzler, Alicia M. Sheill, Seila Gonzalez Estrecha, Catherine Foley, Duncan Tarr und Dean Rehberger. 2020. „The Enslaved Dataset: A Real-world Complex Ontology Alignment Benchmark using Wikibase“. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3197–3204. CIKM '20. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3340531.3412768>.

»textklang« – Ein Mixed-Methods-Workshop zu Lyrik in Text und Ton

Ketschik, Nora

nora.ketschik@ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Bernhart, Toni

toni.bernhart@ilw.uni-stuttgart.de

Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Gärtner, Markus

markus.gaertner@ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Koch, Julia

julia.koch@ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Schauffler, Nadja

nadja.schauffler@ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Kuhn, Jonas

jonas.kuhn@ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Einleitung

Die Überlieferung von Texten ist vorwiegend an die schriftliche Form gebunden, die bis zur Erfindung von

Tonaufnahmetechniken in der zweiten Hälfte des 19. Jahrhunderts die einzige Möglichkeit war, von Sprachbeiträgen nicht nur den Inhalt, sondern weitgehend auch die Darbietung festzuhalten. So haben sich literarische Traditionen und die wissenschaftliche Auseinandersetzung mit Literatur überwiegend entlang der schriftlichen Überlieferung entwickelt. Selbst bei Gattungen wie der Lyrik, in der Klang eine wichtige inhaltliche und ästhetische Rolle spielt (vgl. Richter et al. 2022), steht die Textform im Zentrum der kanonischen Überlieferung. Erst am Ende des 20. Jahrhunderts (u.a. angeregt durch die Sound Studies) haben sich in der Literaturwissenschaft Forschungsfelder zu Stimme, Klang, Akustik, Auditivität und Audioliteralität etabliert (Göttert 1998, Meyer-Kalkus 2001, Schulz 2018, Meyer-Kalkus 2020, Meyer-Sieckendiek 2020). Bis auf wenige Ausnahmen (z.B. Rhythmicalizer, vgl. Meyer-Sieckendiek et al. 2017) folgen die Digital Humanities bislang recht stark dieser eingespielten Zugangsweise – obgleich seit etwa 1900 unzählige Tonaufnahmen von Rezitationen vorliegen. Auch in der linguistischen Prosodieforschung und in der Sprachtechnologie wurde über die letzten Jahrzehnte ein Methodeninventar entwickelt, das eine sehr differenzierte Formulierung von Hypothesen zur Beziehung zwischen Text und lautlicher Realisierung erlaubt. Unser Workshop führt empirische Methoden aus der Phonetik mit aktuellen Technologien der Sprachsynthese und literaturwissenschaftlicher Forschung zur Lyrik der Romantik in einem Mixed-Methods-Workflow zusammen und bietet den Teilnehmenden auf diese Weise die Möglichkeit, das Wechselspiel von Textlichkeit und lautlicher Realisierung im Gedichtekorpus explorativ zu erkunden.¹

Der Workshop knüpft an Arbeiten aus dem BMBF-geförderten Projekt »textklang«² an. In »textklang« kooperieren das Deutsche Literaturarchiv (DLA) Marbach sowie das Institut für Maschinelle Sprachverarbeitung und das Institut für Literaturwissenschaft der Universität Stuttgart, die Expertise in unterschiedlichen relevanten Fachgebieten vereinen. Der Fokus des Projekts liegt auf der Erschließung und Analyse lyrischer Texte der Romantik, wobei der Zusammenhang zwischen dem geschriebenen Text und seiner lautlichen Realisierung in Rezitationen und Vertonungen in den Blick genommen wird.

Das beim Workshop verwendete Forschungskorpus zur Lyrik der Romantik speist sich aus der Mediendokumentation des DLA Marbach, die etwa 2700 Audioaufzeichnungen verschiedener Sprecher*innen seit den 1920ern beherbergt. Diese werden im Zuge des Projekts digitalisiert und um die dazugehörigen Metadaten und Transkripte ergänzt; darüber hinaus werden Texte und Rezitationen mit automatisch erzeugten Annotationen angereichert (siehe Schauffler et al. 2022b für eine Übersicht). Aktuell umfasst das »textklang«-Korpus 1261 Audioaufnahmen zu 786 Gedichten. Metadaten, Textdateien und lizenzfreie Audiodaten werden kontinuierlich über eine interaktive Webseite veröffentlicht.³

In unserem Workshop kommen alle Bereiche des Mixed-Methods-Workflows zum Einsatz, indem Ansätze aus traditionell sehr unterschiedlich arbeitenden Disziplinen zusammengeführt werden. Das Analysetool ICARUS (Gärtner et al. 2015) unterstützt den korpus- und textorientierten Zugang, bildet dabei aber neben morphosyntaktischen Annotationen der Texte auch die

phonetischen Annotationen der Rezitationen ab. Hierfür kommen Verfahren aus der Phonetik zum Einsatz, die die Eigenschaften des Sprachsignals systematisch erfassen. Sprachtechnologische Verfahren der Signalanalyse und -manipulation ermöglichen es so dann, bestimmte Annahmen über ein Re-Synthese-Tool kontrolliert zu testen. Der Bedarf für ein so weit gefasstes Methodenspektrum folgt aus den Grundeigenschaften des Untersuchungsgegenstands selbst. Der Workshop leistet einen Beitrag, die fachspezifischen Ansätze methodologisch zusammenzuführen und auf diese Weise den insbesondere für Lyrik zentralen Zusammenhang von Text und Klang in den Blick zu rücken.

Use-Cases

Idee des Workshops ist, dass die Teilnehmenden ihre eigenen Fragestellungen an Rezitationen von Lyrik der Romantik mitbringen können und darauf aufbauend während der Datenexploration Hypothesen entwickeln. Alternativ können die von uns vorgeschlagenen Fragestellungen aufgegriffen werden. Im Workshop thematisieren wir mehrere Use-Cases aus dem Projektkontext, darunter die Realisierung paralleler Strukturen (z.B. Reim, Satzbau), die unter strukturellen, semantischen und melodischen Aspekten von Interesse sind. Eine andere Fallstudie untersucht unterschiedliche Realisierungen von Enjambements (Schauffler et al. 2022a), die im Spannungsfeld von Vers- und Satzstruktur stehen. In Rezitationen können Sprecher*innen die syntaktische Einheit betonen, die Versgrenze markieren oder einen Mittelweg wählen (vgl. Tsur und Gafni 2019).

Ein weiterer Anwendungsfall, der exemplarisch etwas näher erläutert werden soll, beschäftigt sich mit Interjektionen. Interjektionen bezeichnen Ausrufe- oder Empfindungsworte (z. B. „ach“, „oh“, „juchhe“) und stehen im Grenzbereich von Schriftlichkeit und Mündlichkeit (Wharton 2003, Liedtke 2019). Sie nehmen eine syntaktische Sonderrolle ein und werden in der Linguistik als eigenständige Klasse behandelt, den Partikeln zugeordnet oder als Satzäquivalente angesehen (Liedtke 2019). Sie tragen einerseits denotativ keine Bedeutung, bringen andererseits Emotionen verschiedenster Art und in unterschiedlichen Intensitätsgraden zum Ausdruck (Schwarz-Friesel 2013, 155-157). Mit dem hier vorgestellten Mixed-Methods-Ansatz soll der Spielraum und der besondere textlich-klangliche (Zwischen-)Status von Interjektionen untersucht werden. Dabei interessiert zum einen die syntaktische Stellung von Interjektionen, zum anderen ihr Bedeutungsspektrum sowie, als dritter Aspekt, ihre lautliche Ausprägung. Die „Offenheit“ dieser Wortart legt die Hypothese nahe, dass die verschiedenen Ebenen sich gegenseitig beeinflussen können, beispielsweise das syntaktische Umfeld die lautlichen Realisierungen in der Rezitation prägt oder bestimmte klangliche Merkmale die Bedeutung von Interjektionen ausmachen.

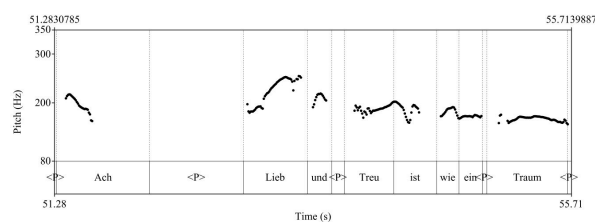


Abb.1

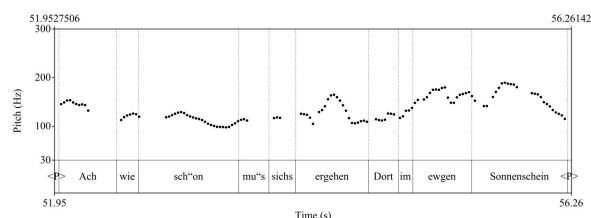


Abb.2

Die abgedruckten Beispiele deuten die syntaktisch-lautlichen Spielräume der Interjektion „Ach“ im Gedichtekorpus an: Während sie im ersten Beispiel syntaktisch isoliert steht (markiert durch den Tonhöhenverlauf und die Sprechpause), wird sie im zweiten Beispiel syntaktisch und lautlich in den Satz integriert. Auch die mit dem „Ach“ ausgedrückten Emotionen (im ersten Beispiel Schermerut, im zweiten Freude) changieren und werden – neben dem semantischen Kontext des Wortes – von der jeweiligen sprachlichen Realisierung beeinflusst. Mögliche Leitfragen für weitere Untersuchungen könnten sein: Welche syntaktischen Merkmale von Interjektionen gehen mit welchen lautlichen Merkmalen einher? Werden Interjektionen in gleicher (syntaktischer) Position lautlich parallel realisiert? Welche Varianz ist zwischen unterschiedlichen Sprecher*innen zu beobachten? Inwiefern beeinflusst die lautliche Realisierung die Bedeutung oder Wahrnehmung von Interjektionen?

Tools

Icarus

Für die Exploration und Visualisierung des Korpus mit allen Annotationsebenen verwenden wir ICARUS (Gärtner 2015) als Anfrageschnittstelle. ICARUS erlaubt eine gemeinsame Visualisierung von prosodischen Informationen und klassischen morphosyntaktischen Annotationen. Darüber hinaus können gezielt Anfragen unter Einbeziehung aller im Korpus verfügbaren Annotationsebenen gestellt werden, um Instanzen bestimmter Phänomene zu finden. An Annotationen stehen sämtliche für das GRAIN Korpus (Schweitzer et al. 2018) beschriebenen morphosyntaktischen und prosodischen Ebenen zur Verfügung. Darüber hinaus sind die Gedichte auch mit Markierungen zu Vers- und Strophenenden versehen, welche ebenfalls in Abfragen benutzt werden können. Je nach Entwicklungsfortschritt wird ICARUS als Desktop-Applikation⁴ eingesetzt oder in der Variante ei-

ner auf das »textklang«-Korpus zugeschnittenen Web-Oberfläche bereitgestellt.

IMS Speech Synthesis Toolkit Toucan

Die durch die Datenexploration entwickelten Hypothesen über Zusammenhänge zwischen Text und lautsprachlicher Realisierung sollen in Perzeptionsexperimenten untersucht werden. Mittels Sprachsynthese erstellen wir zu diesem Zweck eine prosodische Replikation der Originalaufnahmen, wobei phonetische Details (z.B. Lautdauer, Tonhöhe) gezielt manipuliert werden können (Koch et al. 2022). Unser Synthesemodell basiert auf der Modellarchitektur von FastSpeech 2 (vgl. Ren 2021), für die Implementierung nutzen wir das open-source Toolkit IMS Toucan⁵ (Lux et al. 2021, Lux und Vu 2022). Die Workshopteilnehmer*innen können über eine Bedienoberfläche mit dem Modell interagieren, indem sie spezifische, mit einem Phänomen verbundene Merkmale verändern und anschließend die Effekte der veränderten Parameter in der Perzeption testen. Beispielsweise kann die Länge, mit der ein Sprecher etwa das Versende markiert, verkürzt werden, die Tonhöhe an einer bestimmten Stelle angepasst oder die Dauer von Pausen verändert werden.

Ablauf und Ziele

Wir beginnen den Workshop mit einer Einführung in den multimodalen Ansatz und adressieren die methodologisch wie wissenschaftstheoretisch relevante Frage, wie die Spezialisierungen der Fachgebiete innerhalb der DH sinnvoll zusammengeführt werden können. Anschließend präsentieren wir mögliche Forschungsbeispiele und führen in die verwendeten Tools ein.

In zwei Praxisrunden haben die Teilnehmenden die Möglichkeit, das Lyrikkorpus zu erforschen, eigene Forschungsfragen zu entwickeln sowie diese exemplarisch zu untersuchen. Dies kann individuell oder in Kleingruppen geschehen. Die erste Praxisrunde dient der Exploration des Korpus und der Entwicklung möglicher Hypothesen. Hierfür kommt das Tool ICARUS zum Einsatz, über das die Teilnehmer*innen die verschiedenen Annotationsebenen (u.a. morphosyntaktisch, phonetisch) sichten und komplexe Suchanfragen an die Texte modellieren können. Auf Grundlage der Annotationen zur Text- und Lautgestalt können Forschungsfragen entwickelt oder eine der vorgestellten Fragestellungen aus der theoretischen Einführung exploriert werden. Nach einer Zusammenschau der Hypothesen dient die zweite Praxisrunde dazu, ausgewählte Fragestellungen probeweise zu validieren, indem die Annahmen in das Sprachsynthesemodell überführt werden. Wenn beispielsweise die Annahme besteht, dass die Länge und die Tonhöhe einen Einfluss darauf haben, ob die "bedeutungsfreie" Interjektion "Ach" negativ oder positiv konnotiert ist, können eben diese Merkmale in der Sprachsynthese gezielt modifiziert und die Effekte dieser Veränderungen getestet werden.

Die Ziele des Workshops bestehen folglich darin, die Möglichkeiten des Mixed-Methods-Ansatzes auszu-

schöpfen und Lyrik in ihrer Multimodalität erforschbar zu machen. Dabei liegt ein besonderer Schwerpunkt darauf, zu zeigen, wie fruchtbar das Zusammenspiel von textlicher und klanglicher Ebene sein kann. Zwar können die zu behandelnden Fragestellungen im Rahmen des Workshops nur ansatzweise durchgespielt werden, sie können dabei aber die Potenziale des interdisziplinären Ansatzes offenlegen.

Anhang

Zeitplan

1. Einführung und Ablauf (15 Min)
2. Theoretischer Teil (30 Min)
 1. Vorstellung der Projektidee
 2. Einführung in die Use-Cases
 3. Einführung in die verwendeten Tools (anschließende Pause, 15 Min)
3. Praktischer Teil
 1. Erste Praxisrunde: Exploration der Daten, Entwicklung von Hypothesen (45 Min)
 2. Sammeln der Ergebnisse, Vorstellung möglicher Fragestellungen (15 Min) (anschließende Pause, 30 Min)
 3. Zweite Praxisrunde: Bearbeitung der Fragestellungen, Syntheseexperimente (45 Min)
 4. Sammeln der Ergebnisse (15 Min)
4. Abschlussdiskussion (30 Min)

Teilnehmer*innen

Unser Workshop ist für ca. 20 Teilnehmer*innen geeignet und richtet sich an Interessierte aus den digitalen Geisteswissenschaften. Bestimmte technische Vorkenntnisse sind nicht erforderlich.

Technische Ausstattung

Die Teilnehmenden arbeiten an ihren eigenen Laptops. Ausreichend Steckdosen, stabiles Wifi und ein Beamer sollten vorhanden sein. Installationshinweise werden im Vorfeld an die Teilnehmer*innen verschickt.

Beitragende

Nora Ketschik (Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, nora.ketschik@ims.uni-stuttgart.de) ist wissenschaftliche Mitarbeiterin an der Universität Stuttgart. Sie promoviert zu Netzwerkanalysen von mittelhochdeutschen Romanen und setzt sich kritisch mit der Verwendung computergestützter Methoden für literaturwissenschaftliche Analyse Zwecke auseinander.

Toni Bernhart (Institut für Literaturwissenschaft, Universität Stuttgart, toni.bernhart@ilw.uni-stuttgart.de) ist Privatdozent für Neuere deutsche Literatur und wissenschaftlicher Mitarbeiter der Abteilung Digital Humanities an der Universität Stuttgart. Seine Forschungsschwer-

punkte sind die Imaginationsgeschichte von 'Volks-poesie', Auditivität und Literatur, Quantitative Literaturwissenschaft und Wissenschaftsgeschichte der Digital Humanities.

Markus Gärtner (IMS, Universität Stuttgart, markus.gaertner@ims.uni-stuttgart.de) ist wissenschaftlicher Mitarbeiter und Doktorand an der Universität Stuttgart und regelmäßig in der technischen Konzeption und Umsetzung von infrastrukturell fokussierten Projekten tätig.

Julia Koch (IMS, Universität Stuttgart, julia.koch@ims.uni-stuttgart.de) ist wissenschaftliche Mitarbeiterin und Doktorandin an der Universität Stuttgart. In ihrer Promotion arbeitet sie an Deep Learning Modellen für Sprachsynthese mit besonderem Fokus auf Kontrollierbarkeit.

Nadja Schauffler (IMS, Universität Stuttgart, nadja.schauffler@ims.uni-stuttgart.de) ist wissenschaftliche Mitarbeiterin an den Instituten für Maschinelle Sprachverarbeitung und Linguistik an der Universität Stuttgart und Postdoc im Projekt »textklang«, wo sie sich vor allem mit prosodischer Varianz beschäftigt.

Jonas Kuhn (IMS, Universität Stuttgart, jonas.kuhn@ims.uni-stuttgart.de) ist Professor für Computerlinguistik am Institut für Maschinelle Sprachverarbeitung und seit vielen Jahren an interdisziplinären Projekten zur Methodenentwicklung für die Digital Humanities beteiligt. Er ist federführender Projektleiter des BMBF-Projekts »textklang«.

Fußnoten

1. Rollen der Beitragenden: Nora Ketschik (Writing - original draft, Investigation, Methodology), Toni Bernhart (Writing - review and editing), Markus Gärtner (Software), Julia Koch (Software), Nadja Schauffler (Writing - original draft, Investigation, Methodology), Jonas Kuhn (Conceptualization, Methodology, Supervision).
2. <https://textklang.org/> (19.07.2022).
3. Interaktive Übersicht des »textklang«-Korpus: <https://clarin03.ims.uni-stuttgart.de/keshif/demo/textklang.html> (19.07.2022). Die Übersicht ist auch unter "Data" auf der Projektseite (<https://textklang.org/>) abrufbar.
4. ICARUS ist unter <https://github.com/ICARUS-tooling/icarus1-platform> (19.07.2022) bereits open source verfügbar und kann im Voraus von Teilnehmer*innen heruntergeladen werden.
5. <https://github.com/DigitalPhonetics/IMS-Toucan> (19.07.2022).

Bibliographie

Gärtner, Markus, Katrin Schweitzer, Kerstin Eckart und Jonas Kuhn. 2015. "Multi-modal Visualization and Search for Text and Prosody Annotations." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*.

Göttert, Karl-Heinz. 1998. *Geschichte der Stimme*. München: Fink.

Koch, Julia, Florian Lux, Nadja Schauffler, Toni Bernhart, Felix Dieterle, Jonas Kuhn, Sandra Richter, Gabriel Viehhauser und Ngoc Thang Vu. 2022. "PoeticTTS - Controllable Poetry Reading for Literary Studies." In *Proceedings of Interspeech 2022*.

Liedtke, Frank und Lena Rosenbaum. 2019. "Interjektionen und Kontextbezug. Pragmatische Templates als Analysemodell." In *Expressivität im Deutschen*, hg. von Franz d'Avis und Rita Finkbeiner, 129-148. Berlin/Boston: De Gruyter. 10.1515/9783110630190.

Lux, Florian, Julia Koch, Antje Schweitzer und Ngoc Thang Vu. 2021. "The IMS Toucan system for the Blizzard Challenge 2021." In *Proceedings of the Blizzard Challenge Workshop*.

Lux, Florian und Thang Vu. 2022. "Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Meyer-Kalkus, Reinhart. 2001. *Stimme und Sprechkünste im 20. Jahrhundert*. Berlin: Akademie Verlag.

Meyer-Kalkus, Reinhart. 2020. *Geschichte der literarischen Vortragskunst*. Berlin: Metzler. <https://doi.org/10.1007/978-3-476-04802-8>.

Meyer-Sickendiek, Burkhard. 2020. *Hörlyrik. Eine interaktive Gattungstheorie*. Paderborn: Fink.

Meyer-Sickendiek, Burkhard, Hussein Hussein und Timo Baumann. 2017. „Rhythmicalizer. Data Analysis for the Identification of Rhythmic Patterns in Readout Poetry." In *INFORMATIK 2017. Lecture Notes in Informatics (LNI) - Proceedings*, hg. von Maximilian Eibl und Martin Gaedke, 2189-2200. Bonn: Köllen Druck + Verlag GmbH (Series of the Gesellschaft für Informatik 275).

Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao und Tie-Yan Liu. 2021. "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech." In *International Conference on Learning Representations*.

Richter, Sandra, Toni Bernhart, Felix Dieterle, Gabriel Viehhauser, Gunilla Eschenbach, Jonas Kuhn, Nadja Schauffler, André Blessing, Markus Gärtner, Kerstin Jung, Nora Ketschik, Anna Kinder, Julia Koch, Thang Vu und Andreas Kozlik. 2022. "Der Klang der Lyrik. Zur Konzeptualisierung von Sprecher und Stimme, auch für die computationale Analyse." *Poema. Jahrbuch für Lyrikforschung / Annual for the Study of Lyrical Poetry / La recherche annuelle en poésie lyrique* 1 (im Erscheinen).

Schauffler, Nadja, Fabian Schubö, Toni Bernhart, Gunilla Eschenbach, Julia Koch, Sandra Richter, Gabriel Viehhauser, Thang Vu, Lorenz Wesemann und Jonas Kuhn. 2022a. "Prosodic realisation of enjambment in recitations of German poetry." In *Proceedings of the 11th international Conference on Speech Prosody*, 530-534. 10.21437/SpeechProsody.2022-108

Schauffler, Nadja, Toni Bernhart, André Blessing, Gunilla Eschenbach, Markus Gärtner, Kerstin Jung, Anna Kinder, Julia Koch, Sandra Richter, Gabriel Viehhauser, Thang Vu, Lorenz Wesemann und Jonas Kuhn. 2022b. "»textklang« - Towards a Multi-Modal Exploration Platform for German Poetry." In *Proceedings of the 13th edition of the Language Resources and Evaluation Conference (LREC)*, 5345-5355.

Schulz, Miklas. 2018. *Hören als Praxis. Sinnliche Wahrnehmungsweisen technisch*

(re-)produzierter Sprache. Wiesbaden: Springer (Auditive Vergesellschaftungen Hörsinn - Audiotechnik - Musikerleben). <https://doi.org/10.1007/978-3-658-19654-7>.

Schwarz-Friesel, Monika. 2013. *Sprache und Emotion*. 2. Aufl. Tübingen: Narr Francke Attempto Verlag.

Schweitzer, Katrin, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwen und Jonas Kuhn. 2018. "German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection." In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC)*.

Tsur, Reuven und Chen Gafni. 2019. "Enjambment - irony, wit, emotion. A case study suggesting wider principles." *Studia Metrica et Poetica* (5): 7-28.

Wharton, Tim. 2003. "Interjections, Language, and the 'Showing/Saying' Continuum." *Pragmatics and Cognition* 11(1): 39-91. [10.1075/pc.11.1.04wha](https://doi.org/10.1075/pc.11.1.04wha).

Wunsch und Wirklichkeit – Forschungsinfrastrukturen in den Computational Literary Studies: interdisziplinär, modular, vernetzt?

Jung, Kerstin

kerstin.jung@ims.uni-stuttgart.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg

Helling, Patrick

patrick.helling@uni-koeln.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

Kababgi, Daniel

daniel.kababgi@stud-mail.uni-wuerzburg.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg

Einleitung

Die Computational Literary Studies (CLS) befinden sich als Disziplin zwischen der Literaturwissenschaft, der Informatik und der Computerlinguistik. Sowohl aus einer theoretischen als auch einer methodischen Perspektive spielen unterschiedliche Aspekte aus allen angrenzenden Fachbereichen in den CLS eine Rolle. So kommen in den CLS bspw. computergestützte Verfahren wie maschinelles Lernen und Annotationen mit literaturwissenschaftlichen Fragestellungen zum Einsatz und sorgen für eine starke heterogene Prägung des Forschungsfeldes. Entsprechend ergibt sich auch eine diverse Landschaft an genutzten Datentypen und -formaten sowie lebender Systeme, bspw. Software, Tools, Visualisierungen und Plattformen, die es im Sinne der FAIR Prinzipien (Wilkinson et al. 2016) zu managen gilt. Ebenso kombinieren sich Konventionen der unterschiedlichen, fachlichen Teildisziplinen der CLS in der Nutzung von Technologien, Infrastrukturen und der Publikation von Ergebnissen und Daten. Hieraus ergibt sich eine Heterogenität des Forschungsfeldes sowie des spezifischen Forschungsdatenmanagements, wie sie sich auch grundsätzlich in den Geisteswissenschaften allgemein darstellt (Pempe 2012).

Ausgangslage

Das DFG Schwerpunktprogramm 2207 „Computational Literary Studies“ (SPP CLS)¹ bildet mit 10 geförderten Einzelprojekten sowie einem assoziierten Projekt seit 2020 einen Teil der deutschsprachigen CLS-Community. Im Rahmen des Zentralprojekts des Programms werden die einzelnen Teilprojekte individuell und projektübergreifend beim fachspezifischen Forschungsdatenmanagement (FDM) unterstützt.

Zu diesem Zweck wurde unter anderem eine Landschaftsvermessung vorgenommen, bei der in drei durch einen Leitfaden (Helling et al. 2020) gestützten Interviewrunden und einer Reviewrunde die Teilprojekte zum Umgang mit Forschungsdaten und lebenden Systemen, aber auch zu disziplinspezifischen Methoden und alltäglicher Projektarbeit befragt wurden. Zentrale Ziele der Landschaftsvermessung im SPP CLS sind die Entwicklung und Umsetzung einer möglichst fachspezifischen FDM-Strategie für das SPP CLS, sowie die Entwicklung einer Handreichung zu Best Practices und eines Anforderungsprofils für relevante/benötigte Infrastrukturen in den CLS.

Identifizierte Herausforderungen im FDM für die CLS

Neben einer grundsätzlichen Heterogenität in Bezug auf Datenformate und lebende Systeme in den CLS, für deren langfristige Sicherung und Verfügbarmachung kaum fachspezifische Lösungen existieren,² konnte im Rahmen der Landschaftsvermessung insbesondere auch die zentrale Herausforderung kollabo-

rativer Arbeit identifiziert werden. Viele Fragestellungen der CLS werden, wie in vielen anderen Fachdisziplinen auch, mit Hilfe diverser Methoden von interdisziplinären Teams aus Forschenden, gegebenenfalls an verschiedenen Standorten, bearbeitet. Dabei ist es wichtig gemeinsam an Daten und Dokumenten zu arbeiten, teilweise sogar gleichzeitig auf denselben Dateien.

Um die interdisziplinäre Zusammenarbeit zu fördern ist es daher unabdingbar, dass gemeinsam nutzbare Infrastrukturelemente verfügbar und leicht zugänglich sind. Institutionell aufgesetzte Versionskontrollsysteme, die den Zugang von Institutions-externen Forschenden nur über restringierte Gastzugänge zulassen, können dabei ebenso Hürden schaffen, wie verschiedene Vorgaben bezüglich der Nutzung von kommerziellen Angeboten oder proprietären Formaten.

Das Vorgehen und diese bisherigen Zwischenergebnisse der Landschaftsvermessung sowie pragmatische Lösungsstrategien zum Umgang mit Forschungsdaten in den CLS, wie bspw. der Betrieb einer gemeinsamen Gitlab-Instanz, wurden bereits mit den Communities der Digital Humanities (Helling et al. 2022a; Helling et al. 2022b) und des geisteswissenschaftlichen Forschungsdatenmanagements (Helling et al. 2021) diskutiert.

Ziele des vorgeschlagenen Workshops

Der Workshop soll eine oft implizit angenommene Ebene beleuchten, die im alltäglichen Umgang mit Forschungsdaten regelmäßig für kleine oder größere Ärgernisse sorgt oder sogar bestimmte Vorgehensweisen verhindert: Gemeinsames Arbeiten auf interaktiven Plattformen, Datenaustausch, unterschiedliche Datenformate, fehlende fachspezifische Infrastrukturangebote für die Publikation und Archivierung von Forschungsergebnissen sowie nicht mehr verfügbare oder lauffähige lebende Systeme wie Werkzeuge und Plattformen – dieser Ist-Zustand führt unter Umständen an entscheidenden Stellen zu pragmatisch-technischen Entscheidungen. Wir möchten die Community einladen, Erfahrungen aus ihrem Forschungsalltag zu teilen und Hürden aufzuzeigen, um dann gemeinsam eine Vision zu entwickeln, was wir benötigen um Wunsch und Wirklichkeit in Bezug auf Forschungsinfrastrukturen für die CLS in Einklang zu bringen, damit die technisch unterstützende Ebene ihre Rolle erfüllt und nicht zum Verhinderer wird.

Entsprechend möchten wir mit unserem Workshop die Ergebnisse der FDM-Landschaftsvermessung im SPP CLS als Ausgangspunkt nehmen und die damit verknüpften Fragestellungen mit der breiteren CLS-Community diskutieren, um das bisher entwickelte FDM-Anforderungsprofil der CLS um bisher ungesehene Aspekte ebenso zu erweitern wie die Konturen der identifizierten Best Practices zu schärfen. Der Workshop soll als offenes Forum verstanden werden, in dem die CLS-Community einerseits konkrete Bedarfe und Herausforderungen im FDM adressiert und an einem spezifischen, praxis- und community-getriebenen FDM-Bedarfsprofil arbeitet. Andererseits soll der Workshop auch auf operativer Ebene einen konstruktiven Austausch zwischen Fach-

wissenschaftler*innen der CLS und Datenmanager*innen ermöglichen.

Vor dem Hintergrund einer Ausgangslage mit Datentypen, Formaten und Methoden die - bedingt durch die Diversität der spezifischen Forschungsfragen - hochgradig heterogen ist, sollen unter anderem folgende Fragen in den Fokus genommen werden:

- Was benötigen Forschende der Computational Literary Studies für die tägliche Arbeit mit Forschungsdaten?
- Wie gelingt die Zusammenarbeit über Fach- und Institutionsgrenzen hinweg?
- Welcher Angebote bedarf es für die Sicherung, den Zugang, die Reproduzierbarkeit und Nachnutzbarkeit von CLS Forschungsergebnissen?

Dabei soll der Blick nicht nur auf disziplinspezifischen Werkzeugen und Infrastrukturen, wie sie z.B. über Initiativen wie DARIAH-DE³ und CLARIAH-DE⁴ zur Verfügung gestellt werden, liegen, sondern auch auf der disziplinspezifischen Nutzung von generischer Infrastruktur wie bspw. dem Forschungsdatenrepositorium Zenodo⁵ und der Softwareentwicklungs- und Versionskontrollplattform GitHub⁶.

Ein besonderes Augenmerk soll in diesem Zusammenhang auf der Unabhängigkeit von kommerziellen / proprietären Infrastrukturen liegen:

- Gibt es Zusammenhänge zwischen erzeugten Datenformaten und -strukturen und genutzten, proprietären Systemen?
- Welche Bedingungen verhindern möglicherweise die Nutzung spezifischer FDM-Lösungen, bspw. aufgrund von rechtlichen und finanziellen Hürden oder mangelnder Nachhaltigkeit?

Dabei ist es ein Anliegen des Workshops die Erfahrungen der Forschenden der CLS zu nutzen um strukturell wie anekdotisch den Ist-Zustand im Bezug zu den Ergebnissen aus dem Schwerpunktprogramm zu kartografieren und dabei Wunsch und Wirklichkeit einer interdisziplinären, modularen und vernetzten Infrastruktur in Beziehung zu setzen. Nicht zuletzt soll es um die Aussicht gehen, was von den digitalen Erzeugnissen der CLS die Chance hat auch in mehr als zehn Jahren noch nachvollziehbar zu sein.

Entsprechend möchten wir alle CLS-Community-Mitglieder und Interessierte einladen mit uns eine Bedarfskizze für die Vision einer fachspezifischen und für alle zugänglichen Forschungsinfrastrukturlandschaft anzufertigen, die

- Zusammenarbeit über Institutions- und (Bundes-)Ländergrenzen ermöglicht,
- Nachnutzbarkeit, Zugänglichkeit und Reproduzierbarkeit unterstützt sowie
- (Langzeit)Archivierung in den Blick nimmt.

Ablauf des Workshops

Der halbtägige Workshop wird in drei Teile gegliedert sein (siehe Tab. 1), die durch zwei 15-minütige Pau-

sen strukturiert werden. Im ersten Teil führen wir in Thema und Begriffe ein und berichten über Erfahrungen und Ergebnisse aus dem Forschungsdatenmanagement im DFG Schwerpunktprogramm „Computational Literary Studies“. Dieser Teil endet mit einer kurzen Onlineumfrage, in der der bisherige Umgang mit Methoden des Forschungsdatenmanagements sowie typische Problemfälle der Teilnehmenden abgefragt werden. Ähnlich dem Format der CRETA-Werkstatt (Reiter et al. 2020) werden wir im zweiten Teil Thementische zur Archivierungsinfrastruktur, Arbeitsinfrastruktur und lebenden Systemen anbieten, an denen in vor Ort gebildeten Gruppen Erfahrungen, Herausforderungen, Lösungen sowie Visionen und Wünsche formuliert und diskutiert werden können. Dabei wird jeder Tisch von einer*in der Workshop-Organisator*innen begleitet, um im dritten Teil des Workshops Umfrage und Ergebnisse der Thementische gemeinsam auszuwerten und Wunsch und Wirklichkeit in einem gemeinsamen Anforderungsprofil zu beschreiben, das wir im Anschluss an den Workshop über Zenodo veröffentlichen werden.

Tabelle 1: Zeitplan des Workshops.

	Dauer	Inhalt
Teil 1		
	0-30 Min.	Begrüßung und Einführung in das Thema / den Workshop
	30-45 Min.	Durchführung Onlineumfrage
	45-60 Min.	Kaffeepause
Teil 2		
	60-150 Min.	Durchführung Thementische (jeweils 30 Min.)
	150-165 Min.	Kaffeepause
Teil 3		
	165-240 Min.	Zusammenführung der Ergebnisse: Formulierung eines gemeinsamen Anforderungsprofils

Neben dem unmittelbaren Bezug zum Forschungsdatenmanagement in den Computational Literary Studies und der Erweiterung der Ergebnisse aus der Landschaftsvermessung im SPP CLS soll der Workshop grundsätzlich zur Sichtbarkeit der FDM-Bedarfe der CLS-Community in Infrastrukturinitiativen wie dem NFDI-Konsortium Text+⁷ und dem EU-geförderten CLS INFRA⁸ Projekt beitragen.

Adressat*innen des Workshops

Der Workshop richtet sich an etablierte und potentielle Mitglieder der CLS-Community und Interessierte, die Erfahrungen auf ähnlichen Gebieten, mit interdisziplinären Methoden zur Untersuchung von Textgrundlagen haben und sich für die Methoden und Fragestellungen der CLS interessieren. Darüber hinaus möchten wir explizit auch Expert*innen im Bereich des geisteswissenschaftlichen Forschungsdatenmanagements einladen am Workshop teilzunehmen. Die maximale Teilnehmendenzahl beträgt 20. Bei größerem Interesse können die interaktiven Teile des Workshops ggf. in zwei bis drei Iterationen durchgeführt und für die Teilgruppen mit der Onlineumfrage verschachtelt werden.

Als technische Ausstattung wird vor Ort der Zugang zu Strom, stabilem Internet und einem Projektor mit Lein-

wand/Projektionsfläche benötigt. Um sich an der Online-Umfrage beteiligen zu können, sollten die Teilnehmenden über ein digitales Endgerät verfügen.

Organisator*innen des Workshops

Kerstin Jung (Conceptualization, Writing – original draft) promovierte in der Computerlinguistik zum Thema der aufgabenbezogenen Kombination von automatisch erstellten Syntaxanalysen. Sie arbeitet im Zentralprojekt des SPP CLS zur disziplinspezifischen Unterstützung des FDM und bringt Erfahrung aus verschiedenen Infrastrukturprojekten und der Koordination kollaborativer Annotationsvorhaben ein. Ihre Forschungsinteressen liegen im Bereich der Nachhaltigkeit von Sprachressourcen und Abläufen sowie Metadaten- und Annotationsformaten.

Steffen Pielström (Conceptualization, Writing – review & editing) ist promovierter Biologe und arbeitet seit fast 10 Jahren im Bereich der Evaluation, Entwicklung und Vermittlung von quantitativen Methoden für die computergestützte Textanalyse in den Geisteswissenschaften. Er hat an verschiedenen Infrastrukturprojekten für die Digital Humanities mitgewirkt und ist zur Zeit im Zentralprojekt des SPP CLS tätig.

Patrick Helling (Conceptualization, Writing – review & editing) ist Medienwissenschaftler und Medieninformatiker. Er arbeitet im Zentralprojekt des SPP CLS und ist für die Entwicklung einer umfassenden FDM-Strategie für das Schwerpunktprogramm zuständig. Darüber hinaus ist er bereits seit 2017 am Data Center for the Humanities (DCH) an der Universität zu Köln tätig und dort Teil des FDM-Beratungsteams. Patrick Helling verfügt über Expertise im geisteswissenschaftlichen Forschungsdatenmanagement. Im Rahmen seiner Promotion arbeitet er an der Entwicklung eines formalen Beschreibungsmodells für das Management von Forschungsdaten.

Daniel Kababgi (Writing – review & editing) ist Masterstudent an der Universität Würzburg für Digital Humanities und Germanistik. Sein Studienschwerpunkt liegt auf NLP und dessen Anwendung innerhalb der Literaturwissenschaften. Das Hauptaugenmerk liegt auf der distanzierten Betrachtung der Literatur des 18. und 19. Jahrhunderts im Bezug auf die literarischen Epochen der Aufklärung und der Romantik.

Fußnoten

1. <https://dfg-spp-cls.github.io/> (letzter Zugriff: 02. August 2022).
2. Im Gegensatz zu anderen Fachbereichen und Disziplinen, wie bspw. die Linguistik (siehe u. a. Blumtritt und Rau 2018) oder die Medienwissenschaften (siehe u. a. Matuszkiewicz 2022).
3. <https://de.dariah.eu/> (letzter Zugriff: 02. August 2022).
4. <https://www.clariah.de/> (letzter Zugriff: 02. August 2022).
5. <https://zenodo.org/> (letzter Zugriff: 02. August 2022).
6. <https://github.com/> (letzter Zugriff: 02. August 2022).

7. <https://www.text-plus.org/> (letzter Zugriff: 02. August 2022).
8. <https://clsinfra.io/> (letzter Zugriff: 02. August 2022).

Bibliographie

Blumtritt, Jonathan und Felix Rau. 2018. "Nutzerunterstützung und neueste Entwicklungen in Forschungsdatenrepositorien für audiovisuelle (Sprach-)Daten." In *DHd2018: Kritik der digitalen Vernunft*. <https://doi.org/10.5281/zenodo.4622314>.

Helling, Patrick, Kerstin Jung und Steffen Pielström. 2021. "Disziplinspezifisches Forschungsdatenmanagement. FDM-Bedarfserfassung in den Computational Literary Studies." In *FORGE 2021 - Forschungsdaten in den Geisteswissenschaften: MAPPING THE LANDSCAPE - Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen. Konferenzabstracts*. 83-95. <https://doi.org/10.5281/zenodo.5379629>.

Helling, Patrick, Kerstin Jung und Steffen Pielström. 2022a. "Making Research Data FAIR. Seriously? Reflections on Research Data Management in the Computational Literary Studies." In *Digital Humanities 2022 Conference Abstracts*, 230-233.

Helling, Patrick, Kerstin Jung und Steffen Pielström. 2022b. "Pragmatisches Forschungsdatenmanagement - Qualitative und Quantitative Analyse der Bedarfslandschaft in den Computational Literary Studies". In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*, 193-199. <https://doi.org/10.5281/zenodo.6328021>.

Helling, Patrick, Kerstin Jung, Nils Reiter und Steffen Pielström. 2020. "Interviewleitfaden zur FDM-Bestandsaufnahme im Schwerpunktprogramm Computational Literary Studies." Zenodo. <https://doi.org/10.5281/zenodo.4269639>.

Matuszkiewicz, Kai. 2022. "Forschungsdaten in den Medienwissenschaften: Eine Auswertung von qualitativen Interviews zur Bedarfsermittlung für die Gestaltung eines medienwissenschaftlichen Forschungsdatenrepositoriums." In *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern 2/2022*, 1-14. <https://doi.org/10.17192/bfdm.2022.2.8433>.

Pempe, Wolfgang. 2012. "Geisteswissenschaften." In *Langzeitarchivierung von Forschungsdaten: eine Bestandsaufnahme*, 137-60. Boizenburg: wvh, Verlag Werner Hülsbusch.

Reiter, Nils, Gerhard Kremer, Kerstin Jung, Jansi Pangel, Axel Pichler und Benjamin Krautter. 2020. "Reaching out: Interdisziplinäre Kommunikation und Dissemination: Ein CRETA-Erfahrungsbericht" In *Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt* hg. von Nils Reiter, Axel Pichler und Jonas Kuhn, 467-484. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110693973-019>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3. Article number: 160018. <https://doi.org/10.1038/sdata.2016.18>.

Panels

Digitalisierung kulturellen Erbes und postkoloniale Perspektiven

Elwert, Frederik

frederik.elwert@rub.de
Ruhr-Universität Bochum, Deutschland

Berger, Claudia

claudia.berger@uni-erfurt.de
Universität Erfurt

High-Steskal, Nicole

nicole.high-steskal@donau-uni.ac.at
Universität für Weiterbildung Krems

Neudecker, Clemens

clemens.neudecker@sbb.spk-berlin.de
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz

Pons, Jessie

jessie.pons@rub.de
Ruhr-Universität Bochum

Akhlaq, Sara

sara.akhlaq@liu.se
Linköpings universitet

Panelthema

Der Nutzen der Digitalisierung von Objekten kulturellen Erbes scheint auf den ersten Blick selbstverständlich: Der Zugang zu Objekten wird unabhängig vom physischen Zugriff und somit die artefaktbezogene Forschung deutlich erleichtert. Artefakte, die bislang an verschiedenen Orten aufbewahrt werden, können digital zusammengeführt, kontextualisiert und miteinander in Beziehung gesetzt werden. Und nicht zuletzt kann die umfassende digitale Dokumentation auch angesichts der Bedrohung des kulturellen Erbes durch Raub, Krieg oder Umweltzerstörung zumindest ein digitales Gedächtnis sicherstellen.

Zugleich steht diese vermeintliche Selbstverständlichkeit einer kritischen Auseinandersetzung mit den epistemologischen Grundlagen und politischen Konsequenzen im Wege. Oder, wie es Monika Stobiecka zusammenfasst: "Digital archaeology has been used largely to avoid the politics and ethics of dealing with difficult questions concerning the field of heritage studies" (Stobiecka 2020; siehe auch Thompson 2017). Dabei bietet

die Praxis der Digitalisierung von Kulturerbe genügend Anknüpfungspunkte für eine kritische (Selbst-)Reflexion, insbesondere dann, wenn sie in postkolonialen Kontexten stattfindet. In diesem Panel geht es uns daher darum, einen Diskussionsraum zu eröffnen, in dem Anfragen an die Praxis der Digitalisierung und die eigene Rolle gestellt werden können.

Eine solche Auseinandersetzung muss letztlich schon beim Konzept des kulturellen Erbes ansetzen. Wie Stuart Hall in seinem Essay "Whose Heritage?" schreibt, hat der scheinbar unverdächtige Begriff eine unsichtbare, eingeschriebene Agenda, indem darin ein bestimmtes Bild der Vergangenheit und die Konstruktion einer oft national gefassten Identität in der Gegenwart eingeschrieben ist. Um diese unsichtbare Funktion kulturellen Erbes sichtbar zu machen, sei zu fragen: Wofür Kulturerbe, und für wen (Hall 2004)? Dies ist vor allem dann von besonderer Bedeutung, wenn das Kulturerbe im Kontext postkolonialer Machtgeflechte thematisiert wird. So wird etwa die Digitalisierung des kulturellen Erbes in Krisenregionen des Mittleren Ostens in erster Linie von westlichen Akteuren vorangetrieben, was Fragen nach einem neuen „digitalen Kolonialismus“ aufwirft (Thompson 2017, 155). Diese Fragen können und müssen dabei auf ganz verschiedenen Ebenen verhandelt werden. Grundlegend kann etwa gefragt werden, inwieweit Digitalisierungsprojekte hier in der Tradition der Archäologie des 18. und 19. Jahrhunderts stehen, in der westliche Akteure als wahre Kenner und Bewahrer vergangener Kulturen auftraten und vergleichbare Narrative der Errettung des Kulturerbes eine Rolle spielten (Thompson 2017, 162). In der konkreten Praxis kann gefragt werden, welche Kriterien für die Auswahl für Digitalisierungsprojekte eine Rolle spielen oder ob Open-Access-Mandate mit indigenen Vorstellungen von Zugang und Sichtbarkeit bestimmter Objekte in Einklang zu bringen sind (Man#uch 2017, 4–5). Auch rechtliche Fragen spielen hier eine Rolle, wenn etwa die physischen Objekte selbst aufgrund ihres Alters als gemeinfrei klassifiziert werden, auf die Digitalisate aber neue Schutzrechte reklamiert werden (Thompson 2017, 172).

Als Reaktion auf einige dieser Herausforderungen und als Gegengewicht zu den stark technisch formulierten FAIR-Prinzipien (Findability, Accessibility, Interoperability, and Reusability; Wilkinson u. a. 2016) wurden die CARE-Prinzipien (Collective Benefit, Authority to Control, Responsibility, Ethics) für "Indigenous Data Governance" formuliert (Carroll u. a. 2021). Diese Prinzipien lenken den Blick auf die Rechte und den Nutzen der Herkunftsgemeinschaften. In der Diskussion sind sie zunehmend als Korrektiv zu einem rein technischen Blick auf Daten eingeführt worden. Sie beziehen sich allerdings explizit auf die konkreten Bedarfe indigener Gemeinschaften und können daher nicht unterschiedslos auf andere postkoloniale Konstellationen übertragen werden.

Während die FAIR- und CARE-Prinzipien zumeist in Forschungsprojekten zum Tragen kommen, ist ihre Umsetzung in der Digitalisierung von großen Sammlungsbeständen, wie etwa in Museen, Archiven oder Bibliotheken, oft von vielen Ungewissheiten geprägt und stellt Institutionen vor enorme Herausforderungen. Die Rückgabe von Benin-Bronzen hat etwa die Frage des Umgangs mit den Objektdaten sowie der intellektuellen Autorität, die durch die Daten ausgeübt wird, aufgeworfen

(Geismar 2018, 111-2; Pavis und Wallace 2019; Wallace und Euler 2020). Die intellektuelle Autorität zeigt sich beispielsweise in den Objektbeschreibungen und den eingesetzten Vokabularen, die in kleinteiliger Arbeit von Kurator*innen erstellt wurden, um Objekte beschreiben und finden zu können. Einerseits ist es notwendig koloniale Begriffe in Sammlungsdatenbanken zu entfernen, damit koloniale Weltbilder durch die rasche Verbreitung von offenen Daten nicht weiter reproduziert werden. Andererseits sollten diese Begrifflichkeiten nicht vollständig eliminiert werden, da sie über die historische Entwicklung von Institutionen Auskunft geben können. Noch weitgehend lässt sich fragen, wie Metadaten schemata gestaltet werden müssten, um nicht nur einen westlichen Blick auf die Artefakte zum Ausdruck bringen. Das digitale Medium eröffnet neue Wege, um die intellektuelle Autorität neu zu verteilen, nämlich die Aufnahme von unterrepräsentierten Stimmen in Sammlungsdatenbanken und Objektbeschreibungen (Risam 2019, 9), doch stellt sich hier die Frage wer an der Wissensproduktion beteiligt sein darf, kann oder soll und in wie weit Infrastrukturen anderes Wissen und Wissensstrukturen zulassen (Scholz et al. 2021, 299-315).

Dieses Panel will diese und ähnliche theoretische Fragen in Beziehung setzen zu unserer eigenen Praxis in den Digital Humanities und in Digitalisierungsprojekten. Dabei geht es uns darum, einen Schritt zurückzumachen und mit einer gewissen Distanz noch einmal auf unser eigenes Tun und seine epistemologischen Grundlagen zu blicken. Das Ziel ist nicht, moralisch eindeutige Urteile zu fällen, sondern einen Raum zu eröffnen, in dem kritische Fragen gestellt, aber auch eigene Ansätze zum Umgang mit ihnen vorgestellt werden können.

Beiträge

Claudia Berger: Im Projekt „Kartographien Afrikas und Asiens“ (KarAfAs) digitalisieren wir einen ganz besonderen Bestand, der von Kolonialismus in verschiedenen Vermittlungsgraden durchdrungen ist. Unser Digitalisierungsvorhaben speist sich daher zu nicht unerheblichem Anteil aus dem Anliegen, dieses Material, das teils autoritativ koloniale Weltbilder zu manifestieren geholfen hat, teils durch indigene Mitproduzent*innen entstanden ist, global verfügbar zu machen und damit zugänglich für eben jene, die in jenen Gegenden leben und forschen, die von diesem kartographischen Material beschrieben wurden und betroffen waren. Die Frage ist allerdings, ob ein Digitalisierungsvorhaben diesem Anspruch gerecht werden kann. Hierin spielen Fragen der Digitalisierungskultur und des digital divide, aber auch der Aufbereitung der Sammlung zur Zugänglichmachung und Kommunikationsstrategien.

Nicole High-Steskal: Das Linking Viennese Art through Artificial Intelligence - Projekt beschäftigt sich mit dem Einsatz von Künstlicher Intelligenz, um die offenen Bestände von drei Museen in Wien zusammenzuführen. Die Digitalisierungsgrundlage der Museen ist sehr gut und ein Großteil der Objekte wurden zwischen 2002 und 2010 digital erfasst als koloniale Bezüge in Objektbeschreibungen und Vokabularen noch nicht thematisiert wurden, gleichzeitig lag der Fokus der Provenienzforschung vielfach auf Objekten, die zwischen 1938 und

1945 in die Sammlungen gelangt sind und nicht auf koloniale Bezüge. Im Rahmen des LiviaAI-Projektes stellt sich daher die Frage, wie wir sicherstellen können, dass unser Zugang rassistische oder koloniale Sichtweisen nicht reproduziert oder verstärkt, insbesondere bei großen Datensätzen, wo es nicht möglich ist alle Datensätze durchzuschauen. Um mögliche Biases in den Datensätzen besser einschätzen zu können, wurde im Projekt ein spezieller Fokus auf die Aufarbeitung der Digitalisierungsgeschichte und Kontextualisierung der einzelnen Datensätze gelegt. Der Beitrag stellt das LiviaAI-Projekt vor und greift ausgehend davon theoretische Fragestellungen auf.

Clemens Neudecker: Die Staatsbibliothek zu Berlin - Preußischer Kulturbesitz verfügt durch den Vollständigkeitsanspruch der „Sammlung Deutscher Drucke“ der SBB über besonders dichte Bestände aus den Jahren 1871 bis 1912. Im Zuge eines geplanten Digitalisierungsvorhabens „Digitalisierung von Quellen zur deutschen Kolonialzeit 1876-1919“ beabsichtigt die SBB die reichen Quellen aus der deutschen Kolonialzeit zu digitalisieren um so insb. Forschung und Projekte zu Fragen der Dekolonisation, aber auch der Herausbildung des Kolonialgedankens und der Gründung der deutschen Kolonien, zu unterstützen. Die einschlägigen Titel sind mittlerweile ganz überwiegend urheberrechtsfrei und können via Open Access digital bereitgestellt werden. Parallel wird in einem Forschungsprojekt der SBB zu Künstlicher Intelligenz (KI) für das digitale Kulturelle Erbe untersucht, wie digitalisierte Kulturdaten als Datensets („collections as data“) besser Eingang in die Entwicklung von Verfahren und Modellen aus dem Bereich der KI finden können, wobei insbesondere Fragen zur Provenienz und Kontextualisierung der Daten eine Rolle spielen, da im Bereich der KI verbreitete Daten und Modelle oft bereits über ethisch und sozial problematischen Bias verfügen (vgl. z.B. Bender et al. „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?“, 2021, und Inke Arns: „Can Artificial Intelligence be biased? On the critique of AI's 'algorithmic bias' in the arts“, 2022). Besonderer Fokus wird daher auf die Dokumentation und Kontextualisierung von digitalen Kulturdaten aus der Kolonialzeit als Datensets gerichtet, so dass die im Zuge des Digitalisierungsvorhabens entstehenden Daten als Use cases dienen können, um gemeinsam mit der Community Empfehlungen und Richtlinien zu erarbeiten, wie digitalisierte Kulturdaten aus kolonialen Kontexten angemessen dokumentiert, kontextualisiert und als Datensets für die Forschung zugänglich gemacht werden können.

Jessie Pons: Das Projekt „Digitalisierung Gandharischer Artefakte“ (DiGA) digitalisiert buddhistische Skulpturen, die in zwei Sammlungen pakistanischer Museen in der Provinz Khyber-Pakhtunkhwa aufbewahrt werden. Die Digitalisate und beschreibende Metadaten werden frei zugänglich gemacht. Dennoch stellen wir uns im Projekt die Frage, inwieweit unser Vorhaben Teil eines neuen digitalen Kolonialismus ist, wenn Infrastrukturen, Metadatenstandards und Einkommen innerhalb des westlichen Wissenschaftssystems verbleiben. Wir sind daher im engen Austausch mit lokalen Stakeholdern, insbesondere dem Direktorat für Archäologie und Museen (KP), um die Interessen und Bedürfnisse der pakistanischen Partner in der Planung und Umsetzung zu berücksichtigen.

sichtigen. Im Rahmen des Panels möchte ich einige unserer Ansätze vorstellen, aber auch offene Fragen und Herausforderungen diskutieren.

Zeitplan

Einführung ins Thema: 6 Minuten
 Vier Kurz-Inputs à 6 Minuten: 24 Minuten
 Diskussion im Panel: 30 Minuten
 Diskussion mit dem Publikum: 30 Minuten

Bibliographie

Arns, Inke. 2022. "Can Artificial Intelligence be biased? On the critique of AI's 'algorithmic bias' in the arts". Working paper. <https://zenodo.org/record/6797469>.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, und Margaret Mitchell. 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*, 610-623. <https://doi.org/10.1145/3442188.3445922>.

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, und Shelley Stall. 2021. "Operationalizing the CARE and FAIR Principles for Indigenous Data Futures". *Scientific Data* 8 (1): 108. <https://doi.org/10.1038/s41597-021-00892-0>.

Geismar, Haidy. 2018. *Museum Object Lessons for the Digital Age*. London: UCL Press. <https://doi.org/10.14324/111.9781787352810>.

Hall, Stuart. 2004. "Whose heritage? Un-settling 'the heritage', re-imagining the post-nation". In *The Politics of Heritage*, 37-47. Routledge.

Man#uch, Zinaida. 2017. "Ethical Issues In Digitization Of Cultural Heritage". *Journal of Contemporary Archival Studies*, Governance of Digital Memories in the Era of Big Data, 4 (4): 1-17.

Pavis, Mathilde, und Andrea Wallace. 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". *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3378200>.

Risam, Roopika. 2019. *New digital worlds: postcolonial digital humanities in theory, praxis, and pedagogy*. Evanston, Illinois: Northwestern University Press. Stobiecka, Monika. 2020. "Archaeological Heritage in the Age of Digital Colonialism". *Archaeological Dialogues* 27 (2): 113-25. <https://doi.org/10.1017/S1380203820000239>.

Scholz, Andrea, Thiago da Costa Oliveira, und Marian Dörk. 2021. "Infrastructure as digital tools and knowledge practices: Connecting the Ethnologisches Museum Berlin with Amazonian Indigenous Communities", in *Digitalisierung ethnologischer Sammlungen: Perspektiven aus Theorie und Praxis*, Hans Peter Hahn, Oliver Lueb, Katja Müller, und Karoline Noack, Bielefeld: transcript Verlag: 299-316. <https://doi.org/10.14361/9783839457900-017>

Thompson, Erin. 2017. "Legal and Ethical Considerations for Digital Recreations of Cultural Heritage". *Chapman Law Review* 20 (1/6): 153-76.

Wallace, Andrea, und Ellen Euler. 2020. "Revisiting Access to Cultural Heritage in the Public Domain: EU and International Developments". *IIC - International Review of Intellectual Property and Competition Law*. <https://doi.org/10.1007/s40319-020-00961-8>.

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". *Scientific Data* 3 (März): 160018. <https://doi.org/10.1038/sdata.2016.18>.

Forschungsdaten- infrastruktur als offene Werkstatt: Community Building zwischen generischen und datenspezifischen Praktiken

Hagener, Malte

malte.hagener@uni-marburg.de
 Philipps-Universität Marburg, Deutschland

Stanicka-Brzezicka, Ksenia

ksenia.stanicka@uni-marburg.de
 Philipps-Universität Marburg, Marburg Center for Digital Culture and Infrastructure, Herder-Institut für historische Ostmitteleuropaforschung

Krause, Celia

celia.krause@fotomarburg.de
 Deutsches Dokumentationszentrum für Kunstgeschichte, Bildarchiv Foto Marburg

Eggersgluß, Christoph

christoph.eggersgluess@uni-marburg.de
 Philipps-Universität Marburg, Marburg Center for Digital Culture and Infrastructure

Anliegen

Sowohl die in der Forschungslandschaft inzwischen gut etablierten Digital Humanities, als auch das gewünschte Zusammenwachsen der Prozesse in den verschiedenen GLAM-Einrichtungen wären ohne Community Building jenseits der tradierten Communities nicht denkbar. Der Aufbau dieser neuen Fachgemeinschaften zwingt zu einer Reflexion der etablierten Arbeitsweisen

und bezieht die Verwendung von Richtlinien und Werkzeugen für den Umgang mit Daten ein. Unser Anliegen ist es zu verdeutlichen, auf welche Weise sich eine gemeinsame Infrastruktur für Forschungsdaten und das Community Building gegenseitig beeinflussen können. Eine solche Infrastruktur kann etwa in Analogie zu einer "offenen Werkstatt" verstanden werden, die Raum für einfache Zugänge und Hilfsmittel zur gemeinschaftlichen Nutzung in strukturierter Form anbietet.

Die Frage nach den Alleinstellungsmerkmalen einzelner Disziplinen steht vor allem im Kontext globaler Digitalisierung und Standardisierung von Forschungsprozessen, der Einführung digitaler Curricula sowie aufgrund der zunehmenden Forderung nach Interoperabilität im Raum (Balsinger 2005). Ebenso besteht die Anforderung, durch die Integration „analoger“ und „digitaler“ Aspekte in die Forschung oder die berufliche Praxis neue, „hybride“ Praktiken zu etablieren (Zaagsma 2013). Innerhalb vornehmlich digital arbeitender Fachcommunities stehen außerdem zunehmend generisch einsetzbare Werkzeuge und Methoden im Spannungsfeld mit den speziell auf bestimmte Datendomänen zugeschnittenen Tools.

Angesichts der Entwicklung gemeinsamer Infrastrukturen haben gemeinschaftlich entwickelte und genutzte Tools mittlerweile eine rasante Dynamik erfahren. Viele Projekte ließen sich ohne kollaborative und transparente Arbeitsgestaltung kaum realisieren und werden daher von Virtuellen Forschungsumgebungen, Wikis, Messaging-Diensten, Cloud- und Ticketsystemen, Sync&Share Systemen, Versionierungstools und spezieller Software unterstützt. Das Community Building kann innerhalb dieser Infrastrukturen als Prozess verstanden werden, der zwischen generischer Offenheit und Zielgruppenspezifität anzusiedeln ist und Fächergrenzen überschreitet. Das Panel möchte in der Diskussion rezenten Problemlagen nachgehen und eruieren, wie sich Prozesse des Community Building im Digitalen an Orten des Lernens, der Forschung oder Vernetzung gestalten und welche Veränderungen, Herausforderungen sowie Chancen und Risiken dies für Universitäten wie GLAM-Einrichtungen birgt.

Impulsvorträge

Von geisteswissenschaftlichen Forschungsdaten zu *NFDI4Culture* als Open Community

Blickt man zurück auf die Entwicklung der letzten Jahre, so hat die Diskussion um den Begriff „Forschungsdaten“ das Community Building in den Geisteswissenschaften maßgeblich befeuert (Andorfer 2015). Die ausgeprägte Unschärfe dieses Begriffs stand einer von allen Communities akzeptierten Definition lange entgegen. Umfragen halfen einer Konkretisierung näherzukommen und legten Sichtweisen und Umgang mit Daten offen. Auf diese Weise hat man versucht, Kategorien zu bilden, um eine übergreifende Ordnung zu finden. Etwa zeitgleich wurden allgemeine und fachspezifische Policies und Empfehlungen zum Umgang mit Forschungsdaten für die Communities herausgegeben (z. B. DFG 2015ff.). Diese mündeten schließlich – getragen vom Aufbau

spezieller Repositorien und Datenzentren (<https://dhd-ag-datenzentren.github.io/>) – in den konkreten Bedarf einer Infrastruktur, welche eine langfristige Sicherung und nachhaltige Bereitstellung digitaler Forschungsdaten garantieren kann, wobei der Akzent auf einer fachübergreifenden Zusammenarbeit liegt.

Resultat dieser Bestrebung ist seit Herbst 2020 die Nationale Forschungsdateninfrastruktur (NFDI), darunter das Konsortium *NFDI4Culture* für das (im)materielle Kulturerbe (<https://nfdi4culture.de/>). Die Zielgruppen dieser Infrastruktur sind einerseits die Datenerzeuger und Vertreter verschiedener Wissensgebiete, andererseits die Datenanbieter (i.d.R. Infrastruktureinrichtungen, GLAM), deren Funktionsbereiche sich teilweise überlappen und für einen reibungslosen Datenfluss ineinander greifen. Entscheidend ist die Konsolidierung der Communities: Es gilt, die Interessen unterschiedlicher Teilgruppen zusammenzuführen und zielgruppenspezifische Angebote für qualitativ hochwertige Forschungsdaten zu entwickeln. Eine Verständigung zwischen den Disziplinen wird gefördert, indem Fachvertreter:innen in gemeinsamen Task Areas zusammenarbeiten (Altenhöner et al. 2020). Für eine solche Zusammenarbeit sind auch fachübergreifende Angebote gefragt, weshalb die NFDI gemeinsame Querschnittsthemen identifiziert hat, die über die Konsortien hinweg in (bislang) vier Sektionen („Gemeinsame Infrastruktur“, „Metadaten, Terminologien, Provenienz“, „Recht und Ethik“ sowie „Training und Ausbildung“) bearbeitet werden.

Kleine Tools und schwach strukturiertes Publizieren (Code and Data Literacy am Beispiel medienkulturwissenschaftlicher Lehre)

Data Literacy kann als Indikator dienen, wenn es um die Zugehörigkeit zu digital geprägten Communities geht, ebenso bietet sie Möglichkeiten, diese zu verbinden. Der Vortrag möchte den kritischen Umgang mit Code und Daten als „ability to collect, manage, evaluate, and apply data, in a critical manner“ (Ridsdale et al. 2015) in einem erweiterten Sinne anhand ausgewählter Beispiele verdeutlichen und für einen offenen Umgang mit Methoden, Tools sowie Lerneinheiten plädieren. Zunächst sollen die Vorteile von Open Educational Resources (OER) aufgezeigt werden. Etwas globaler formuliert soll damit auch für das zügige und eher schwach strukturierte Publizieren unfertiger, kürzerer und kleinerer Recherchen und Übungen durch Studierende argumentiert werden. Dies führt nicht nur über das gegenseitige Lesen, Kommentieren und Korrigieren im Seminar hinaus, sondern öffnet die Publikation von Unterrichtsmaterial auch für einen Austausch mit den Fachcommunities (Bsp. <https://zfmedienwissenschaft.de/online/open-media-studies-blog>). Ressourcen, Tutorials und Anleitungen auch außerhalb des angestammten Lehrbetriebs gibt es viele (u.a. <https://programming-historian.org/> oder <https://digital-history-berlin.github.io/Python-fuer-Historiker-innen/>). Ihre Auffindung und Nutzung gehört auch zur Data Literacy. Mit der Öffnung der Seminare, insbesondere durch die gezielte Ansprache von Infrastruktureinrichtungen/GLAM (Bsp.

DNBLab sowie lokale Angebote der Universitätsbibliotheken), können nicht nur Lehr- und Lerneinheiten, sondern auch die eingesetzten Tools kontextabhängig und datenspezifisch weiterentwickelt werden. So werden neben allgemein nützlichen Redaktionsabläufen auch eine Vielzahl von Herangehensweisen an unterschiedlichste Medieninhalte und Schnittstellen auf allen Ebenen wissenschaftlichen Arbeitens eingeübt, also Datenkompetenz wie Datenkritik praktisch erprobt. In der Diskussion verweisen wir auf Trainings- und Ausbildungsszenarien, die nicht nur von einer Einrichtung allein geleistet werden können, sondern sich entlang von Forschungsdateninfrastrukturen und dem Austausch von OER bewegen. Schwerpunkte sind medien- und kulturwissenschaftliche Kontexte und die Ausprägung bestimmter Datenbegriffe.

Humanities im Wandel – Neue Möglichkeiten durch Community Building

Das letzte Statement liefert unter Einbezug des genannten Praxisfalls eine zusammenfassende Definition des Begriffs „Community Building“, skizziert nochmals seinen praktischen Nutzen und legt offen, welche Kriterien und Aspekte einer Forschungsdateninfrastruktur dafür entscheidend sind.

Da im Digitalen viele Prozesse in den Geistes- und Kulturwissenschaften gemeinsam, also fachübergreifend, erfolgen, soll auch allgemein darüber reflektiert werden, auf welche Weise die Vernetzung von Personen und Einrichtungen in diesem Feld spezifisch verläuft. Einerseits wird diese unmittelbar durch die projektbedingte Entwicklung bestimmter Tools und Services für die User angeregt („If we build it, they will come“: Ramsay 2016). Andererseits tragen länger andauernde, übergreifende Prozesse wie die Entwicklung von Auszeichnungsformaten, Vokabularen, Ontologien und Metadatenprofilen, die Ausarbeitung von Daten- und Qualitätsmanagement-Strategien (etwa mit Datenmanagementplänen), die digitale Kuratierung oder Archivierung oder die Herstellung von Linked Open Data zum Community Building bei. Daneben ist eine Übertragbarkeit von Methoden in der Datenverarbeitung zu beobachten, und zwar nicht nur zwischen einzelnen Disziplinen innerhalb der Digital Humanities, sondern darüber hinaus zwischen diesen und der Informatik sowie anderen MINT-Fächern (Musikwissenschaft: Plaksin 2021).

Besonders relevant wird die Frage sein, ob es sich um eine interdisziplinäre oder vielmehr eine transdisziplinäre, sehr offene Form der Zusammenarbeit handelt (Balsiger 2005, 140ff. 148f. 166f. 174. 179, Jungert 2010). In einem Fall werden Wissen und Methoden aus dem anderen Wissensgebiet genutzt, ohne den Methoden- und Erkenntnisraum des eigenen Gebietes zu verlassen. Im anderen Fall können Wissen und Methoden der eigenen Disziplin einen Gegenstand der Nachbardisziplin erschließen, wobei der Methoden- und Erkenntnisraum überschritten wird. Quasi disziplin-unabhängig – wie in einer offenen Werkstatt – kann dann an Aspekten gearbeitet werden, die jede Disziplin für sich allein womöglich nicht behandeln würde (z. B. Mittelstraß 1992).

Leitfragen für die Diskussion

1. Wo finden sich Beispiele für generische und spezifische Bedarfe innerhalb der/einer Forschungsdateninfrastruktur?
2. Inwieweit werden konkrete Interaktions- und Partizipationsmöglichkeiten an *NFDI4Culture* als offene Werkstatt von den Zielgruppen bereits wahrgenommen?
3. Abseits unseres Beispiels: Welche konkreten Lösungen bieten Forschende, um ihre (datenorientierten und auf digitalen Tools beruhenden) Arbeitsweisen in den NFDI-Aufbauprozess zu integrieren und inwieweit ist es überhaupt möglich und sinnvoll?
4. Wie gelingt es, Forschungscommunities über ein Service-Portfolio dauerhaft in den NFDI-Entwicklungsprozess einzubinden?
5. Wie wirken sich – datenspezifisch betrachtet – die community-übergreifend genutzten Werkzeuge und Methoden auf das Community Building aus?
6. Wie sollen Forschungs- und Infrastrukturprojekte auf die zunehmende praxisorientierte Diversität reagieren?
7. Wie beeinflusst das Community Building den Transfer in die Öffentlichkeit und die Rezeption in anderen Bereichen der (Wissens-)Gesellschaft, die ebenfalls dem digitalen Wandel unterliegen (GLAM als „Knowledge Broker“, Simon 2018, 320)?

Bibliographie

- Altenhöner, Reinhard et al. 2020. *NFDI4Culture - Consortium for research data on material and immaterial cultural heritage. Research Ideas and Outcomes*. <https://doi.org/10.3897/rio.6.e57036> (zugegriffen: 3. Dezember 2022).
- Andorfer, Peter. 2015. "Forschungsdaten in den (digitalen) Geisteswissenschaften. Versuch einer Konkretisierung." *DARIAH-DE Working Papers*: 14. <http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2015-14.pdf> (zugegriffen: 03. Dezember 2022).
- Balsiger, Philipp W. 2005. *Transdisziplinarität: systematisch-vergleichende Untersuchung disziplinübergreifender Wissenschaftspraxis*. München: Wilhelm Fink Verlag.
- DFG 2015ff. *Fachspezifische Empfehlungen zum Umgang mit Forschungsdaten*. https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/empfehlungen/index.html (zugegriffen: 03. Dezember 2022).
- Jungert, Michael. 2010. "Was zwischen wem und warum eigentlich. Grundsätzliche Fragen der Interdisziplinarität." In *Interdisziplinarität. Theorie, Praxis, Probleme*. Darmstadt: WBG, 1-12.
- Mittelstraß, Jürgen. 1992. "Auf dem Weg zur Transdisziplinarität." *Gaia* 1: 250.
- Plaksin, Anna. 2021. "Modelle zur computergestützten Analyse von Überlieferungen der Mensuralmusik – Empirische Textforschung im Kontext phylogenetischer Verfahren." PhD diss. Univ. Münster und TU Darmstadt. <https://doi.org/10.26083/tuprints-00017211> (zugegriffen: 03. Dezember 2022).
- Ramsay, Stephen. 2016. "Who's In and Who's Out." In *Defining Digital Humanities. A Reader*, hg. von Melissa

Terras, Julianne Nyhan and Edward Vanhoutte, 239-242. London : Routledge.

Ridsdale, Chantel et al. 2015. *Strategies and Best Practices for Data Literacy Education, Knowledge Synthesis Report*. <http://hdl.handle.net/10222/64578> (zugegriffen: 3. Dezember 2022).

Simon, Holger. 2018. "Digitales Ökosystem - Eine Antwort auf die digitale Transformation in den Kulturinstitutionen am Beispiel der Museen", in: Kuroczy#ski, Piotr, Bell, Peter und Dieckmann, Lisa (Hrsg.). *Computing Art Reader: Einführung in die digitale Kunstgeschichte*, Heidelberg: arthistoricum.net. <https://books.ub.uni-heidelberg.de/arthistoricum/catalog/book/413>, 319-328 (zugegriffen: 3. Dezember 2022).

Zaagsma, Gerben. 2013. "On Digital History". *BMGN - Low Countries Historical Review* 128,4: 3-29.

Herausforderung, Lesson Learned oder Chance? Der Zusammenhang zwischen Kulturen des Scheiterns und Open-Bewegungen in den Digital Humanities

Wuttke, Ulrike

ulrike.wuttke@gmx.net

Fachhochschule Potsdam, Deutschland

Kampkaspar, Dario

dario.kampkaspar@tu-darmstadt.de

Universitäts- und Landesbibliothek, Technische Universität Darmstadt

Müller-Laackman, Jonas

jonas.mueller-laackman@sub.uni-hamburg.de

Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky

Gengnagel, Tessa

tessa.gengnagel@uni-koeln.de

Universität zu Köln

Lang, Sarah

sarah.lang@uni-graz.at

Karl-Franzens-Universität Graz

Karcher, Stefan

stefan.karcher@dfg.de

DFG

Schrade, Torsten

Torsten.Schrade@adwmainz.de

Akademie der Wissenschaften und der Literatur | Mainz

Hintergrund und Beschreibung des Themas

'Scheitern' und 'Misserfolg' sind von Haus aus negativ konnotierte Ausdrücke, die oft den Eindruck eines Scherbenhaufens oder zumindest von vergeudeter Zeit und verschwendetem Geld hervorrufen. Die wenigsten Vorhaben in den Digital Humanities (DH) enden vollständig ohne Ergebnisse. Dennoch sind Projekte, die eines von mehreren Deliverables nicht oder nicht vollständig erbringen konnten, in denen Tools nicht die gedachten Ergebnisse lieferten oder eine versprochene Entwicklung nicht vollständig erbracht werden konnte, durchaus häufiger anzutreffen. Während sich die meisten Akteur*innen der Digital Humanities jetzt – etwas verstohlen – an solche Fälle erinnern werden, spiegelt sich dieser Umstand kaum in (projektbezogenen) Konferenzbeiträgen, Veröffentlichungen oder Projektberichten wider.

So vielfältig wie die Ursachen für das Scheitern sind auch die Ursachen für die verhaltene – in den meisten Fällen eher abwesende – Kommunikation. Dass über das Scheitern und seine vielfältigen Ursachen aber so wenig gesprochen wird, ist ein nicht zu unterschätzendes Problem. Negative Ergebnisse – um gleich die Erkenntnis, dass ein bestimmter Weg nicht zum gewünschten Ziel geführt hat, von der vollständigen Ergebnislosigkeit abzugrenzen – sind für die Forschung relevant: sie führen zu alternativen Ansätzen und schließen unfruchtbare Wege aus der Suche aus.

Erst die offene Kommunikation über 'Sackgassen' verhindert, daß Andere die gleichen Pfade erneut beschreiten. Überspitzt ausgedrückt: Das Verstecken von sogenannten Misserfolgen beraubt die Forschung eines Ergebnisses und führt unweigerlich dazu, dass Fehler wiederholt werden – und damit weitere Projekte scheitern können. Nicht das einzelne negative Ergebnis kostet Geld, sondern die fehlende Fehlerkultur. Deshalb bedarf es unter anderem einer offenen Kommunikation über positive wie negative Projektergebnisse und nicht zuletzt auch eines guten Projektmanagements, das mögliche Fehlermodi rechtzeitig erkennt und diesen gegensteuert sowie aktuelle Diskussionen und relevante Veröffentlichungen auswertet und in das Projekt einbringt. Negative Ergebnisse müssen mehr Akzeptanz erfahren, um den Nimbus des (vielleicht gar persönlichen) Scheiterns (mit entsprechenden, befürchteten wie tatsächlichen, Konsequenzen für die eigene Zukunft) zu verlieren.

Motivation und Leitfragen bzw. Themencluster

Verständlicherweise ist in den Digital Humanities das Phänomen des Scheiterns nicht unbekannt. Problematisiert wurde das Scheitern von Projekten im DHd-Kontext zuletzt in einigen projektspezifischen sowie projektübergreifenden Kontexten (z. B. Frank 2022, Gengnagel 2021-2022, RaDiHum20 2022, Schuhmacher 2022, Zarei et al. 2022). Auch darüber hinaus gibt es neben dem wegweisenden Artikel von Dombrowski (2014) weitere interessante Beispiele des Umgangs mit diesem Diskurs, wie z. B. Drucker 2014, Graham 2019. Diese Beispiele nehmen jedoch nicht weg, dass der Diskurs über das Scheitern noch spärlich und angstbehaftet ist (siehe FuReSH 1-2 2022) und relativ wenig Eingang in die Publikationskultur findet, wobei Ausnahmen hier die Regel zu bestätigen scheinen, wie die Rezension von Dombrowski (2019) von *Failing Gloriously and Other Essays* (Graham 2019) zu denken gibt: "Graham acknowledges upfront the privilege that underpins his ability to talk so openly about failure. He's a white man with tenure, which counts for a lot [...]"

Es stellt sich die Frage, welche unterschiedlichen Fehlerkulturen aus den Ursprungsdomänen in den DH zusammenkommen und ob z. B. stärker technisch geprägte Felder (und damit auch Publikationsoutlets wie z. B. die *Proceedings der Computational Humanities Tagung*¹) von vornherein und aus wissenschaftlichen Notwendigkeiten eine andere Fehlerkultur haben. Während in diesen Bereichen das Scheitern bzw. der Umstand, unterschiedliche Lösungswege ausprobiert zu haben, von technischer Kompetenz zeugen und ihnen somit auch für Early Career Researcher ein Potential innezuwohnen scheint, die eigene Technik- und Methodenkompetenz unter Beweis zu stellen, scheint es in den sogenannten Buchwissenschaften selbst für etablierte Wissenschaftler*innen weniger Raum zu geben, um über Scheitern zu sprechen. Dabei gibt es mittlerweile sogar Open-Access-Journals, wie JOTE (Journal of Trial and Error)², die auch für Digital-Humanities-Forschung die Möglichkeit bieten, für alle Beteiligten nutzbringend ihre *Errores* zu publizieren.

Vor diesem Hintergrund will das Panel Bezüge zwischen Kulturen des Scheiterns und Open-Bewegungen in den Digital Humanities Praktiken sondieren und relevante Positionen kartieren. Das geplante Panel will nicht nur einfach die eben skizzierten Probleme beschreiben, sondern den Austausch über diese Themen speziell im Kontext der Digital Humanities durch einen strukturierten thematischen Impuls forcieren, bei dem insbesondere folgende Themenbereiche bzw. Leitfragen adressiert werden:

- Was bedeutet es eigentlich, dass ein Projekt gescheitert ist?
- Warum ist eine gute Fehlerkultur für Open Humanities wichtig bzw. ist die Offenlegung des Forschungsprozesses eine Gefahr für die Karriere?
- Ist in den technischeren Feldern der DH das Beschreiben von verschiedenen Lösungswegen inklusi-

sive Irrwegen bereits stärker etabliert als in den sogenannten Buchwissenschaften?

- Welche Rolle spielen wissenschaftssoziologische Faktoren, wie die Forschungshierarchie und Leistungsdruck, für die Möglichkeiten, um über Scheitern zu sprechen?
- Welche Verantwortung haben Forschungsförderer beim Etablieren einer besseren Fehlerkultur? Was können wir aus anderen Disziplinen lernen?
- Inwieweit ist Scheitern ein inhärenter Bestandteil des wissenschaftlichen Prozesses und inwieweit brauchen wir deshalb eine andere Fehlerkultur?
- Welche Chancen und Herausforderungen bestehen bezüglich der Verbesserung bestehender Kulturen und Strukturen?

Primäres Ziel des Panels ist neben der Herausarbeitung momentaner Schmerzpunkte die Sichtbarmachung der Vielschichtigkeit des Diskurses und die Entwicklung von multiperspektivischen Handlungsoptionen.

Ablauf und Organisation des Panels

Um eine lebendige Diskussion anzuregen, verzichtet das Panel auf die sonst üblichen Kurzreferate der Teilnehmer*innen. Nach einer kurzen Einführung in die Thematik des Panels durch die Moderator*innen werden die Panelist*innen in pointierten Statements ihren Bezug zum Panelthema vorstellen (15 Minuten). Dabei stellt jede Teilnehmer*in eine spezifische Perspektive vor. Dann soll ein multiperspektivischer Austausch über bestehende Kulturen des Scheiterns sowie die Frage nach Veränderungsbedarfen und möglichen Lösungsansätzen anhand der oben skizzierten und im Vorfeld des Panels (u. a. anhand der im Vorfeld eingehenden Stellungnahmen bzw. Problematisierungen, siehe unten) ggf. weiter zu konkretisierenden Leitfragen, die den Teilnehmer*innen des Panels im Vorfeld zur Verfügung gestellt werden, im Mittelpunkt stehen (45 Minuten). An diesem Punkt wird bereits frühzeitig der Diskurs in Richtung Publikum geöffnet und dieses interaktiv für direkte Erwidern in die Diskussion einbezogen. Die letzten 30 Minuten sind explizit für die Diskussion zwischen dem Panel und dem Publikum vorgesehen.

Während des gesamten Panels wird die Moderation auf eine sachliche und gewaltfreie Kommunikation achten. Das Panel wird unterstützt durch die Möglichkeit der Beteiligung via Twitter oder anderer Social-Media-Kanäle, vor, während und nach der Diskussion. Insbesondere soll auch im Vorfeld bzw. während des Panels eine niedrigschwellige Beteiligung ermöglicht werden, z. B. durch das Einbringen anonymierter kürzerer Stellungnahmen und Problematisierungen (ggf. mittels digitaler Feedbacklösungen), wobei sich die Organisator*innen und Moderator*innen die Freiheit der Auswahl erlauben. Auf diese Weise kann schon im Vorfeld des Panels die Diskussion gebündelt und um weitere Perspektiven erweitert werden. Es wird angestrebt, die Ergebnisse des Panels der Fachöffentlichkeit zur Verfügung zu stellen (als Blogbeitrag, White Paper, Thesenpapier, Fachartikel o. ä.).

Teilnehmende (Vorstellung der Panelist*innen und Moderation)

Teilnehmende:

Tessa Gengnagel (internationale Perspektive) ist als Postdoc am CCeH (Universität zu Köln) in der Geschäftsführung tätig. Für das Panel wird sie zu Failure als Topos einer konstituierenden Kraft in DH-Narrativen beitragen (siehe Gengnagel 2021-2022) und dazu als Folie eine wissenschaftssoziologische Perspektive kontrastieren, die intersektionale Kriterien berücksichtigt.

Sarah Lang (Perspektive AG Empowerment) ist Computational Humanist am Grazer Zentrum für Informationsmodellierung. Sie lotet im Rahmen des Panels wissenschaftstheoretische Aspekte des Umgangs mit Fehlern vor dem Hintergrund breiterer Fragen, wie z. B. unsichtbaren Machtstrukturen, Prekariat und Diskriminierung, aus.

Stefan Karcher (Perspektive Fördergeber) ist Referent bei der Deutschen Forschungsgemeinschaft (DFG) und dort unter anderem für den Bereich Digital Humanities zuständig. Im Panel wird er die Frage von für die Forschung zuträglichen Kulturen des Scheiterns aus der Perspektive eines Drittmittelgebers diskutieren.

Torsten Schrade (Perspektive Research Software Engineering und wissenschaftliche Infrastruktur) leitet die Digitale Akademie der Akademie der Wissenschaften und der Literatur | Mainz und ist Spokesperson von NF-DI4Culture. Zu den Leitsprüchen seiner Forschungs- und Entwicklungsaktivitäten zählen "Weniger schlecht programmieren" (Passig/Jander) und "Fail better" (Beckett).

Moderation:

Ulrike Wuttke (Fachhochschule Potsdam, Fachbereich Informationswissenschaften) lehrt und forscht am Fachbereich Informationswissenschaften der Fachhochschule Potsdam. Zu ihren Schwerpunkten gehören Open Science Advocacy und Training.

Dario Kampkaspar (Technische Universität Darmstadt, Universitäts- und Landesbibliothek) leitet das Zentrum für digitale Editionen und ist seit über 10 Jahren in digitalen Projekten sowohl öffentlicher wie privater Fördergeber in Deutschland und Österreich inhaltlich wie organisierend tätig.

view-failing-gloriously-and-other-essays-shawn-graham (zugegriffen 12. Juli 2022).

Drucker, Johanna. 2021. "Sustainability and Complexity: Knowledge and Authority in the Digital Humanities," Digital Scholarship in the Humanities 36, Nr. Supplement_2: ii86-94. <https://doi.org/10.1093/llc/fqab025>.

Frank, Markus. 2022. "Projektmanagement für die Digital Humanities (Workshop)." In DHd2022: Kulturen des digitalen Gedächtnisses: Konferenzabstracts, Universität Potsdam & Fachhochschule Potsdam, 07. bis 11. März 2022, hg. von Michaela Geierhos, Potsdam, 400-2. <https://doi.org/10.5281/zenodo.6304590>.

FuReSH I+II. 2022. "Veranstaltungsreihe: Kulturen des Scheiterns," Mai 2022. <https://blogs.hu-berlin.de/furesh/2022/05/19/veranstaltungsreihe-kulturen-des-scheiterns/> (zugegriffen 12. Juli 2022).

Gengnagel, Tessa. 2021-2022. "Vom Topos des Scheiterns als konstituierender Kraft: Ein Essay über Erkenntnisprozesse in den Digital Humanities." In Fabrikation von Erkenntnis – Experimente in den Digital Humanities, hg. von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis und Ulrike Wuttke, Wolfenbüttel, 2021-2022 (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5) https://doi.org/10.17175/SB005_011.

Graham, Shawn. Failing Gloriously and Other Essays, with a foreword by Eric Kansa and afterword by Neha Gupta. 2019. Grand Forks, ND, 2019. <https://doi.org/10.31356/dpb015>

RaDiHum20. 2022. "RaDiHum20 spricht mit Markus Frank über Projektmanagement," 20. April, 2022. <https://radium20.de/radium20-markus-frank-projektmanagement/> (zugegriffen 12. Juli 2022)

Schumacher, Mareike. 2022. "Wie Wölkchen im Morgenlicht". Zur automatisierten Metaphern-Erkennung und der Datenbank literarischer Raummetaphern laRa." In DHd2022: Kulturen des digitalen Gedächtnisses: Konferenzabstracts, Universität Potsdam & Fachhochschule Potsdam, 07. bis 11. März 2022, hg. von Michaela Geierhos, Potsdam, 232-6. <https://doi.org/10.5281/zenodo.6304590>

Living Handbook "Digitale Quellenkritik"

Deicke, Aline

aline.deicke@adwmainz.de

Akademie der Wissenschaften und der Literatur | Mainz / Philipps-Universität Marburg

Wachter, Christian

christian.wachter@uni-bielefeld.de
Universität Bielefeld

Feichtinger, Moritz

moritz.feichtinger@unibas.ch
Universität Basel

Fußnoten

1. <http://ceur-ws.org/Vol-2989/>
2. <https://archive.jtrialerror.com>

Bibliographie

Dombrowski, Quinn. 2014. "What Ever Happened to Project Bamboo?" Literary and Linguistic Computing 29, Nr. 3: 326-39. <https://doi.org/10.1093/llc/fqu026>.

Dombrowski, Quinn. 2019. "Book Review: 'Failing Gloriously and Other Essays' by Shawn Graham," Digital Humanities @ Stanford, 18. November 2019. <https://digitalhumanities.stanford.edu/book-re>

Lemaire, Marina

marina.lemaire@uni-trier.de
Universität Trier

Schmunk, Stefan

stefan.schmunk@h-da.de
Hochschule Darmstadt

Hall, Mark

mark.hall@open.ac.uk
Open University

Harvey, Francis

f_harvey@leibniz-ifl.de
Leibniz-Institut für Länderkunde

Durdağı, A. Nursen

ndurdagi@sakarya.edu.tr
Sakarya Üniversitesi

Geistes- und Kulturwissenschaften bauen in ihren Erkenntnisprozessen wesentlich auf der Befragung von Quellen unterschiedlichster Materialität und Medialität auf: So bezeichnet der Begriff sämtliche Objekte und Überreste, die zum Erkenntnisgewinnungsprozess über das Vergangene beitragen, z. B. Gemälde, Musiknotenblätter, Texte, Fotografien, Münzen, Inschriften, Kleidung oder andere Alltagsgegenstände. Quellen liefern aber keine Wahrheiten, sondern müssen gedeutet und in die Sprache der (historischen) Wissenschaften übersetzt werden. Zudem können Quellen subjektiv, fehlerhaft, verfälscht oder auch nur in Teilen erhalten sein. Um diesen Herausforderungen zu begegnen, hat sich in den historischen Wissenschaften die Methode der Quellenkritik etabliert (vgl. Koselleck 1977). Sie dient dazu, die Aussagekraft einer Quelle für ein gegebenes Forschungsvorhaben (insbesondere in Relation zu anderen Quellen) zu beurteilen und stellt damit letztlich die Grundlage zu ihrer Analyse dar. Hierfür werden sie z. B. beschrieben, indiziert, kontextualisiert, übersetzt und daran anschließend ausgewertet. Die Quellenkritik ist damit eine der Säulen des Forschens schlechthin, sowohl in den historischen Wissenschaften als auch darüber hinaus (vgl. z. B. Arnold 2001).

Die digitale Transformation verändert alle Bereiche der Gesellschaft – dies schließt die Wissenschaft insgesamt und die historischen Wissenschaften im Speziellen ein. Neben Quellen genuin digitaler Natur, sog. “born digital” Quellen wie z. B. Software, Websites, Social Media Beiträge und persönliche Textnachrichten, stehen “traditionelle” Quellen im zunehmenden Maße digitalisiert zur Verfügung. Diese digitalen Repräsentationen stellen die Quellenkritik allerdings vor neue Herausforderungen: Wer hat die Quelle wie digitalisiert und zu welchem Zweck? Welche Formate und Transformationsalgorithmen wurden verwendet? Wer hostet die digitale Quelle und gewährleistet die Langzeitverfügbarkeit sowie ihre Integrität? Wie wird die Quelle auffindbar für diejenigen, die sie in ihrer Forschung verwenden wollen? Zu all diesen Fragen müssen sich die Geistes- und Kulturwissenschaft-

ten verhalten und dabei sowohl philosophische Überlegungen (Was ist eigentlich ein digitales Objekt?), als auch Überlegungen zu Methodologie, Langzeitarchivierung, manueller / semi-automatischer / automatischer Erschließung, Forschungsethik und viele andere mehr berücksichtigen.

Die traditionelle Quellenkritik muss vor diesem Hintergrund um die Dimension einer digitalen Quellenkritik erweitert werden. Hierzu sind in jüngerer Zeit bereits einige Beiträge vorgelegt worden (vgl. z. B. Fickers 2020; Föhr 2017; Hering 2014; Pfanzer 2015), doch mit dem rasanten technischen Wandel und vor dem Hintergrund sich stark verändernder digitaler Infrastrukturen und Arbeitsprozesse in der Wissenschaft (z. B. im Zusammenhang mit der Nationalen Forschungsdateninfrastruktur) braucht es eine stete und kritische Begleitung des Themas, wie es nur ein kontinuierlich und kooperativ geführter wissenschaftlicher Diskurs gewährleisten kann. Dieser Aufgabe hat sich der Arbeitskreis Digitale Quellenkritik (der lose an den DHd-Verband und die Arbeitsgruppe digitale Geschichtswissenschaft des Historikerverbands angeschlossen ist) verschrieben. Die Gruppe setzt sich aus Vertreter*innen mit unterschiedlichen institutionellen Hintergründen (Forschungs- und Infrastruktureinrichtungen) und aus diversen geistes- und kulturwissenschaftlichen Disziplinen zusammen, um die verschiedenen Aspekte digitaler Quellenkritik möglichst vielseitig zu beleuchten. Der Arbeitskreis hat es sich zur Aufgabe gemacht, die Thematik aufzufächern, die verschiedenen theoretischen, methodischen und inhaltlichen Aspekte zu identifizieren und sie in einem living handbook zusammenzuführen. Das living handbook stellt dabei sowohl eine Kondensationsfläche für den status quo als auch eine Einführung in die Thematik und eine Diskussionsgrundlage dar.

Aktuell befinden sich bereits mehrere Kapitel des living handbooks im Publikationsprozess, andere Kapitel sind noch in der Aufarbeitung bzw. Planung. Um den derzeitigen Stand im Detail vorzustellen und im Rahmen von Diskussionen weitere Impulse aus der DH-Community einzuholen und in die Kapitel zu integrieren, erscheint ein Panel auf der DHd2023 als ideales Format. Entsprechend soll dort nachstehende Auswahl an Kapiteln kurz präsentiert und zur Diskussion gestellt werden. Sowohl das Projekt des living handbooks als auch die öffentliche Diskussion mit der DH-Community stehen damit ganz im Geiste des Tagungsmottos “Open Humanities – Open Culture”.

Beitrag 1: Offenes, community-getriebenes Publikationsformat

Aline Deicke

Das Handbuch ist ein Community-Projekt und per Definition nie abgeschlossen: Es hält sich offen für Korrekturen, Ergänzungen und Aktualisierungen. Die Prozesse der Arbeit mit Quellen, ihre kritische Reflexion und Interpretation und der Weg von einer Quelle zu einer wissenschaftlichen Aussage sollen damit transparent gemacht und aktualisiert werden. Das living handbook zur digita-

len Quellenkritik ist so ein Stück gelebte Open Culture der digitalen Geistes- und Kulturwissenschaften.

Anhand der Arbeiten und Debatten um das Handbuch lassen sich Aspekte diskutieren, die Open Science in den Digital Humanities allgemein betreffen: Wie lässt sich eine möglichst weite Partizipation von Stimmen und Perspektiven aus der Gesellschaft und verschiedenen Fachcommunities mit der Sicherung von Expertise vereinbaren? Welche Verfahren der redaktionellen Überarbeitung erfordert eine offene und stetige Erweiterung von Texten? Wie lassen sich Elemente des traditionellen Publikationsbetriebs (Reputationsmetriken, Reviewverfahren etc.) in einen community-getriebenen, kollaborativen Prozess überführen?

Beitrag 2: Theorie der digitalen Quellenkritik

Moritz Feichtinger

Das Projekt des living handbooks ist getrieben von der gemeinsamen Einsicht, dass Quellenkritik im digitalen Zeitalter um neue Herangehensweisen und Fragestellungen erweitert werden muss. Unklar bleibt jedoch, inwieweit dieser digitale Wandel die theoretischen und methodischen Grundlagen der Disziplinen erfassen wird oder sollte. Handelt es sich bei digital unterstützten Praktiken des Befragens, Analysierens und Interpretierens lediglich um den Einsatz eines modernisierten Werkzeugkastens, oder ändert sich die Gewinnung historischer Erkenntnis fundamental? Wenn der Umgang mit Quellen von der Suche über die Bearbeitung bis zur Interpretation in erheblichem Umfang von digitalen Methoden bestimmt ist, bedeutet dies letztlich eine Erweiterung oder Erneuerung der Grundkenntnisse und Grundwissenschaften historischen Forschens?

Der Beitrag fragt also nach den Erfordernissen digitaler Quellenkritik und deren Konsequenzen für die Methodik und Hermeneutik historischen Forschens. Debattiert werden soll, welche Zugänge anderer Disziplinen übernommen werden können und unter welchen Bedingungen dies geschehen kann (oder sollte). Zudem möchten wir zur Diskussion stellen, welche Gestalt und welches Ausmaß der Einfluss digitaler Quellenkritik auf die Hermeneutik haben kann (oder sollte).

Erleben wir einen Wandel oder eine Pluralisierung der Erkenntnisverfahren? Bedeutet dies eine Schärfung des Profils akademischer Forschung (und ihrer hergebrachten Hermeneutik) oder eine stärkere Annäherung an andere Disziplinen?

Beitrag 3: Algorithmenkritik und die Grenzen des Algorithmus

Mark Hall

Im digitalen Raum werden Quellen zwangsläufig algorithmisch verarbeitet: von der Suche relevanter Quellen über ihre Analyse bis hin zur Visualisierung der Analyseergebnisse – Algorithmen sind allgegenwärtig. Jeder die-

ser Algorithmen hat das Potenzial, Quellen und Ergebnisse zu verzerren respektive zu beeinflussen (Van Es 2018). Eine kritische Reflexion der im Forschungsprozess verwendeten Algorithmen ist daher integraler Bestandteil der digitalen Methoden- und Quellenkritik (Dobson 2015).

In der Informatik wird der Algorithmus generell als unabhängig von den bearbeiteten Daten gesehen. Diese künstliche Trennung ist aber für die Algorithmenkritik aus mehreren Gründen problematisch: Erstens lässt sich die Frage nach potenziellen Verzerrungen nur im Hinblick auf die spezifischen Eigenschaften der zu verarbeitenden Daten beantworten. Ein Algorithmus kann für das eine Quellenkorpus geeignet sein und den Aussagewert eines anderen verzerren. Zweitens sind bei vielen Methoden die Grenzen zwischen Daten, Modell und Algorithmus nicht trennscharf. So nutzt Maschinelles Lernen etwa mehrere Algorithmen, um zum einen aus Daten ein Modell zu trainieren und zum anderen, um das Modell auf die Quelle anzuwenden. Diese Aspekte können nicht unabhängig voneinander betrachtet werden.

Es ist daher notwendig, eine holistische Algorithmenkritik zu entwickeln, die den Algorithmus im Kontext der Daten und Modelle analysiert. Zu diskutierende Fragen sind unter anderem: Ist diese enge Verknüpfung von Algorithmus, Daten und Modellen wirklich notwendig? Was für Methoden der Algorithmenkritik kann man ohne eine detaillierte Analyse des Codes anwenden? Wo zieht man die Grenze zur Datenkritik?

Beitrag 4: Für die verräumlichten Geisteswissenschaften: Von Karten zu Standorten

Francis Harvey

In einer digitalen Quellenkritik (Fickers 2020; Pfanzer 2015) kann die Aufarbeitung von Orten und Standorten wichtige epistemologische und ontologische Hinweise für die Forschung geben (Bodenhamer et al 2010). Digitalisierte Karten und raumbezogene Daten stellen neue Herausforderungen dar. Forschende können zwar mit Einsichten aus der traditionellen Geographie und Kartographie die eigene Quellenkritik oft vertiefen, aber mit dem digitalen Wandel geht ein epistemologischer Wandel einher, der neue Herausforderungen an den verräumlichten Erkennungsgewinnungsprozess stellt (Capurro 2010). So kommen beispielsweise Fragen dazu auf, wie unterschiedlichen Zugängen zu Digitalisaten durch Informationsinfrastrukturen entgegengewirkt oder wie eine digitalisierte Kartenkritik in eine digitale Quellenkritik integriert werden kann.

Beitrag 5: Populäre Wissensproduktion – digitale Quellenlücke?

A. Nursen Durdağı

Die flächendeckende Nutzung von Computern in Wissenschaft und Gesellschaft, die rasante Entwicklung des Internets mit einer enormen Zunahme der Digitalisierung u.a. mit Hilfe von künstlicher Intelligenz führt zu einer unumkehrbaren Veränderung der menschlichen Informationsrezeption. Wissen wird nicht mehr ausschließlich über traditionelle Wege rezipiert, sondern oftmals über populär (wissenschaftlich)e Zugänge wie z. B. YouTube-Videos, Social Media Beiträge etc. Je mehr Informationen vorliegen und generiert werden, umso schwieriger wird es zu identifizieren, welche Beiträge wissenschaftlichen, peer-geprüften Forschungsprozessen entstammen. Insbesondere geht oft verloren, wie und von wem Informationen zu Sachverhalten erstellt, kommuniziert und verändert wurden. Für eine digitale Quellenkritik und -analyse bedeutsam sind u.a. soziale und geografische Unterschiede in der Verfügbarkeit von Ressourcen, in der Zugänglichkeit von Dokumenten angesichts (staatlicher) Geschichtspolitik und Zensur, eine sorgfältige Dokumentation der Datenhistorie sowie die Gewährleistung einer langfristigen Auffindbarkeit. Dennoch stellt sich die Frage, ob und in welchem Umfang es in Zukunft für Institutionen, Gremien oder auch Einzelakteur*innen überhaupt möglich sein wird, die Genese von Informationen umfassend zu recherchieren und nachzuvollziehen, und welchen Einfluss diese Unsicherheiten auf Prozesse der Wissensproduktion haben werden.

Bibliographie

Arnold, Klaus. 2001. "Der wissenschaftliche Umgang mit Quellen." In *Geschichte. Ein Grundkurs*, hg. von Hans-Jürgen Goertz, 2. Auflage, 42-58. Hamburg: Rowohlt.

Bodenhamer, David, John Corrigan und Trevor M. Harris (Hg.). 2010. *The Spatial Humanities*. Bloomington: Indiana University Press.

Capurro, Rafael. 2010. "Digital Hermeneutics—An Outline." *AI & Society* 35, Nr. 1: 35-42. <https://doi.org/10.1007/s00146-009-0255-9> (zugegriffen: 02. August 2022).

Dobson, James E. 2015. "Can an algorithm be disturbed? Machine learning, intrinsic criticism, and the Digital Humanities." *College Literature* 42, Nr. 4: 543-564. <https://doi.org/10.1353/lit.2015.0037> (zugegriffen: 02. August 2022).

Es, Karin van, Maranke Wieringa und Mirko Tobias Schäfer. 2018. "Tool Criticism: From Digital Methods to Digital Methodology." In *Proceedings of the 2nd International Conference on Web Studies*, hg. von Everardo Reyes, Mark Bernstein, Giancarlo Ruffo und Imad Saleh, 24-27. WS.2 2018. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3240431.3240436> (zugegriffen: 02. August 2022).

Fickers, Andreas. 2010. "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?" *Zeithistorische Forschungen – Studies in Contemporary History* 17, Nr. 1: 157-68. <https://doi.org/10.14765/zzf.dok-1765> (zugegriffen: 02. August 2022).

Föhr, Pascal. 2017. *Historische Quellenkritik im Digitalen Zeitalter*. Thesis, University of Basel. <https://doi.org/>

[info:doi/10.5451/unibas-006805169](https://doi.org/10.5451/unibas-006805169) (zugegriffen: 02. August 2022).

Hering, Katharina. 2014. "Provenance Meets Source Criticism." *Journal of Digital Humanities* 3, Nr. 2. <http://journalofdigitalhumanities.org/3-2/provenance-meets-source-criticism/> (zugegriffen: 02. August 2022).

Koselleck, Reinhart. 1977. "Standortbindung und Zeitlichkeit. Ein Beitrag zur historiographischen Erschließung der geschichtlichen Welt." In *Objektivität und Parteilichkeit*, hg. von Reinhart Koselleck, Wolfgang Mommsen, und Jörn Rüsen. München: Deutscher Taschenbuch-Verlag.

Knowles, Anne. 2014. "The Contested Nature of Historical GIS." *International Journal of Geographical Information* 28, Nr. 1: 206-211. <https://doi.org/10.1080/13658816.2013.850696> (zugegriffen: 02. August 2022).

Pfanzelter, Eva. 2015. "Die historische Quellenkritik und das Digitale." *Archiv und Wirtschaft. Zeitschrift für das Archivwesen der Wirtschaft* 1: 5-19. <http://diglib.uibk.ac.at/ulbtirolfodok/866898> (zugegriffen: 02. August 2022).

Open DH? Mapping Blind Spots

Gengnagel, Tessa

tessa.gengnagel@uni-koeln.de
Universität zu Köln, Deutschland

Lang, Sarah

sarah.lang@uni-graz.at
Universität Graz, Österreich

Probst, Nora

nora.probst@uni-koeln.de
Universität zu Köln, Deutschland

Gerber, Anja

anja.gerber@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Dang, Sarah-Mai

sarah-mai.dang@uni-marburg.de
Philipps-Universität Marburg, Deutschland

Duan, Tinghui

tinghui.duan@uni-jena.de
Universität Trier

Grallert, Till

till.grallert@fu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Keck, Jana

keck@ghi-dc.org

Deutsches Historisches Institut Washington (GHI), USA

Nyhan, Julianne

julianne.nyhan@tu-darmstadt.de

Technische Universität Darmstadt, Deutschland

Harkening back to the “big tent” metaphor (e.g. Terras 2013) which characterized debates about the inclusivity of the field ten years ago, the topic of ‘openness’ in the conference theme invites associations of ‘blue skies’, endless horizons, and the sense that everything is possible – in terms of participation, dissemination and objects of observation. This notion is complicated by several issues that discourses of cultural criticism have identified in the Digital Humanities in recent years (although they are not exclusive to the field): Among them monolingualism (Fiormonte 2021), a heritage of colonialism (Risam 2019) and gender imbalance (Gao et al. 2022, 330), to name but a few.

It is as of yet unclear whether and how the aspirations of the field can be reconciled with the realities of its practices. Supposing, however, that there is something to be learned from shifting the gaze towards the “borderlands” (Earhart 2018) of the field’s perceived center of activity – especially in a continental European context –, it would appear paramount to interrogate the theme of the conference and probe the boundaries of its ‘openness’ against this backdrop of socio-economic, political and infrastructural inequalities.

In order to remedy the *invisibility* of feminist, postcolonial and multilingual approaches or indeed their marginalization as topics with niche interest only intended for those directly affected by them, the panel will consist of six panelists that will each bring a different thematic focus to bear:

- Dr. Sarah-Mai Dang, Philipps University Marburg | data feminism
- Tinghui Duan, University of Trier | multilinguality
- Dr. Till Grallert, Humboldt University of Berlin | decolonization
- Jana Keck, German Historical Institute Washington (GHI) | transnationality
- Prof. Dr. Julianne Nyhan, Technical University of Darmstadt | hidden history
- Dr. Antonio Rojas Castro, Berlin-Brandenburg Academy of Sciences and Humanities | cultural context

All of these statements will have their own unique perspective but they will also intersect in many ways, which is why the panel chose a broad view on potential borders and blind spots: Sarah-Mai Dang will address film-historical desiderata while Julianne Nyhan’s statement on the history of DH will also take gender aspects into account. Antonio Rojas Castro’s presentation of the German-Cuban collaboration in the *Proyecto Humboldt Digital* project will argue for the necessity of decolonial and postcolonial considerations. Till Grallert’s statement, which will focus on the neo-colonial invisibility of the cultural heritage of large parts of the societies of the Global South, will also provide input on the challenges for DH of un-

der-resourced languages. This in turn connects to Tinghui Duan who will address challenges of multilinguality in the Global East. Jana Keck will cover perspectives of the Global North by presenting the transnational forum of the DH working group of the German Historical Institutes that engage, for example, in transnational discussions of data feminism which leads us back to Sarah-Mai Dang’s statement.

Each panel member will give a short statement (5 minutes each, 30 minutes in total). These will be interwoven with the first round of discussion among the panelists (30 minutes). Due to the broad range of topics, the expert statements as well as this first round of discussion will address a shared set of key points to anchor and focus the conversation. After this initial hour, the discussion will be open to questions from audience members (30 minutes), making for a total length of 90 minutes.

The panel will be held in English due to it being an international panel with an international topic. This does not mean, however, that the panel will be oblivious to the cultural academic context in which it will take place. Moderation will make the conscious effort of connecting the panel discussion with discourses of the German-speaking DH community and situating the outcome of the panel both within and beyond the conference setting of the DHd.

It is important to note that the panel does not presume to speak for marginalized groups whom it does not represent. For this reason, it primarily aims to provide a forum for self-reflection, as a first step towards a dialogue that should be extended in the future and should, as its long-term goal, seek to amplify voices that would otherwise not be heard.

Details of the panelist statements which will initiate the discussion are as follows (in alphabetical order):

Sarah-Mai Dang, New Insights and Old Blind Spots: Visualizing Film Historical Research Data

With the increasing production and use of data in the wake of digitalization, the goal of feminist film historians to increase the visibility of women’s work in film history has taken on a new urgency (Dang 2020). Through the production, processing and dissemination of data, blind spots can be maintained or amplified, but also minimized (Wreyford/Cobb 2017; D’Ignazio/Klein 2020). In her presentation, Sarah-Mai Dang will show how digital data visualization can help us to open up research on women in early cinema and thus make their impact in film history more visible. By presenting a case study on the Women Film Pioneers Explorer (<https://www.online.uni-marburg.de/women-film-pioneers-explorer>, Dickel et al. 2021), she seeks to discuss how visualized visions about the past can be defined as situated knowledges.

Tinghui Duan, Multilinguality as Challenge

Multilinguality in DH can refer to the inclusion of multilingual research objects in digital resources. It also refers to the need for multilingual academic exchange (which is often limited to English), especially since many terms are difficult to translate. Both aspects come with major challenges (cf. Fiormonte 2021). Tinghui Duan will address these challenges by using the example of the Romantic Period Poetry Archive (<https://t.co/QdLYAaWiCZ>), an "open access digital platform of global Romantic-period poetry". Since the project itself notes that its "selection process is largely Western/Northern-centric" (e.g. Görner 2021, Matuschek 2021), it is not surprising that not a single Chinese romantic writer was registered originally (<https://t.co/o8MkasqQ0A>). While three Chinese writers have now been admitted at the suggestion of Tinghui Duan (for romanticism in Asia cf. Long et al. 2018, Rabut 2014), he will address possibilities of acknowledging languages more equally.

Till Grallert, Neo-Colonial Layers of Invisibility in a Digitised World: Arabic Cultural Production and the Affordances of the Digital

Against the backdrop of his work on Arabic periodicals from the late nineteenth-century Eastern Mediterranean, Till Grallert will challenge the equation of "digitisation = access" by outlining the layers of inaccessibility inherent to existing digitisation efforts and infrastructures concentrated, for a large part, in the Global North (Grallert 2022, Risam 2018, 2019). These are a) a knowledge gap regarding the material artefacts themselves; b) a digitisation bias rooted in collection and survival biases and the direct costs of digitisation; c) the hegemony of socio-technical infrastructures rooted in Anglo-American neoliberal capitalism and the Western cultural canon (cf. Piron 2018); d) insufficient computational tools for layout and optical character recognition (cf. Wrisley et al. forthcoming); and e) insufficient access to basic digital infrastructures and utilities (cf. Aiyegbusi 2019).

Jana Keck, The Working Group Digital Humanities of the Max Weber Foundation and its Role as a Transnational Forum for German DH Scholars

The Working Group Digital Humanities of the Max Weber Foundation brings together German DH scholars

from 11 institutes worldwide with diverse backgrounds in the humanities (cf. Keck/Rohden 2021). The group provides not only a platform to discuss DH methods, theories and projects, but it also allows for the reflection on how the respective cultural, political, social, or economic conditions of the different locations shape the everyday experiences of DH research. Discussions in the group about upcoming trends and topics in DH have revealed a unanimous view: We need more ethics in DH! Jana Keck's statement will elaborate on the group's findings that it is not so much the debate about technological innovations or newest tools that brings DH-scholars together worldwide, but rather discussions on ethical frameworks in DH (Proferes 2020).

Julianne Nyhan, On Hidden and Devalued Labour in the Incunabular Digital Humanities: The Index Thomisticus Project c. 1954–67

Julianne Nyhan will seek to 'represence' the role of gender and invisible labour in the incunabular DH. Between 1954–1967, the Index Thomisticus project, an influential project in the early years of the field of Humanities Computing, had a workforce that numbered, at its peak, about 65 individuals. Yet many of those who worked in this project, especially the young women handling the project data, remain absent from histories of the development of DH. As her forthcoming book (Nyhan 2023) explores, these labour absences were neither inconsequential nor thoughtless, but can be understood to distill processes of exclusion/inclusion and expressions of social and epistemological hierarchy that would shape not only the emerging field of humanities computing, and in due course DH, but aspects of the wider development of computing too.

Antonio Rojas Castro, Diversifying the User Experience in Digital Editions

User experience research and design is very common in libraries, archives, and universities to assess the ease with which the community uses informational resources (Azadbakht, Blair, and Jones 2017; Seale, Hicks, and Nicholson 2022). In the field of DH, researchers have recognized the important role of the graphical user interface in digital editions (Bleier et al. 2018) and how software and tools are shaping the presentation of texts (Alvite-Díez and Rojas Castro 2022). This statement aims to take a critical approach to user experience and to interrogate how we can diversify the targeted users of Proyecto Humboldt Digital (<https://habanaberlin.hypotheses.org/>) to broaden the access and usage of our digital editions.

Bibliographie

- Aiyegbusi, Babalola Titilola. 2019. "Decolonizing Digital Humanities: Africa in Perspective." In *Bodies of Information: Intersectional Feminism and Digital Humanities*, ed. by Elizabeth Losh and Jacqueline Wernimont, 434–46. Debates in the Digital Humanities 4. Minneapolis: University of Minnesota Press. <https://doi.org/10.5749/j.ctv9hj9r9.26>.
- Alvite-Díez, M. Luisa, and Antonio Rojas Castro. 2022. "Ediciones digitales académicas: concepto, estándares de calidad y herramientas de publicación." *El profesional de la Información* 31 (2). <https://doi.org/10.3145/epi.2022.mar.16>.
- Azadbakht, Elena, John Blair, and Lisa Jones. 2017. "Everyone's Invited: A Website Usability Study Involving Multiple Library Stakeholders." *Information Technology and Libraries* 36 (4): 34–45. <https://doi.org/10.6017/ital.v36i4.9959>.
- Bleier, Roman, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider, eds. 2018. *Digital Scholarly Editions as Interfaces*. Schriften des Instituts für Dokumentologie und Editorik 12. Norderstedt: Books on Demand.
- Dang, Sarah-Mai. 2020. "Unknowable Facts and Digital Databases: Reflections on the Women Film Pioneers Project and Women in Film History." *Digital Humanities Quarterly* 14 (4). <http://www.digitalhumanities.org/dhq/vol/14/4/000528/000528.html>.
- Dickel, Henri, Matija Miskovic, Kharazm Noori, Christian Schmidt, Atefeh Soltanifard, Sarah-Mai Dang, and Thorsten Thormählen. 2021–. "Women Film Pioneers Explorer." <https://www.online.uni-marburg.de/women-film-pioneers-explorer/>.
- D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge, MA: The MIT Press.
- Earhart, Amy E. 2018. "Digital Humanities within a Global Context: Creating Borderlands of Localized Expression." *Fudan Journal of the Humanities and Social Sciences* 11: 357–369. <https://doi.org/10.1007/s40647-018-0224-0>.
- Fiormonte, Domenico. 2021. "Taxation against Overrepresentation? The Consequences of Monolingualism for Digital Humanities." In *Alternative Historiographies of the Digital Humanities*, ed. by Dorothy Kim and Adeline Koh, 333–376. Santa Barbara: Punctum Books. <https://www.jstor.org/stable/j.ctv1r7878x.13>.
- Gao, Jin, Julianne Nyhan, Oliver Duke-Williams, and Simon Mahony. 2022. "Gender Influences in Digital Humanities Co-Authorship Networks." *Journal of Documentation* 78 (7): 327–350. <https://doi.org/10.1108/JD-11-2021-0221>.
- Görner, Rüdiger. 2021. *Romantik: Ein europäisches Ereignis*. Ditzingen: Reclam.
- Grallert, Till. 2022. "Open Arabic Periodical Editions: A Framework for Bootstrapped Scholarly Editions Outside the Global North." In "Minimal Computing," ed. by Roopika Risam and Alex Gil, special issue, *Digital Humanities Quarterly* 16 (2). <http://digitalhumanities.org/dhq/vol/16/2/000593/000593.html>.
- Keck, Jana, and Jan Rohden. 2021. "Virtuelle Reise: Digital Humanities in der Max Weber Stiftung." In *vDHD-Konferenz Abstracts 2021*. <https://doi.org/10.5281/zenodo.5850834>.
- Long, Hoyt, Anatoly Detwyler, and Yuancheng Zhu. 2018. "Self-Repetition and East Asian Literary Modernity, 1900–1930." *Journal of Cultural Analytics* 2 (2). <https://doi.org/10.22148/16.022>.
- Nyhan, Julianne. 2023. *Hidden and Devalued Feminized Labour in the Digital Humanities: On the Index Thomisticus Project 1965–67*. London: Routledge.
- Matuschek, Stefan. 2021. *Der gedichtete Himmel: Eine Geschichte der Romantik*. 1st edition. München: C.H. Beck.
- Piron, Florence. 2018. "Postcolonial Open Access." In *Open Divide: Critical Studies in Open Access*, ed. by Ulrich Herb and Joachim Schöpfel. Sacramento: Litwin Books. <http://hdl.handle.net/20.500.11794/16178>.
- Proferes, Nicholas. 2020. *What Ethics can Offer the Digital Humanities and What the Digital Humanities can Offer Ethics*. London: Routledge.
- Rabut, Isabelle. 2014. "Chinese Romanticism: The Acculturation of a Western Notion". In *Modern China and the West*, ed. by Peng Hsiao-yen and Isabelle Rabut, 201–223. Leiden/Boston: Brill. https://doi.org/10.1163/9789004270220_009.
- Risam, Roopika. 2018. "Decolonizing the Digital Humanities in Theory And Practice." In *The Routledge Companion to Media Studies and Digital Humanities*, ed. by Jentery Sayers, 78–86. London: Routledge.
- Risam, Roopika. 2019. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston, IL: Northwestern University Press. <https://doi.org/10.2307/j.ctv7tq4hg>.
- Seale, Maura, Alison Hicks, and Karen P. Nicholson. 2022. "Toward a Critical Turn in Library UX." *College & Research Libraries* 83 (1): 6–24. <https://doi.org/10.5860/crl.83.1.6>.
- Terras, Melissa. 2013. "Peering Inside the Big Tent." In *Digital Humanities: A Reader*, ed. by Melissa Terras, Julianne Nyhan, and Edward Vanhoutte, 263–270. Farnham, Surrey: Ashgate.
- Wreyford, Natalie, and Shelley Cobb. 2017. "Data and Responsibility: Toward a Feminist Methodology for Producing Historical Data on Women in the Contemporary UK Film Industry." *Feminist Media Histories* 3 (3): 107–132. <https://doi.org/10.1525/fmh.2017.3.3.107>.
- Wrisley, David Joseph, Masoud Ghorbaninejad, and Nathan P. Gibson. Forthcoming. "RTL." In *Debates in the Digital Humanities 2022*. Minneapolis: University of Minnesota Press.

Open Humanities in
der Filmwissenschaft –
zwischen Wunsch und
Wirklichkeit

Howanitz, Gernot

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

Dang, Sarah-Mai

sarah-mai.dang@staff.uni-marburg.de
Universität Marburg

Diecke, Josephine

diecke@staff.uni-marburg.de
Universität Marburg

Ewerth, Ralph

Ralph.Ewerth@tib.eu
TIB / Universität Hannover

Lameris, Bregt

bregt.lameris@fiwi.uzh.ch
Open University of the Netherlands

Scherer, Thomas

scherer.thomas@fu-berlin.de
FU Berlin

Vukovic, Teodora

teodora.vukovic2@uzh.ch
Universität Zürich

Baresch, Ariadne

baresch@uni-trier.de
Universität Trier

Motivation

Das von der AG Film & Video organisierte Panel setzt sich inspiriert vom Tagungsmotto „Open Humanities, Open Culture“ mit Fragen der Offenheit in der Filmwissenschaft auseinander. In den letzten Jahren wurden verschiedene Aspekte der Open Humanities im Kontext der Filmwissenschaft umrissen, und zwar in Einzelbetrachtungen, die zunächst den Umweg der Medienwissenschaften nehmen (Sondervan 2018; Hirsbrunner 2019). Darüber hinaus gab es Arbeiten zum Potential offener Forschungsdaten für filmwissenschaftliche Fragestellungen (Heftberger et al. 2020), zum Forschungsdatenmanagement in der Filmwissenschaft (Dang 2020) sowie zur Verfügbarmachung digitaler Filme durch das Bundesarchiv (Heftberger 2020). Auch die Frage der Analysevokabulare wurde aufgeworfen (Bakels et al. 2020), ebenso jene des Open Access (Kolleg-Forschungsgruppe *Cinepoetics*, FU Berlin; GfM-AG Open Media Studies: <https://mediastudies.hypotheses.org/2633>)

Das Panel möchte diese Einzelbeobachtungen zusammenführen und gemeinsam mit weiteren Fragestellungen diskutieren. Dabei erweist sich die Frage nach Offenheit für die Filmwissenschaft als besonders virulent. Ein zentrales Thema ist die Wiederverwertbarkeit von Software: Digitale Tools für Bewegtbilder sind komplex und aufwändig, der erhöhte Entwicklungsaufwand ‚rechnet‘

sich erst bei intensiver Nutzung. Forschungsdaten können häufig nicht zur Verfügung gestellt werden, aus urheberrechtlichen Gründen – die milliardenschwere Filmindustrie wirkt hier entgegen – ebenso wie aus technischen – Filme brauchen im Vergleich zu anderen Medien ein Vielfaches an Speicherplatz – und organisatorischen – es fehlen etablierte Annotierungsstandards.

Das Panel nimmt verschiedene Herausforderungen in den Blick: technische Infrastruktur und Standardisierungsbestrebungen, Lehre, Forschungsdatenmanagement und Citizen Humanities. Wir sind überzeugt davon, dass die Filmwissenschaft durch das Angehen ihrer spezifischen Probleme wesentliche Impulse auch für andere Geisteswissenschaften setzen und die Open Humanities entsprechend weiterentwickeln kann.

Organisation des Panels

Wie schon bei der DHd 2022 verzichten wir auf Kurzreferate der Teilnehmer:innen. Um eine lebendige Diskussion zu erleichtern, setzen wir auf pointierte Eröffnungsstatements. Dabei bringt jede:r Teilnehmer:in eine spezifische Perspektive und eigene Forschungsfragen in die Diskussion ein, die im folgenden Abschnitt kurz umrissen werden. Die Fokussierung auf das Gespräch soll helfen, das Publikum verstärkt anzusprechen.

Der Zeitplan sieht eine zehnminütige Einführung in das Thema durch den Moderator vor, die von sechs dreiminütigen Kurzstatements der Teilnehmer:innen abgerundet werden. Es folgt ein offenes Panelgespräch entlang vorher verschickter Leitfragen (30 Minuten), bei dem das Publikum natürlich auch gerne eingreifen darf. Die letzten 30 Minuten sind für Fragen aus und Diskussion mit dem Publikum vorgesehen; dabei möchten wir auch mit digitalen Feedbacklösungen, beispielsweise Umfragetools, experimentieren, die das Eintreten in einen Dialog möglichst niederschwellig gestalten sollen.

Spezifische Perspektiven

Ralph Ewerth (Hannover) wirkt bei der Entwicklung einer offenen Forschungsinfrastruktur zur Film- und Videoanalyse mit (im DFG-geförderten Projekt TIB-AV-A, <https://gepris.dfg.de/gepris/projekt/442397862>). Dabei ergeben sich zahlreiche Herausforderungen: die freie Verfügbarkeit der Infrastruktur über das Web, der große Datenumfang der Videos, die Usability der Benutzerschnittstelle, die Auswahl der anzubietenden Funktionen inkl. Datenimport/-export, die Realisierung als Open-Source-Projekt, die Erweiterbarkeit durch Plugins, die Nachhaltigkeit der Lösung, und nicht zuletzt die Bewerbung in den verschiedenen Communities. Daher benötigt die Entwicklung einer solchen Lösung Expertise in verschiedenen Gebieten, u.a. im Bereich Software-Entwicklung (Web-Interface, Backend-Technologien inkl. Nutzung von GPU-Prozessoren, Hosting der Daten), fundierte Kenntnisse des aktuellen Forschungsstands in der Informatik (Computer Vision, natürliche Sprachverarbeitung, Mustererkennung, maschinelles Lernen, Informationsvisualisierung) und ein interdisziplinäres Verständnis für die Anforderungen, die der Arbeitsweise der Film-

wissenschaft entsprechen. Hieraus ergibt sich die Notwendigkeit, dass die Zielgruppen von vornherein in die Entwicklung der Infrastruktur mit einbezogen werden. Im Panel sollen die Herausforderungen aus der Open-Humanities-Perspektive der Filmwissenschaft diskutiert werden.

Sarah-Mai Dang (Marburg) diskutiert Forschungsdatenmanagement im Bereich des Filmkulturerbes, das intellektuelle Konventionen und institutionelle Rahmenbedingungen widerspiegelt, denen spezifische Vorstellungen von Film, Kanon und Autorschaft eingeschrieben sind. Möchten Wissenschaftler:innen mit filmhistorischen Datenbanken arbeiten, ist zunächst zu verstehen, mit welchen Daten sie es eigentlich zu tun haben. Auf welchen Quellen basieren die Daten? Nach welchen Kriterien wurden sie generiert? Von wem und zu welchem Zweck?

In einem Vergleich zweier Beispiele aus der Arbeit der BMBF-Forschungsgruppe „Datenvisualisierungen in der digitalen Filmgeschichtsschreibung am Beispiel der Forschung zu Frauen im Frühen Kino“ (DAVIF) sollen diese Fragen näher erläutert werden. Konkret geht es um die filmwerksbezogenen Daten des DFF – Deutsches Institut & Filmmuseum sowie die biographischen Daten des an der Columbia University angesiedelten Women Film Pioneers Project (WFPP), einer kollaborativen Onlineplattform, die mehr als dreihundert Profile aus der Stummfilmzeit versammelt. Während das Forschungsdatenmanagement des DFF von kuratorischen Überlegungen im Sinne einer interoperablen Nachnutzung geleitet wird und gemäß eines standardisierten Verfahrens erfolgt (EN 15907), ist die Arbeit des WFPP von einer Vielzahl an Forschungsinteressen im Sinne einer pluralen Filmgeschichtsschreibung bestimmt. Die Konsequenzen eines solchen ‚offenen‘ FDM für unsere Forschung und die Implikationen standardisierter Taxonomien sollen im Panel diskutiert werden.

Josephine Diecke (Marburg) und *Thomas Scherer* (Berlin) erachten offene Filmanalyse vokabulare zur Videoannotation in Forschung und Lehre als einen neuralgischen Punkt für digitale Open-Humanities-Bestrebungen in der Filmwissenschaft. Die Suche nach geeigneten Tools, Annotationsvokabularen und Austauschformaten ist zentral an eine Kultur des lebendigen und aktiven Methodenaustauschs gebunden. Eine kurze Gegenüberstellung aktueller Ansätze, wie dem interaktiven Thesaurus der *VIAN-Web-App* (<https://www.vian.app/keywords>), der linked-open-data *AdA-Filmmontologie* (Bakels et al 2020) und der Erprobung und Kritik der computergestützten Filmanalyse im Kontext des *Digital Cinema Hubs* (Diecke 2022) soll aufzeigen, welche Praktiken es im Forschungsfeld zu offenen, modularen und skalierbaren Filmanalyse vokabularen gibt, die als Fundament für die Vermittlung von Filmanalysekompetenzen und die wechselseitige (Nach-)Nutzung von Forschungsdaten fungieren können. Diecke und Scherer leisten damit einen Beitrag zu den zuletzt im Workshop der AG ‚Film und Video‘ debattierten engeren und früheren Rückkopplungsschleifen zwischen Entwicklungen in den Digital Humanities und filmwissenschaftlicher Lehr- und Forschungspraxis, die dem Vorbehalt des Überstülpens fachfremder Systematisierungszwänge entgegenwirken. Ein Schritt in diese Richtung ist die Entwicklung von Lehrmodulen zur Einführung in die annotationsba-

sierte Filmanalyse sowie die Konzeption und Verbreitung übergreifender Analysestandards, wie sie Vukovic und Baresch im Folgenden vorschlagen. Im Austausch mit den Panel-Teilnehmenden und dem Publikum sollen die vorgestellten Perspektiven und Bedürfnisse gegengeprüft werden, um diese in weiterführende Bestrebungen zur Anwendung und Verbesserung der digitalen Filmanalyse einbinden zu können.

Bregt Lameris (Open University of The Netherlands) untersucht, wie das im Rahmen von Barbara Flueckigers FilmColors-Projekt entwickelte Annotationstool VIAN als Plattformen für Citizen-Humanities-Projekte dienen kann. Die kontinuierliche Weiterentwicklung digitaler Werkzeuge ermöglicht es, Bürger:innen in geisteswissenschaftliche Forschungsprojekte einzubeziehen, oft in Zusammenarbeit mit Kulturerbeinstitutionen. Dadurch wird das Wissen über geisteswissenschaftliche Forschung in der Bevölkerung gefördert; gleichzeitig erheben Citizen Humanists wichtige Daten – etwa durch Taggen – und produzieren Wissen auf neue, ungewohnte Weise.

Die Implementierung von VIAN als Citizen-Humanities-Tool ermöglicht es Bürger:innen, audiovisuelle Videoanalysen ohne fachspezifische Ausbildung vorzunehmen. Den Teilnehmer:innen muss dabei der Umgang mit dem Tool und auch mit filmanalytischen Konzepten nähergebracht werden. Eine für VIAN bereits umgesetzte Möglichkeit ist die Entwicklung eines Bildglossars für filmwissenschaftliche Begriffe. Darüber hinaus können Citizen-Humanities-Projekte auch über eine Vereinfachung der Begrifflichkeiten bzw. das Ziel einer weniger detaillierten Analyse an verschiedene soziale Gruppen angepasst werden.

Nach einer Präsentation des aktuellen Glossars und einiger Ergebnisse meiner Arbeit, VIAN für Forscher:innen mit intellektuellen Beeinträchtigungen zugänglich zu machen, hoffe ich, die Podiumsteilnehmer:innen und das Publikum zu einer Diskussion jener komplexen Probleme anzuregen, mit denen uns die Citizen Humanities konfrontieren.

Teodora Vukovic (Zürich) konzentriert sich auf die technischen Aspekte einer möglichen Standardisierungsbestrebung für Filmannotation. Für die multimodale Analyse in der Filmwissenschaft müssen verschiedene Datentypen kombiniert und synchron gehalten werden, z.B. Kameraperspektiven mit Transkripten von Sprache und visuellen Inhalten. Die Komplexität dieser multimodalen Daten erschwert es, allgemeingültige und umfassende Standards vorzuschlagen. Es fehlt an standardisierten Softwarelösungen, und schließlich war die multimodale Analyse lange ein weitgehend qualitatives Gebiet, das keine ausgefeilten Datenstrukturen benötigte, wie sie in Gebieten mit einer längeren quantitativen Tradition, wie etwa der Korpuslinguistik, zu finden sind. Mit dem Aufkommen neuer Verarbeitungstechnologien, KI-Tools und der Zunahme multimodaler Software wie VIAN-DH wird die Frage nach Standards und Konventionen immer drängender. Eine der wichtigsten Anforderungen ist ein Datenrepräsentationsformat, das flexibel genug ist, um den verschiedenen Annotationstypen gerecht zu werden, und konkret genug, um in eine Datenbank implementiert oder für die Korpuskompilierung verwendet werden zu können. Eine weitere Anforderung sind Annotationskonventionen für multimodale Transkripte von nonverbalen Daten, die die vermittelte Bedeutung ein-

deutig und klar identifizieren, aber auch skalieren und korpusübergreifend verwendet werden können. Diese Aspekte sollen im Panel diskutiert werden.

Ariadne Baresch (Trier) nimmt die strukturellen Aspekte potentieller Standardisierungsbestrebungen in den Blick. Eine Festlegung auf ein „open vocabulary“ für die digitale Filmannotation spiegelt den Bedarf der Community wider, sich auf Standards, gemeinsame Annotationskonzepte und ein stabiles, umfassendes Tool zu einigen. Das beim Annotieren entstehende Spannungsfeld zwischen idiosynkratischer Betrachtungsweise und einer möglichst objektiven Auszeichnung wird derzeit innerhalb jedes, sich damit beschäftigenden Forschungsprojekts im Bereich Film neu verhandelt. Die Erarbeitung eines Annotationsstandards könnte die Überlegungen und Erfahrungen, die bei diesen Forschungsprojekten erarbeitet werden, bündeln und zugleich den Einstieg in die Filmannotation erleichtern. Der Standard muss dabei folgenden Ansprüchen genügen: er sollte in einem offenen Format angeboten werden, interoperabel und von der Community bei Bedarf flexibel erweiterbar sein, um die entstandenen Daten und Metadaten zur Weiterverarbeitung verfügbar zu machen. Das Annotationsvokabular sollte unterschiedliche Annotationsphänomene und Analyserichtungen innerhalb der Filmwissenschaft beinhalten, sodass Nutzende, egal ob Anfänger:innen oder Fortgeschrittene, den Standard für ihre individuellen Bedürfnisse verwenden können. Im Panel soll daher der potenzielle Bedarf für eine Film Encoding Initiative in die Diskussion gebracht werden. Innerhalb dieser könnten, ähnlich der Modelle der TEI und der MEI xml-basiert Elemente für in der Filmannotation relevante Phänomene angeboten werden, welche die Frage der Multimodalität von Film widerspiegeln und strukturieren.

Bibliographie

Bakels, Jan-Hendrik, Thomas Scherer, Jasper Stratil und Henning Agt-Rickauer. 2020. „AdA Filmontology – A Machine-Readable Film Analysis Vocabulary for Video Annotation.“ In *Book of Abstracts DH2020 Conference* 443–445. https://dh2020.adho.org/wp-content/uploads/2020/07/488_AdAFilmontologyamachinereadable-FilmAnalysisVocabularyforVideoAnnotation.html (zugegriffen: 2. August 2022).

Dang, Sarah-Mai. 2020. „Forschungsdatenmanagement in der Filmwissenschaft: Daten, Praktiken und Erkenntnisprozesse.“ *montage AV* (Januar 2020), 119–40.

Dang, Sarah-Mai. 2022. „o.J. – Recherchepraktiken, Datenquellen und Modellierungen.“ In *Doing Research. Wissenschaftspraktiken zwischen Positionierung und Suchanfrage*, hrsg. von Sandra Hofhues und Konstanze Schütze, 330–37. Bielefeld: Transcript. [im Erscheinen]

Diecke, Josephine. 2022. „Teaching Digital Methods in Film Studies: Managing Tools and Expectations.“ In *2022 NECS Conference*, University of Theatre and Film I.L. Caragiale (UNATC), Bucharest.

Heftberger, Adelheid. 2020. „Eine lohnende Mammutaufgabe – Rahmenbedingungen der digitalen Filmbenutzung im Bundesarchiv.“ *Bibliothek Forschung und Praxis* 44.3, 404–410. <https://doi.org/10.1515/bfp-2020-2035>(zugegriffen: 2. August 2022).

Heftberger, Adelheid, Jakob Höper, Claudia Müller-Birn und Niels-Oliver Walkowski. 2020. „Opening up Research Data in Film Studies by Using the Structured Knowledge Base Wikidata.“ *Digital Cultural Heritage*, hg. von Horst Kremers, 401–410. Cham: Springer. https://doi.org/10.1007/978-3-030-15200-0_27(zugegriffen: 2. August 2022).

Heinisch, Barbara et al. 2021. „Citizen Humanities.“ *The Science of Citizen Science*, hg. von Katrin Vohland, Anne Land-Zandstra, Luigi Ceccaroni, Rob Lemmens, Josep Perelló, Marisa Ponti, Roeland Samson und Katherin Wagenknecht, 97–118. Cham: Springer. https://doi.org/10.1007/978-3-030-58278-4_6(zugegriffen: 2. August 2022).

Hirsbrunner, Simon David. 2019. „Open Your Heart – On Reasons Why Media Scholars Might be Reluctant to Open their Research Data.“ *Open Media Studies* 8. 4. 2019. <https://mediastudies.hypotheses.org/1237>(zugegriffen: 2. August 2022).

Jannidis, Fotis und Julia Flanders. 2019. „A Gentle Introduction to Data Modeling.“ In *The Shape of Data in the Digital Humanities: Modeling Texts and Text-Based Resources*, hrsg. von Julia Flanders und Fotis Jannidis, 26–94. London, New York: Routledge.

Sondervan, Jeroen. 2018. „Open Science and Open Media Studies – Question on a Culture in Transition.“ *Open Media Studies* 29. 10. 2018. <https://mediastudies.hypotheses.org/867>(zugegriffen: 2. August 2022).

Opening Sources – modulare Wege zur Quellenbereitstellung und -edition

Burckhardt, Daniel

burckhardt@ghi-dc.org

Deutsches Historisches Institut Washington

Hörnschemeyer, Jörg

hoernschemeyer@dhi-roma.it

Deutsches Historisches Institut in Rom

König, Mareike

MKoenig@dhi-paris.fr

Deutsches Historisches Institut Paris

Schulz, Julian

Schulz@MaxWeberStiftung.de

Geschäftsstelle der Max Weber Stiftung, Bonn

Grallert, Till

till.grallert@hu-berlin.de

Humboldt-Universität zu Berlin

Keck, Jana

keck@ghi-dc.org

Deutsches Historisches Institut Washington

Hintergrund

Das Erstellen von digitalen Quelleneditionen gehört mit zu den etabliertesten und einflussreichsten Praktiken und Methoden im Bereich der Humanities Computing bzw. Digital Humanities (Burnard 2014). Die Vielzahl praktischer Erfahrungen und eine fundierte theoretische Reflexion führten zu einer Konvergenz, die sich sowohl an der Stabilität und weiten Verbreitung technischer Standards wie der TEI, in der zunehmenden Akzeptanz von Kriterien für die Bewertung digitaler Editionen (z.B. Sahle 2014) und einem breiten Konsens zu den methodischen, organisatorischen und sozialen Rahmenbedingungen ihrer Erstellung (Fritze et al. 2022) ablesen lässt.

Neben dieser Konvergenz zeigt sich in der Praxis allerdings ein sehr breites Spektrum digitaler Bereitstellungsarten, von Formaten und Zugängen zu historischen Quellen, die wir im Panel „Opening Sources“ adressieren möchten. Im Sinne der Open Humanities steht dabei der offene Zugang zu einem bestimmten Quellenbestand als Hauptziel im Zentrum des editorischen Tuns. Anhand von vier exemplarischen editorischen Projekten, angesiedelt an den Instituten der Max Weber Stiftung (MWS), soll diese Vielfalt von Formen und Zugängen verdeutlicht und als Ausgangspunkt für eine umfassendere Diskussion genutzt werden.

Die Mehrheit dieser Projekte erfüllt die Kriterien an eine wissenschaftliche Edition gemäß des erwähnten Kriterienkataloges nicht oder nicht ausreichend.¹ Auch fallen sie nicht unbedingt unter den Begriff *born digital editions*, den Patrick Sahle dahingehend definiert, dass solche Projekte nicht ohne Verlust an wesentlichen Informationen und Funktionen in eine herkömmliche Papierform übertragen werden können (2013, Bd. 2, 149). Dabei teilen die Projektbeteiligten diese Kriterien durchaus. Jedoch erlauben die vorhandenen organisatorischen, finanziellen oder personellen Ressourcen nur Teilschritte der Quellenpublikation und -edition. Dank der Konvergenz von Formaten für Metadaten (MODS/teiHeader), Digitalisate (TIFF/PNG/JPEG), Transkripte und kritische Apparate (TEI) sowie deren Verschränkung (METS, IIIF) handelt es sich – so die erste These des Panels – beim datenzentrierten Ansatz, der den verschiedenen digitalen Zugangsformen zugrunde liegt, weniger um klar voneinander abzugrenzende Genres, sondern um ein gestuftes Editionsmodell, dessen Modularität wir ausgehend von vier kurzen Input-Referaten zur Diskussion stellen. Denn – so die zweite These des Panels – sehr viele Editionsprojekte dürften im Spannungsfeld von Anspruch und Wirklichkeit vor ähnlichen Herausforderungen stehen, sodass ein gemeinsames Nachdenken über Lösungsansätze lohnend erscheint (vgl. Sayers 2017).

Leitfragen

Die folgenden Punkte stehen im Zentrum der Input-Referate und sollen in der Diskussion im Panel und mit den Teilnehmer_innen vertieft werden:

- **Linearität des Editionsprozesses**
Wie weit deckt sich das informationswissenschaftliche Paradigma eines linearen Vorgangs von der Erfassung der Metadaten und Bereitstellung des Digitalisats über die Erstellung des Transkripts und dessen Annotation mit der Praxis des Edierens in den Geisteswissenschaften? Oberbichler et al. (2021) zeigen am Beispiel von Zeitungsarchiven, dass eine forschungsgeleitete digitale Hermeneutik immer wieder die prozessuale Linearität durchbricht. Nicht allein die Auswahl der zu edierenden Quellen, auch deren Datierung und Transkription setzt häufig Wissen voraus, das erst beim *close-reading* bei der Quellenannotation entsteht. Ein zeitlich gestrecktes oder institutionell verteiltes Verfahren droht den Rückfluss neuer Erkenntnisse in einen vorhergehenden Arbeitsschritt zu erschweren, so dass auch bei digitalen Editionen eine Abkehr vom linearen Wasserfall-Modell hin zu agilen, iterativen Projekten zu beobachten ist (Ferraro et al. 2018). Und würden solche Teilschritte überhaupt von Forschungsförderern Finanzierung erhalten?
An welchen Stellen lässt sich der Publikationsprozess unterbrechen und zu einem späteren Zeitpunkt weiterführen? Was sind Mindestanforderungen im Sinne der Open Humanities für dieses Vorgehen (freie Lizenzen, breit akzeptierte Standards, offene Schnittstellen, lückenlose Dokumentation)? Sind das bloß notwendige, oder auch hinreichende Bedingungen? Welche Parameter müssen von vornherein festgelegt werden (z.B. Browsing-, Such- und Auswertungsmöglichkeiten), welche können nachträglich ohne Nacharbeiten an Metadaten und Auszeichnung eingebracht werden?
- **Vom Digitalisat zum Text**
(Putnam 2016) hat die fundamentale Bedeutung der Findbarkeit von (Voll-)Texten für den digitalen Wandel in den Geisteswissenschaften herausgearbeitet. „Digitale Suche bietet eine vermittlungsfreie Entdeckung“. Deshalb ist die einfache Durchsuchbarkeit für eine breite wissenschaftliche Rezeption eines Quellenbestandes meist wichtiger als dessen minutiöse Annotation. Der Durchbruch bei der automatisierten Handschriftenerkennung (HTR) in den letzten Jahren ermöglicht eine Automatisierung, die zuvor Druckerzeugnissen in lateinischen Zeichensätzen vorbehalten war. Da je nach Schreibhand passende Trainingsdaten die Voraussetzung sind, kann sich hier die Verschränkung mit Citizen-Science-Ansätzen zum Crowd-Sourcing anbieten.
- **Vom Text zur Edition**
Die zentrale Frage beim Übergang zur Edition ist die Frage des *Edendum*. Soll und kann der Gesamtbestand an Transkripten inhaltlich annotiert und umfassend kommentiert werden? Falls nicht, sind Kriterien wie „Repräsentativität“, die „Vollständigkeit in Teilen“ beispielsweise für bestimmte Textgattungen

oder Zeiträume oder gar „Popularität“ (häufig zitiert oder im Textarchiv oft aufgerufen) zielführend? Ist es auch in diesem Schritt hilfreich, automatisierte Verfahren oder Bürgerwissenschaftler_innen zur Annotation und Verlinkung mit Normdaten in Erwägung zu ziehen?

- **Verlässlichkeit**

Eine schrittweise Edition ist keineswegs eine „Edition light“. Die Streckung des Prozesses kann nur gelingen, wenn jeder einzelne Schritt passende Qualitätssicherungsverfahren und die dauerhafte Sicherung der Zwischenergebnisse gewährleistet. Im Vordergrund stehen hier die FAIR-Prinzipien, die vor allem die Zugänglichkeit und die spätere Weiternutzung gewährleisten. Auch bei Quellenbeständen aus nicht-indigenen Kontexten, sollten zusätzliche Aspekte wie *Collective Benefit*, *Responsibility* und *Ethics* aus den CARE-Prinzipien berücksichtigt werden.

Die Dauerhaftigkeit bezieht sich auf die langfristige Sicherung der Forschungsdaten wie auf die einfache Zugänglichkeit in einer niedrighschwellig verfügbaren Präsentationsumgebung (vgl. Fritze et al. 2022, Abschnitt 12).

Format

Nach der Einführung durch die Organisator_innen (5 Min.) wird jede_r Panelist_in das eigene Projekt mit Bezug auf diese Leitfragen kurz vorstellen (je 5 Minuten). In einem zweiten Teil suchen wir Antwortvorschläge, zunächst in einer Diskussion unter den Panelist_innen (30 Min) und anschließend gemeinsam mit den anwesenden Kolleg_innen (30 Min.). Die Publikation der Ergebnisse in einem Blogbeitrag ist geplant.

Statements

Einleitung und Moderation

Daniel Burckhardt (Deutsches Historisches Institut Washington) und Julian Schulz (Geschäftsstelle der Max Weber Stiftung, Bonn)

Wenn Technik und Bewusstsein voranschreiten: die drei Leben der Korrespondenz der Constance de Salm (1767-1845)

Mareike König (Deutsches Historisches Institut Paris)

Bisweilen verhindern zögerliche Projektpartner die freie Bereitstellung von Digitalisaten, begrenzen die Mittel das Maximalziel bei Digitalisierung und Erschließung von Quellen und eröffnet die technische Entwicklung neue Möglichkeiten. So geschehen beim Projekt der Digitalisierung und Inventarisierung der Korrespondenz von Constance de Salm. Zu Projektbeginn war weder die besitzende Institution bereit, die Digitalisate ohne Anmeldung und ohne Wasserzeichen online zu stellen, noch gab es die Mittel für eine zweisprachige Erschließung geschweige denn für eine vollständige Transkription der

Briefe. Der Impulsbeitrag zeichnet exemplarisch nach, wie dennoch in den oben genannten Einzelstufen – vom Digitalisat zum Text und vom Text zur Edition – trotz Pausen die Korrespondenz der Constance de Salm schrittweise ediert werden konnte: durch einen langen Atem beim Verhandeln mit Partnern, Vernetzung mit ähnlichen Projekten, Nachnutzung von Workflows und Verfolgen kleiner Etappenziele.

Quellen ohne Ressourcen: Ansätze des minimal computing für die Erschließung arabischer Periodika

Till Grallert (Humboldt-Universität zu Berlin, ehem. Orient-Institut Beirut)

„Open Arabic Periodical Editions“ ist ein Projekt, das die Rahmenbedingungen für die digitale Erschließung von Quellen in und aus Gesellschaften des Globalen Südens mit den Ansätzen des *minimal computing* in der Praxis adressiert (Gil und Ortega 2016, Risam 2019). Hierbei geht es um mehrere, miteinander verwobene Schichten der Unzugänglichkeit, die DH-Projekte des Globalen Südens konstant verhandeln müssen (Grallert 2022): Hegemonie des anglophonen Globalen Nordens über die Infrastrukturen der Wissensproduktion (Wissensorganisation, Forschungsfelder, Fördergelder, Arbeitssprachen, vgl. Fiorimonte 2021); computationale Methoden und Werkzeuge der digitalen Editorik, die für nicht-lateinische Schriften und Sprachen des Globalen Südens nicht erprobt oder verfügbar sind (z.B. OCR, NER, vgl. Auddy 2022); begrenzter Zugang zur technischen Grundversorgung mit Strom, Internet und Hardware.

The Wisdom of the Crowd: Quellen öffnen mit Citizen Scholars

Jana Keck (Deutsches Historisches Institut Washington)

„Migrant Connections“ ist eine digitale Forschungsinfrastruktur (Omeka S) des DHI Washington, welche Zugang zu diversen Quellenmaterialien wie Briefen, Tagebüchern, und Zeitungsartikeln zur deutschen Migration in die USA bietet. Gemeinsam mit Citizen Scholars werden kontinuierlich Quellen gesammelt, digitalisiert, transkribiert, übersetzt, Metadaten erstellt und kontextualisiert. Durch die Expertise der Citizens aus Deutschland und den USA konnten bisher hunderte von Quellen erschlossen werden, die Einblicke vor allem in die Alltagsgeschichte von bisher wenig beachteten, „gewöhnlichen“ historischen Akteur_innen geben. Der Beitrag möchte die Potentiale einer schrittweisen Edition bei offener Kollaboration mit Citizen Scholars im Forschungsprozess beleuchten, dabei die Frage der Qualitätssicherung thematisieren und gleichzeitig hervorheben, wie diese Offenheit uns als Wissenschaftler_innen dazu bringt, Vertrauen zu schenken und Kontrolle abzugeben.

„Königsweg“ Digitale Edition. Offen, aber für wie lange?

Jörg Hörnschemeyer (Deutsches Historisches Institut Rom)

Der Impulsbeitrag geht der Frage nach, wie offen und nachhaltig die „letzte Stufe“ des eingangs skizzierten Modells der Bereitstellung digitaler Quellen, die Digitale Edition, wirklich ist. „Ferdinand Gregorovius – Poesie und Wissenschaft. Gesammelte deutsche und italienische Briefe“ ist eine nach „allen Regeln der Kunst“ erarbeitete Digitale Edition, die über XML/TEI-annotierte, text- und sachkritisch erschlossene Brieftranskriptionen verfügt und mit umfangreichen Normdatenverknüpfungen, einem projektspezifisch entwickelten User Interface und einer REST-API versehen wurde. Trotz (oder auch gerade wegen?) dieser komplexen Architektur ist absehbar, dass die Edition in 10-20 Jahren vor der zentralen Frage stehen wird, ob und wie ein offener Zugang über die verschiedenen Zugänge auch weiterhin gewährleistet werden kann und wie neue Forschungsergebnisse in eine als abgeschlossen angesehene Edition einfließen können. Könnten z.B. Semantic Web-Technologien (Wettlaufer 2018), Social Editing (Crompton et al. 2016) und kollaborative, (Fach)community-getriebene Ansätze Antworten auf diese beiden Fragebereiche liefern? Und welche Ressourcen benötigt dann eine Langzeitstrategie für Editionen auf institutioneller Ebene?

Fußnoten

1. Vgl. IDE-Richtlinien, 1. Geltungsbereich, Definitiorik, wonach Digitale Editionen „nicht nur in digitaler Form publiziert [werden], sondern ... in ihrer Methodologie einem digitalen Paradigma [folgen]“.

Bibliographie

Auddy, Purbasha. 2022. „Mining Verbal Data from Early Bengali Newspapers and Magazines: Contemplating the Possibilities.“ In *Global Debates in the Digital Humanities*, hg. von Domenico Fiormonte, Sukanta Chaudhuri, und Paul Ricarte, 117–26. Debates in the Digital Humanities. Minneapolis: University of Minnesota Press.

Burnard, Lou. 2014. *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*. Marseille: OpenEdition Press. DOI: <https://doi.org/10.4000/books.oep.426>.

Crompton, Constance, Daniel Powell, Alyssa Arbuckle, Ray Siemens, Maggie Shirley und Devonshire Manuscript Editorial Group. 2016. „Building a Social Edition of the Devonshire Manuscript“ In *Digital Scholarly Editing: Theory and Practice*, hg. von Matthew James Driscoll und Elena Pierazzo, 19–39. DOI: <https://doi.org/10.11647/OBP.0095>.

Ferraro,GINESTRA und Anna-Maria Sichani. 2018. „Design as Part of the Plan: Introducing Agile Methodology in Digital Editing Projects“ In *Digital Scholarly Editions as Interfaces. Schriften des Instituts für Dokumentologie und Editorik* 12, hg. von Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber und Gerlinde Schneider, 83–105. Norderstedt. <https://kups.uni-koeln.de/9113/>; urn:nbn:de:hbz:38-91132 (zugegriffen: 3. August 2022).

Fiormonte, Domenico. 2021. „Taxation against Overrepresentation? The Consequences of Monolingualism for Digital Humanities.“ In *Alternative Historiographies of the Digital Humanities*, hg. von Dorothy Kim und Adeline Koh, 333–76. Earth: punctum books. DOI: <https://doi.org/10.53288/0274.100>.

Fritze, Christiane et al. 2022. „Manifest für digitale Editionen“. <https://dhd-blog.org/?p=17563> (zugegriffen: 3. August 2022).

Gil, Alex und Élika Ortega. 2016. „Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing.“ In *Doing Digital Humanities: Practice, Training, Research*, hg. von Constance Crompton, Richard J Lane und Ray Siemens, 22–34. Abingdon: Routledge.

Grallert, Till. 2022. „Open Arabic Periodical Editions: A Framework for Bootstrapped Scholarly Editions Outside the Global North.“ In *Digital Humanities Quarterly* 16, Bd. 2, „Minimal Computing“ hg. von Roopika Risam und Alex Gil. <http://digitalhumanities.org/dhq/vol/16/2/000593/000593.html> (zugegriffen: 3. August 2022).

Oberbichler, Sarah, Emanuela Boros, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen und Mikko Tolonen. 2021. „Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians.“ *Journal of the Association for Information Science and Technology* 73, Issue 2, 225–239. DOI: <https://doi.org/10.1002/asi.24565>.

Putnam, Lara. 2016. „The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast“ *The American Historical Review* 121, 377–402. DOI: <https://doi.org/10.1093/ahr/121.2.377>.

Risam, Roopika. 2019. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston: Northwestern University Press. <https://doi.org/10.2307/j.ctv7tq4hg>.

Sahle, Patrick. 2013. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik*. Norderstedt: BoD. <https://kups.ub.uni-koeln.de/5352/>; urn:nbn:de:hbz:38-53523 (zugegriffen: 3. August 2022).

Sahle, Patrick. 2014. *Kriterienkatalog für die Besprechung digitaler Editionen*. 1.1. Weitere Schriften. Institut für Dokumentologie und Editorik. <https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/> (zugegriffen: 3. August 2022).

Sayers, Jentery, Hrsg. 2017. *Making Things and Drawing Boundaries: Experiments in the Digital Humanities*. Debates in the Digital Humanities 3. Minneapolis: University of Minnesota Press.

Wettlaufer, Jörg. 2018. „Der nächste Schritt? Semantic Web und digitale Editionen“ In *Digitale Metamorphose: Digital Humanities und Editionswissenschaft. (Sonderband der Zeitschrift für digitale Geisteswissenschaften 2)*, hg. von Roland S. Kamzelak und Timo Steyer. DOI: https://doi.org/10.17175/sb002_007.

Vorträge

Algorithmen- gestützte Analyse visuell-materieller Eigenschaften von Briefen

Nantke, Julia

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

Reul, Christian

christian.reul@uni-wuerzburg.de
Universität Würzburg, Deutschland

Bläß, Sandra

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

Flüh, Marie

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

Der visuell-materiellen Gestaltung von historischen Dokumenten und insbesondere von Briefen kommt in der Forschung ein hoher Stellenwert zu. Visuell-materielle Eigenschaften wie die Papier- und Stiftfarbe, die Anordnung von Schrift auf der Seite und die Schrifttrichtung geben Auskunft über historische Bedingungen und konkrete Umstände des Schreibens und der Schreibenden sowie, im Fall von Briefen, über das Verhältnis der Schreibenden zu den jeweiligen Adressat:innen (vgl. Baasner 2008; Wiethölter/Bohnenkamp 2010; Henzel 2020; Lukas/Osthof 2016).

Obwohl die digitale Repräsentation historischer Dokumente in digitalen Editionen und Archiven entscheidende Vorteile hinsichtlich der Darstellung visuell-materieller Charakteristika bietet (vgl. Bohnenkamp-Renzen 2013; Radecke 2015), spielt deren Analyse in computer-gestützten Verfahren noch immer eine untergeordnete Rolle. Die meisten quantitativen Methoden und hierbei genutzten Tools beziehen sich auf den Inhalt und/oder den Schreibstil literarischer und historischer Texte, die bibliografischen Codes (McGann 1991: 77) bleiben jenseits ihrer Repräsentation als Bild-Digitalisate in digitalen Editionen bei der Analyse und Interpretation meistens unberücksichtigt.

Im Rahmen des geplanten Vortrags soll ein Ansatz zur quantitativen, Algorithmen-gestützten Erschließung visuell-materieller Charakteristika in einem Korpus von Briefen aus dem Zeitraum ‚Deutschland um 1900‘ präsentiert werden. Hierbei wollen wir zeigen, wie sich anhand verschiedener (teil-)automatisiert ermittelbarer Werte zur optisch erfassbaren Gestaltung der Briefe inhaltliche Aussagen sowohl über konkrete Dokumente

sowie deren Schreibende und Adressat:innen machen als auch Erkenntnisse über die historischen Dynamiken der Textsorte ‘Brief’ gewinnen lassen. Denn zum einen handelt es sich bei Briefen um einen Dokumententyp, der spätestens seit dem 18. Jahrhundert spezifischen Kodierungen unterworfen ist, bei denen die Ebenen des sprachlichen Ausdrucks und der visuell-materiellen Gestaltung komplex ineinandergreifen (vgl. Baasner 2008). Zum anderen zeichnet sich gerade der untersuchte Entstehungszeitraum der Briefe durch die Lockerung kommunikativer Etikette aus, was im Rahmen des Briefformats zunehmende Spielräume zur individuellen Ausgestaltung eröffnet (vgl. Ehlers 2004).

Bei dem analysierten Korpus handelt es sich um einen Ausschnitt aus dem ca. 35.000 Briefe umfassenden Dehmel-Archiv der Staats- und Universitätsbibliothek Hamburg (SUB), die aktuell im Projekt *Dehmel digital* wissenschaftlich erschlossen werden.¹ Die Briefe richten sich an den um 1900 berühmten, 1920 verstorbenen Dichter Richard Dehmel und stammen von verschiedenen anderen Künstler:innen wie Rainer Maria Rilke, Stefan Zweig, Else Lasker-Schüler, Detlev von Liliencron und Peter Behrens, die in unterschiedlich engem Kontakt zu Dehmel standen und deren Erfolg als Künstler:innen zum Zeitpunkt der Korrespondenzen ebenfalls verschieden groß war. Weiterhin enthält das Korpus auch Briefe von Vertreter:innen des Literaturbetriebs, die mit Dehmel über Publikationsvorhaben, geplante Veranstaltungen oder gemeinsame Projekte verhandelten.

Wir wollen in unserem Beitrag zeigen, wie sich anhand der visuell-materiellen Gestaltung der Briefe Aussagen über die Lebensumstände der Briefschreibenden, deren Beziehung zu Dehmel sowie die Charakteristika epistolarer Kommunikation um 1900 und die Veränderung der Normen der Briefschreibung ableiten lassen, welche sich in dieser Zeit vollziehen.

Die Grundlage für die maschinenlesbare Erschließung visuell-materieller Eigenschaften der Dokumente bilden die Strukturinformationen auf der Briefseite. Diese werden im Rahmen der Layoutanalyse, einem Teilschritt der HTR (Handwritten Text Recognition), semi-automatisch mittels OCR4all² (vgl. Reul et al. 2019) erfasst. Hierbei können Strukturen wie der Seitenspiegel, der Haupttext, Grußformeln, Briefköpfe etc. entweder manuell ausgezeichnet und typisiert oder ein automatisch generierter Vorschlag gezielt korrigiert werden, wobei manuell geprüfte Daten als Trainingsbeispiele für die fortlaufende Verbesserung der algorithmischen Methoden herangezogen werden können. Aus den so erzeugten Daten können anschließend die Informationen über die visuell erfassbaren Merkmale der Briefe automatisiert extrahiert werden, um im Rahmen statistischer Analysen Zusammenhänge zwischen Layout und Briefinhalt aufzudecken (vgl. Busch/Hegel 2017; Hurlbut 2013). So lassen sich z.B. statistische Mittelwerte für den Weißraum der Briefe einzelner Korrespondenzpartner:innen berechnen und miteinander vergleichen oder die Papierfarbe eines Briefs ins Verhältnis zu den innerhalb des Korpus üblichen Färbungen setzen.

Die von uns untersuchten Merkmale lassen sich grob in zwei Kategorien unterteilen: Erstens gibt es grundsätzlich vorhandene visuell-materielle Eigenschaften wie das Format der Briefbögen, das Verhältnis von Weißraum und Textraum sowie der Abstand zwischen Grußformeln

und Textblock, die für die epistolare „Respektsemiotik“ (Ehlers 2004: 21) von großer Aussagekraft sind. Diese können direkt aus den oben erwähnten Auszeichnungen abgeleitet bzw. berechnet werden. Ebenfalls in den Bereich der grundsätzlichen Eigenschaften gehören die Papier- und Stifffarbe, deren „Auswahl und Wirkungsweise [...] in enger Abhängigkeit – historisch wandelbarer – ökonomischer, kultureller, sozialer, ästhetischer u.a. Faktoren sowie des jeweiligen Inhalts und der Funktion des Briefs“ stehen (Henzel 2020: 222). Die computergestützte Identifikation dieser Merkmale erfordert allerdings im Vergleich einen etwas größeren Aufwand: Zunächst wird das originale Farbbild in ein Binärbild umgewandelt, sodass in den zuvor ausgezeichneten Textregionen die Vordergrundpixel weitestgehend der Schrift und die Hintergrundpixel weitestgehend dem unbeschriebenen Papier entsprechen. Nach dem Ausschließen von Übergangspixeln, um Störeffekte zu minimieren, werden die Entsprechungen im Farbbild gesammelt und der jeweiligen Klasse zugewiesen. Abschließend wird separat die durchschnittliche Schrift- und Papierfarbe berechnet, indem zunächst je ein Mittelwert für die drei Farbkanaäle Rot, Grün und Blau gebildet wird und diese zu je einer Farbe kombiniert werden.

Neben diesen generellen visuell-materiellen Eigenschaften beziehen wir zweitens spezifische Charakteristika wie Briefköpfe, Abbildungen und Zeichnungen in unsere Analysen mit ein, die nur in einem Teil der Dokumente enthalten sind und bei denen vorerst lediglich erfasst wird, ob sie vorhanden sind oder nicht. Diese Erfassung zielt darauf, die Dokumente anschließend nach den entsprechenden Charakteristika filtern zu können, um auf breiter Basis Aussagen über die Verbreitung und die konkreten Eigenschaften der Gestaltungsmittel machen zu können.

Zwei Beispiele sollen im Folgenden die dargestellte Vorgehensweise sowie die Aussagekraft der Ergebnisse der visuell-materiellen Analysen exemplarisch illustrieren.

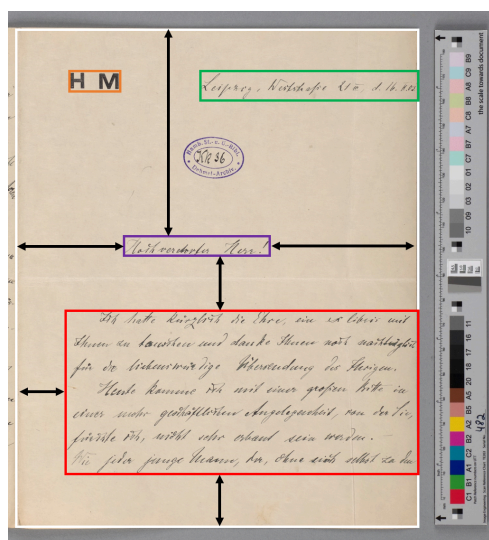


Abb. 1: Brief von Heinz Möller an Richard Dehmel vom 16.02.1903, Dehmel-Archiv der SUB, HANSb18474, S. 1

Der oben abgebildete Brief des Leipziger Herausgebers Heinz Möller vom 16. Februar 1903 ist offensichtlich auf

eigenem Briefpapier verfasst. Wie anhand des Briefkopfs mit den Initialen des Verfassers zu erkennen, tritt Möller im eigenen Namen auf und handelt nicht im Auftrag einer Organisation. Des Weiteren orientiert sich die Gestaltung des Briefs geradezu prototypisch an den Regeln der Respektsemiotik: Das Ausmaß der textfreien Weißräume an den Briefrändern und zwischen den einzelnen Briefteilen (oberer Briefrand und Anrede, Anrede und Textblock) und das äußerst sorgfältige Schriftbild bringen auf visuell-materieller Ebene die Hochachtung des Schreibenden gegenüber dem angeschriebenen Richard Dehmel zum Ausdruck und korrespondieren dabei mit verbalsprachlichen Formeln wie „Hochverehrter Herr“ und „In aufrichtiger Verehrung und Dankbarkeit“. Insbesondere im Vergleich mit dem zweiten Beispiel wird deutlich, dass gerade im Rahmen geschäftlicher Korrespondenz auch zu Beginn des 20. Jahrhunderts die seit der Antike in stetig aktualisierten „Briefstellern“ verbreiteten formalen Regeln zur Abfassung von Briefen noch Gültigkeit besaßen (vgl. Schiegg 2020). Zugleich lässt sich aus der Gestaltung eine Aussage über das Hierarchieverhältnis der beiden Korrespondenzpartner zueinander ableiten: Möller wendet sich hier, wie gesagt in eigener Sache, an Dehmel als „Obmann“ des Kartells lyrischer Autoren, um einen Nachlass für die Honorare einer von Möller geplanten Lyrik-Anthologie zu erbitten. Der Umstand, dass es sich hier um ein Bittschreiben handelt, spiegelt sich eindrucksvoll in der visuell-materiellen Gestaltung des Dokuments.

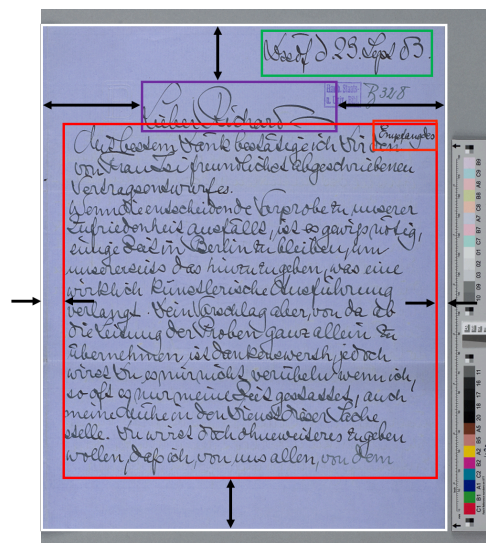


Abb. 2: Brief von Peter Behrens an Richard Dehmel vom 23.09.1903, Dehmel-Archiv der SUB, HANSb313015, S. 1

Das zweite Beispiel, der Brief des Kunsthandwerkers und Designers Peter Behrens an Richard Dehmel vom 23. September 1903, zeigt eine andere Form der signifikanten Individualisierung von Briefpapier: Behrens' Briefbogen enthält zwar keinen aufgedruckten Briefkopf, die farbliche Gestaltung, das lilafarbene Briefpapier hebt ihn aber sehr stark aus dem Feld der breiteren Masse von Briefen aus dem untersuchten Zeitraum hervor und macht ihn, etwa in einem Stapel von Briefen, sofort als einen Brief des Verfassers Behrens erkennbar. Darüber hinaus weist die Gestaltung des Dokuments durch-

aus grundlegende visuell-materielle Charakteristika der Briefschreibung auf, die im Rahmen einer quantitativen Analyse als Eigenschaften des etablierten Brieflayouts der Zeit um 1900 ermittelt werden können: Es besteht, wie für den Brief von Möller festgestellt, ebenfalls ein gewisser Abstand zwischen dem oberen Blattrand und der Anrede sowie zwischen Blattrand und Brieftext, die Anrede ist, wiederum analog zum Brief Möllers, zentriert und der eigentliche Brieftext als Block gesetzt. Im Vergleich mit dem ersten Beispiel zeigen sich allerdings auch deutlich messbare Differenzen, die Rückschlüsse auf das Verhältnis der Korrespondenzpartner erlauben. So besteht ein deutlich geringerer Abstand zwischen oberem Blattrand und Anrede und kein gegenüber dem sonstigen Zeilenabstand vergrößerter Abstand zwischen Anrede und Brieftext. Hinzu kommt eine Wort-Einfügung über der Zeile, die ebenfalls einen Bruch mit der klassischen epistolaren Etikette darstellt und von einer größeren Vertrautheit der Korrespondenzpartner zeugt. Wiederum spiegelt hier die visuell-materielle die verbalsprachliche Gestaltung. Peter Behrens war nicht nur künstlerisch mit Richard Dehmel verbunden, sondern auch ein guter Freund, was sich in der Anrede als „Lieber Richard“ niederschlägt.

Die beiden Beispiele zeigen andeutungsweise, wie die automatisierte, quantitative Auswertung der visuell-materiellen Gestaltungsformen in einem Briefkorpus dazu beitragen kann, bereits vor der genauen inhaltlichen Sichtung der Briefe Hypothesen über das Verhältnis der Korrespondenzpartner:innen und den Zweck der Kommunikation anzustellen sowie darüber hinaus grundsätzliche Charakteristika epistolarer Kommunikation in einem bestimmten Zeitraum zu erschließen.

Fußnoten

1. Vgl. <https://dehmel-digital.de>.
2. <https://www.ocr4all.org>

Bibliographie

Baasner, Rainer. Stimme oder Schrift? Materialität und Medialität des Briefs. Adressat: Nachwelt. Briefkultur und Ruhmbildung. Hg. v. Detlev Schöttker. München: Wilhelm Fink, 2008, S. 53–69.

Bohnenkamp-Renken, Anne, Hg. Medienwandel / Medienwechsel in der Editionswissenschaft. De Gruyter, 2013, <https://doi.org/10.1515/9783110300437>.

Busch, Hannah, und Philipp Hegel: Automatic Layout Analysis and Storage of Digitized Medieval Books. In: Digital Philology. A Journal of Medieval Cultures, 6, 2 (2017), S. 196–212, <https://doi.org/10.1353/dph.2017.0010>.

Dehmel digital. Hg. v. Julia Nantke unter Mitarbeit von Sandra Bläß und Marie Flüh, seit 2021, <https://dehmel-digital.de>.

Ehlers, Klaas-Hinrich: Raumverhalten auf dem Papier. Der Untergang eines komplexen Zeichensystems dargestellt an Briefstellern des 19. und 20. Jahrhunderts. In: Zeitschrift für germanistische Linguistik 32, 1 (2004), S. 1–31.

Henzel, Katrin: Materialität des Briefs. In: Handbuch Brief. Von der Frühen Neuzeit bis zur Gegenwart. Hg. v. Marie Isabel Matthews-Schlinzig, Jörg Schuster, Gesa Steinbrink u. Jochen Strobel. Berlin/Boston: De Gruyter 2020, S. 222–231.

Hurlbut, Jesse: The Manuscript Average, Part 1. Dezember 2013, https://jessehurlbut.net/wp/mssart/?page_id=2097.

Lukas, Wolfgang und Matthias Osthof. „Physische vs. gedutete Räumlichkeit. Zur Auszeichnung spatialer Informationen in der historisch-kritischen Ausgabe C.F. Meyers Briefwechsel“. Jahrbuch für Computerphilologie online, 2016, <http://computerphilologie.digital-humanities.de/jg09/lukasosthof.pdf>.

McGann, Jerome: The Textual Condition. Princeton: Princeton University Press 1991.

Radecke, Gabriele: Materialautopsie. Überlegungen zu einer notwendigen Methode bei der Herstellung von digitalen Editionen am Beispiel der Genetisch-kritischen und kommentierten Hybrid-Edition von Theodor Fontanes Notizbüchern. TextGrid. Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften. Hg. v. Heike Neuroth u. a., 2015, S. 39–56.

Reul, Christian; Christ, Dennis; Hartelt, Alexander; Balbach, Nico; Wehner, Maximilian; Springmann, Uwe; Wick, Christoph; Grundig, Christine; Büttner, Andreas; Puppe, Frank: OCR4all – An open-source tool providing a (semi-) automatic OCR workflow for historical printings. In: Applied Sciences, 9(22), 2019.

Schiegg, Markus: Briefsteller. In: Handbuch Brief Von der Frühen Neuzeit bis zur Gegenwart. Hg. v. Marie Isabel Matthews-Schlinzig, Jörg Schuster, Gesa Steinbrink u. Jochen Strobel. Berlin/Boston: De Gruyter 2020, S. 276–290.

Wiethölter, Waltraud u. Anne Bohnenkamp (Hg.): Der Brief, Ereignis & Objekt. Berlin: Stroemfeld, 2010.

„Auch heute war die Stimmung im Allgemeinen fest.“
Zero-Shot Klassifikation zur Bestimmung des Media Sentiment an der Berliner Börse zwischen 1872 und 1930

Wehrheim, Lino

lino.wehrheim@ur.de
Universität Regensburg

Borst, Janos

borst@informatik.uni-leipzig.de
Universität Leipzig

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de
Universität Leipzig

Niekler, Andreas

aniekler@informatik.uni-leipzig.de
Universität Leipzig

Motivation und Korpus

In einer Szene des Films „The Wolf of Wall Street“ hält der Börsenmakler Jordan Belfort, verkörpert durch Leonardo Di Caprio, eine Ansprache vor seinen Mitarbeitern, woraufhin eine Parade nackter Musiker durch das Großraumbüro zieht, gefolgt von einer Gruppe leicht bekleideter Frauen und einer orgiastischen Feier in den Büroräumen. Auch wenn sich der Film auf eine besonders schillernde Persönlichkeit konzentriert, so veranschaulicht er doch, wie verrückt es an der Börse bisweilen zugehen kann. Tatsächlich spielen dort wie auch in anderen Lebensbereichen nicht nur „harte“ Informationen, sondern eben auch vermeintliche „weiche“ Aspekte wie Stimmungen und Gefühle eine wichtige Rolle, die sich in kollektiven Irrationalitäten, Hypes und Ängsten niederschlagen (Akerlof and Shiller 2009; Shiller 2015). Der Ökonom John Maynard Keynes prägte 1936 dafür den Ausdruck der „animal spirits“ (Keynes 1936). Ein Ansatz, die Bedeutung der Stimmung für die Börse greifbar zu machen, besteht darin, das in der Finanzpresse zum Ausdruck kommende Media Sentiment mittels Sentiment Analyse (Liu 2015) zu messen und dessen Einfluss auf die Kursentwicklung zu bestimmen (Tetlock 2007; García 2013; Hanna, Turner und Walker 2020).¹

Das so gemessene Media Sentiment wird dabei als Spiegel der Stimmung der Börsianer, des Investor Sentiment, interpretiert. Das vorliegende Projekt verfolgt das Ziel, Sentiment Daten für die Berliner Börse für den Zeitraum zwischen 1872 und 1930 zu generieren, die dann für verschiedene inhaltliche Fragestellungen genutzt werden können.² Das Ziel des Beitrags besteht darin, die mit der Erhebung dieser Daten verbundenen Herausforderungen aufzuzeigen und unseren Ansatz der Zero-Shot-Klassifikation vorzustellen.

Datengrundlage ist ein Textkorpus, das aus täglichen Marktberichten der Berliner Börsen-Zeitung (BBZ), der wichtigsten Finanzzeitung dieser Epoche, besteht.³ Diese Berichte, die im Median etwa 540 Wörter umfassen, enthalten eine kompakte verbale Beschreibung des täglichen Geschehens an der Berliner Börse: In welcher Stimmung befand sich die Börse? Was beeinflusste die allgemeine Stimmung? Welche Aspekte wiesen Besonderheiten auf? Wie entwickelten sich bestimmte Teilmärkte, etwa Eisenbahnaktien?

Die Berichte wurden von uns in einem mehrstufigen Prozess aus den ganzseitigen Scans der BBZ extrahiert (Liebl und Burghardt 2020), die die Staatsbibliothek Berlin zur Verfügung stellt.⁴ Dazu wurde mittels einer eigenen OCR-Pipeline und Layout-Detection zunächst die relevante Ausgabe (Morgen vs. Abend), dann die relevante Seite und zuletzt der jeweils relevante Seitenabschnitt identifiziert, was sich aufgrund des wandelnden Layouts der Zeitung und insbesondere der Berichte als durchaus komplexes Problem erwies. Insgesamt wurden so knapp 18 000 Berichte extrahiert, was einem Textvolumen von etwa 9,87 Millionen Wörtern entspricht. Für den Großteil des Untersuchungszeitraums zählt das Korpus knapp 300 Berichte pro Jahr (der Börsenhandel fand von Montag bis Samstag statt). Für die Zeit des Ersten Weltkriegs und die Jahre zwischen 1922 bis 1924 ergeben sich aufgrund eines wenig standardisierten und häufig wechselnden Berichtsformats sowie häufiger Börsenfeiertage bislang nur 150 bis 200 Texte pro Jahr. Einige weitere Datenlücken in den 1870er Jahren, die aus unvollständigen Überlieferungen der BBZ resultieren, wurden mit dem Pendant aus der *Vossischen Zeitung* gefüllt, die ebenfalls einen täglichen Börsenbericht veröffentlichte. Dabei wurde durch verschiedene Tests sichergestellt, dass beide Zeitungen eine weitestgehend übereinstimmende Bewertung der Gesamtstimmung an der Berliner Börse angeben.⁵

Methode

Die Börsenberichte der BBZ weisen im Hinblick auf die Bestimmung des in ihnen enthaltenen Sentiments eine vierfache Herausforderung auf. Zunächst wird die Analyse durch die spezielle Domäne verkompliziert. Die Tonalität eines ökonomischen Texts bzw. eines Finanztexts hängt in hohem Maße vom Kontext bzw. der „Richtung“ der Aussage ab (Loughran und McDonald 2011; Malo et al. 2014; Xing et al. 2020). So induziert bspw. das Wort „Verlust“ nur dann eine negative Tonalität, wenn dieser „steigt“, „sich einstellt“, „verharrt“, etc. Geht er jedoch zurück, resultiert dagegen eine Aussage mit positiver Tonalität. Im Börsen-Kontext besonders problematisch: Für die eine Marktseite mögen fallende Kurse etwas Negatives sein. Für die andere, die darauf gesetzt hat, dass die Kurse zurückgehen, ist der gleiche Vorgang hingegen etwas Positives (Hausse- vs. Baissepartei). Hinzu kommt, dass das Börsenjargon durch ein ganz eigenes Vokabular gekennzeichnet ist („Contremine“, „debarrassieren“, „Reprise“, usw.). Neben solch hochspezifischen Ausdrücken finden sich weiterhin viele Beispiele für die Verwendung eines Standardlexikons um eine spezifische Marktstimmung – teilweise geradezu metaphorisch – zu beschreiben: In der Fachsprache der Börse sind die Geschäfte etwa „flau“, „matt“, „fest“ oder „lustlos“ (Krupke 1904; Kautsch 1912). Einerseits standardisiert derlei Jargon zwar den Sprachgebrauch erheblich; andererseits erschwert es den Einsatz von off-the-shelf Lösungen was beispielsweise die automatische Sentiment Analyse solcher Sprachbelege angeht, da in Standardressourcen die Börsen-spezifische Bedeutung diese Begriffe bzw. die Begriffe selbst nicht berücksichtigt werden. Dies wird zuletzt noch dadurch verstärkt, dass wir

es teils mit einem veralteten Sprachgebrauch zu tun haben, der sich neben terminologischen Besonderheiten durch Schachtelsätze, Verklammerungen, vielfache Verneinungen, Konjunktive, Querverweise und andere Besonderheiten auszeichnet. Historizität und Sprachwandel erschweren jedoch nicht nur die Sentiment-Analyse, sondern auch andere NLP-Ansätze wie etwa Named Entity Recognition (siehe etwa Hellrich et al. 2019, Ehrmann et al. 2021), da im Falle historischer Sprache viele Standard-NLP-Ressourcen nicht ohne Weiteres nutzbar sind.

Vor dem Hintergrund dieser vielfältigen Herausforderungen, die sich durch die spezifische Domäne und die historische Sprachstufe ergeben, waren erste Experimente mit den in Tabelle 1 aufgeführten Sentiment-Lexika nicht erfolgreich. Dies untermauert bestehende Erkenntnisse aus der einschlägigen Fachliteratur bzgl. einer geringeren Performance Wörterbuch-basierter Ansätze gegenüber Machine-Learning-Verfahren (siehe etwa Mishev et al. 2020 und van Atteveldt et al. 2021).⁶

Als Alternative wurden deshalb aktuelle Ansätze aus dem Bereich neuronaler Sprachmodelle erprobt. Durch das Vortrainieren großer Transformer-basierter Sprachmodelle, wie etwa BERT, wurden in den letzten Jahren immer wieder Durchbrüche in verschiedenen Anwendungsbereichen des Natural Language Processing erzielt (Devlin et al. 2019). Das übliche Vorgehen ist es, ein solch vortrainiertes Sprachmodell auf Task-spezifischen Daten nachzutrainieren und es damit an die eigene Anwendungsdomäne anzupassen (Finetuning). Man spricht hier von *transfer learning* (Zhuang et al. 2020). Unter dem Namen Zero-Shot-Klassifikation (früher: Data-less Classification, Chang et al. 2008) existieren seit Kurzem Verfahren, bei denen für die Zielaufgabe gar keine aufwendig erstellten Trainingsdaten mehr für das Finetuning zur Verfügung stehen müssen, aber dennoch ein domänenspezifischer Klassifikator generiert werden kann (vgl. Yin et al. 2019, Veeranna et al. 2016, Brown et al. 2020). Der große Vorteil des Zero-Shot-Verfahrens ist es, dass hierfür keinerlei (also zero) Trainingsdaten vorliegen müssen – deren Generierung je nach Projekt mit hohen Kosten verbunden sein kann –, sondern Klassifikationsergebnisse im Sinne eines Transfer-Lernens erzielt werden. Einer der erfolgreichsten Ansätze für die Zero-Shot-basierte Textklassifikation stützt sich dabei auf sog. *Entailment*-Modelle, die darauf trainiert sind, einen logischen Widerspruch oder Implikationen zweier Sätze zu erkennen (Yin et al. 2019). Dazu werden die Zielkategorien für den Klassifikator in natürlichsprachliche Sätze umformuliert und mit dem zu klassifizierenden Text verglichen.

Tabelle 1: Übersicht zu genutzten Sentiment-Lexika für die deutsche Sprache.

Name	Sprache	URL
BPW dictionary	Deutsch (optimiert für Finanzsprache)	https://www.uni-giessen.de/fbz/fb02/forschung/research-networks/bsfa/textual_analysis/index.html
SentiWS	Deutsch (allgemein)	http://www.ulliwaltinger.de/sentiment/
German Emotion Analysis	Deutsch (allgemein)	https://www.romanklinger.de/emotion/
Affective Norms for German Sentiment Terms (ANGST)	Deutsch (allgemein)	https://link.springer.com/article/10.3758/s13428-013-0426-y
Lexikon von Chen/Skiena	Multilingual	https://aclanthology.org/attachments/P14-2063/Data-sets.zip

In unserem konkreten Anwendungsfall erstellen wir einen Klassifikator für die Sentiment-Kategorien „positiv“, „neutral“ und „negativ“, welche entsprechend in den folgenden Sätzen verbalisiert werden: „Die Stimmung an der Börse ist {positiv|neutral|negativ}“. Der Klassifikator entscheidet durch automatische Analyse aller sprachlichen *Entailments* sodann, welche dieser drei Hypothesen sich am wahrscheinlichsten aus dem Eingabetext schlussfolgern lässt. Verwendet wurde hierzu ein BERT-Modell⁷, das auf dem deutschsprachigen Teil des XNLI (*cross natural language inference*) Datensatzes (Conneau et al. 2018) getestet wurde.

Vor der Klassifizierung wurden zunächst alle Berichte in einzelne Sätze segmentiert. Nach ersten Experimenten fiel auf, dass einige Sätze in den Marktberichten stark deskriptiven Charakter haben, angezeigt bspw. durch eine längere Auflistung von Unternehmen in direkter Abfolge. Solche Sätze enthalten aber keine relevanten Informationen aus Perspektive des Media Sentiment und wurden deshalb schon im Vorfeld heuristisch herausgefiltert. Dabei wurde Domänenwissen in generalisierbare Heuristiken übertragen, die dann automatisiert auf das gesamte Korpus angewendet wurden. Beispielsweise prüft die Heuristik, ob längere Aufzählungen oder gehäuftes Auftreten von numerischen Werten in einem Satz vorkommen. Letzteren kommt eine besondere Bedeutung zu, da sie meist in Aussagen wie „Der Kurs ist um 2 % gefallen“ oder „Der Kurs stieg bis 218“ auftreten. Zwar können diese Sätze, wie das erste Beispiel zeigt, durchaus eine Form von Tonalität enthalten. Allerdings drücken sie aus unserer Sicht keine Stimmungen oder Gefühle, sondern eine finanzwirtschaftliche Information aus, wie sie auch an anderen Stellen der Zeitung, etwa der Kurstabelle, zu finden ist (siehe mehr dazu im Abschnitt „Herausforderungen“). Das Herausfiltern der Sätze mit numerischen Werten soll sicherstellen, dass unser Sentiment Index nicht durch solche Informationsaussagen verzerrt wird. Weiterhin wurden sehr kurze (kürzer als 30 Zeichen) und sehr lange Sätze (länger als 800 Zeichen) sowie Überschriften entfernt. Alle verbleibenden Sätze wurden dann nach dem vorgestellten Zero-Shot-Ansatz klassifiziert und einem Sentiment Score zugewiesen (positiv: 1, negativ: -1, neutral: 0), der letztlich über alle Sätze eines Berichts zu einem Gesamt-Sentiment-Score gemittelt wurde.

Um die Qualität der Zero-Shot-Klassifizierung anekdotisch zu evaluieren, wurden 150 Sätze als Gold Standard durch einen Domänenexperten annotiert. Auf dieser Basis ergab sich eine korrekte Klassifikation des automatischen Ansatzes in 74% der Fälle. Von den 26% falsch erkannten Sätzen entfällt der Großteil auf solche, die fälschlicherweise als positiv bewertet wurden, obwohl sie eigentlich als neutral einzustufen sind. Tabelle 2 fasst verschiedene Metriken zur Evaluation des Zero-Shot-Ansatzes sowie die Verteilung der Klassen zusammen.

Tabelle 2: Evaluationsmetriken und Klassenverteilung des Zero-Shot Klassifikators

	Negativ	Neutral	Positiv
Precision	83.67%	36.84%	82.35%
Recall	70.69%	46.67%	83.17%
F1-score	76.64%	41.18%	82.76%
Distribution	30.69%	15.87%	53.44%

In einer ersten Voranalyse des Korpus (siehe Abb. 1) ergeben sich die nachfolgenden Zeitreihen, die jeweils die diachronen Sentiment-Werte für ein Gleitfenster von 30 und 365 Tagen darstellt. Eine Darstellung der Sentimentausschläge auf Tagesbasis wurde verworfen, da diese starken Fluktuationen unterworfen ist und damit nur schwer interpretierbar ist.

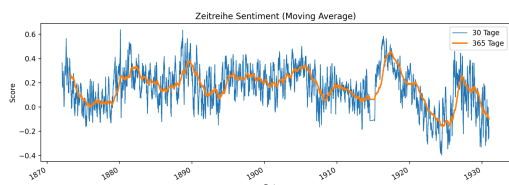


Abbildung 1: Zeitreihe Sentiment über den gesamten Zeitraum des Korpus (Moving Average).

In der geglätteten Darstellung auf Monats- (blau) und Jahresbasis (orange) zeichnet sich dagegen eine Entwicklung mit Phasen unterschiedlicher Stimmungslagen ab, die zumindest auf den ersten Blick plausibel erscheint.⁸

Beispielsweise ist die Spekulationsblase der frühen 1870er Jahre mit einer sehr positiven Stimmung verbunden, die sich nach dem Börsenkrach von 1873 rasch eintrübt. Das gleiche Schema ergibt sich auch für andere Phasen rasch steigender Kurse, wie etwa in den Jahren 1888–90 und 1926/27. Auch der massive Einsturz der Kurse während des Ersten Weltkriegs ist mit einem starken, wenn auch zeitlich verzögerten Rückgang des Sentiment-Index verbunden. Gerade der Anstieg des Sentiment-Index zu Beginn des Ersten Weltkriegs wirft jedoch Fragen auf. Zum einen stellt sich an dieser Stelle die technische Frage, wie belastbar die Ergebnisse angesichts einer Fehlerrate von 26% sind; diesen Punkt führen wir weiter unten aus. Zum anderen zeigt dieser Aspekt inhaltliche Fragestellungen auf: In welchem Verhältnis stehen Stimmung und Kursentwicklung zueinander? Sind beispielsweise steigende (fallende) Kurse immer mit einer positiven (negativen) Stimmung verbunden? Welche Trendwenden lassen sich in der Börsenstimmung identifizieren? Und welche historischen Ereignisse und Entwicklungen lassen sich als Ursachen für veränderte Stimmungslagen festmachen?

Herausforderungen

Unsere aktuellen Experimente zeigen deutlich, dass – eingedenk der eingangs beschriebenen Korpus-Probleme – die Anwendung eines Zero-Shot-Verfahrens auf dem Korpus von Marktberichten bereits vielversprechende Ergebnisse liefert. Gleichzeitig ist die derzeitige Fehlerrate von 26% zu hoch, um belastbare Aussagen treffen zu können. Zudem ist die Klassifizierung der Aussagen in nur eine Kategorie mit drei Ausprägungen angesichts der in den Börsenberichten enthaltenen Aussagen nicht unproblematisch, da sie nur ein sehr grobkörniges und, was schwerer wiegt, potentiell verzerrtes Bild liefert. Konkret scheint uns die Berücksichtigung zweier weiterer Kategorien geboten, die von den uns bekannten Stu-

dien zu Finanzmarktsentiment allerdings ausgeblendet werden. Grundsätzlich beziehen sich die einzelnen Aussagen eines BBZ-Berichts auf drei verschiedene Ebenen. Aussagen wie „Das Aussehen der Börse war heute überaus unfreundlich“ beziehen sich auf die gesamte Börse, andere wie „Schantung-Aktien setzten nach niedrigerem Beginn eine Besserung durch“ dagegen auf einzelne Wertpapiere. Dazwischen rangieren Aussagen wie „Elektrische Werte blieben behauptet“, die sich auf einen bestimmten Teilbereich des Börsenhandels beziehen. Damit ein Sentiment Score die Gesamtstimmungslage an der Börse korrekt widerspiegelt, muss eine Gewichtung dieser Aussagentypen erfolgen, da ansonsten die Gefahr einer Verzerrung der ermittelten Tonalität besteht.

Die Aussagen unterscheiden sich zudem in einer weiteren Dimension. Beispiele wie „Auf dem Rentenmarkte hat sich die Stimmung für Italiener auch heute nicht gebessert“ beschreiben die Börsenstimmung selbst, während Aussagen wie „Die rheinisch-westfälischen Bahnen büßten durchschnittlich nur 3 pCt. ein“ zwar auch Tonalität enthalten, allerdings eher eine (positive, neutrale oder negative) finanzwirtschaftliche Information wiedergeben. Während beispielsweise Takala et al. (2014) solchen Aussagen ein positives Sentiment bescheinigen, würden wir sie als Informationsaussagen mit positiver Tonalität definieren, da es aus unserer Sicht zwischen Sentiment im Sinne der Stimmung an der Börse und Sentiment als Ausdruck einer Informationsaussage zu differenzieren gilt. Beide Unterscheidungen, Aussageebene und Aussagegegenstand, scheinen uns für die Bestimmung eines repräsentativen Sentiment Scores als sehr wichtig. Dementsprechend erfordern die Börsenberichte der BBZ eigentlich eine Aspekt-basierte Sentimentanalyse, bei der neben der Entitätskategorie (Börse, Teilmarkt, Einzeltitel) auch die Aussageart (Informations- vs. Stimmungsaussage) berücksichtigt wird.

Fazit

Die Erkenntnis, dass der Zero-Shot-Ansatz bei einfachen Klassifizierungsaufgaben im Falle einer sehr spezifischen und komplexen Domäne bereits eine hohe Datenqualität liefert, stiftet Zuversicht, da in diesem Fall bereits geringe Mengen von Annotationsdaten ausreichen, um die Datenqualität zu evaluieren und gegebenenfalls durch entsprechendes Nachtrainieren zu erhöhen. Nun stellt sich die weitergehende Frage, inwieweit dieser Ansatz auch für komplexere Aufgaben wie Aspect-Based-Sentiment geeignet ist. Erste Analysen sowie die Arbeit von Shu et al. (2022) stimmen hier optimistisch. Dies wäre insofern eine relevante Erkenntnis, als die meisten praxisbezogenen Aufgaben eher komplexer Natur sind. Komplexere Aufgaben lassen sich im Bereich des maschinellen Lernens häufig durch eine größere Menge an annotierten Trainingsdaten lösen. Ein Vorteil des Zero-Shot Ansatzes läge hier darin, dass das Verfahren auf den Zieldaten zunächst direkt evaluiert werden kann und nur im Bedarfsfall weitere Daten für das Nachtrainieren manuell erstellt werden müssen. Insgesamt scheint dieser neuartige Ansatz aus dem Bereich des Transfer Learning also sehr vielversprechend, auch für sehr heterogene Textkorpora, wie sie in den Digital Humanities häufig vorliegen.

Fußnoten

1. Siehe Raimondo (2019) für einen Literaturüberblick.
2. Siehe Projekt-Homepage <https://media-sentiment.uni-leipzig.de>. Während dieses Zeitraums entwickelte sich die Berliner Börse zu einem international bedeutsamen Handelsplatz (Pohl 2002; Buchner 2019).
3. Zur Geschichte der BBZ, siehe die Beiträge in Bertkau (1930).
4. Siehe <https://zefys.staatsbibliothek-berlin.de/list/>.
5. Trotz dieser Datenlücken haben wir aufgrund ihrer finanzhistorischen Bedeutung an der BBZ als Untersuchungsobjekt festgehalten, zumal nicht ohne Weiteres zu nachvollziehbar ist, ob die *Vossische Zeitung* nicht ähnliche, zeitlich versetzte Lücken aufweist.
6. Daher sehen wir von einer detaillierten Evaluation des Lexikonansatzes gegenüber dem des Maschinellen Lernens an dieser Stelle ab. Zudem soll im weiteren Projektverlauf (siehe Ausblick) eine Aspekt-basierte Sentiment Analyse umgesetzt werden, die sich aus unserer Sicht wesentlich leichter in einen neuronalen Ansatz integrieren lässt als dies bei lexikonbasierten Ansätzen der Fall ist.
7. <https://huggingface.co/svalabs/gbert-large-zero-shot-nli>.
8. An dieser Stelle liegt ein systematischer Vergleich dieser Reihen mit historischen Finanzmarktdaten nahe. Da ein solcher Vergleich eine Vielzahl methodischer Fragen aufwirft, vertagen wir diesen Schritt aus Platzgründen auf eine künftige Publikation. Natürlich gilt es bei der vorläufigen Interpretation das Risiko eines Confirmation Bias im Hinterkopf zu halten.

Bibliographie

- Akerlof, George A. und Robert J. Shiller. 2009. "Animal Spirits." Princeton.
- Bertkau, Friedrich (Hg.) 1930. "75 Jahre Berliner Börsen-Zeitung." Berlin.
- Buchner, Michael. 2019. "Die Spielregeln der Börse: Institutionen, Kultur und die Grundlagen des Wertpapierhandels in Berlin und London, ca. 1860-1914." Tübingen.
- Brown, Tom, Benjamin Mann, Nick Ryder, et al. 2020. "Language models are few-shot learners." NeurIPS.
- Chang, Ming-Wei, Lev Ratinov, Dan Roth und Vivek Srikumar. 2008. "Importance of Semantic Representation: Dataless Classification." Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence.
- Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk und Veselin Stoyanov. "XNLI: Evaluating Cross-lingual Sentence Representations." Proceedings of the 2018 Conference on Empirical Methods Natural Language Processing.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT (1): 4171-4186.
- Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello und Antoine Doucet. 2021. "Named Entity Recognition and Classification on Historical Documents: A Survey". ACM Computing Survey.
- García, Diego. 2013. "Sentiment during Recessions." The Journal of Finance 68 (3): 1267-1300.
- Hanna, Alan J., John D. Turner und Clive B. Walker. 2020. "News Media and Investor Sentiment during Bull and Bear Markets." The European Journal of Finance 26 (14): 1377-95.
- Kautsch, Jacob. 1912. "Handbuch des Bank- und Börsenwesens für Kaufleute, Industrielle, Kapitalisten, Bankiers und Bankbeamte. Mit besonderer Berücksichtigung Deutscher, Österreichischer und Schweizerischer Verhältnisse und den in Deutschland, Österreich und der Schweiz Geltenden Bank- und Börsengesetzen." Berlin.
- Keynes, John Maynard. 1936. "The General Theory of Employment, Interest and Money." London.
- Krupke, Franz. 1904. "Krupkes Konversations-Lexikon der Börse und des Handels und praktischer Führer für Kapitalisten." Berlin.
- Liebl, Bernhard und Manuel Burghardt. 2020. "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline." Proceedings of the 1st Workshop on Computational Humanities Research (CHR).
- Liu, Bing. 2015. "Sentiment analysis: mining opinions, sentiments, and emotions". New York.
- Loughran, Tim und Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." The Journal of Finance 66 (1): 35-65.
- Malo, Pekka, Ankur Sinha, Pekka Korhonen, Jyrki Walenius und Pyry Takala. 2014. "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts." Journal of the Association for Information Science and Technology 65 (4): 782-96.
- Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev und Dimitar Trajanov. 2020. "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers." IEEE Access 8: 131662-131682.
- Pohl, Hans (Hg.) 2002. "Geschichte des Finanzplatzes Berlin." Frankfurt am Main.
- Raimondo, Carlo. 2019. "The Media and the Financial Markets: A Review." Asia-Pacific Journal of Financial Studies 48 (2): 155-84.
- Shiller, Robert J. 2015. "Irrational Exuberance." Revised and expanded third edition. Princeton.
- Shu, Lei, Hu Xu, Bing Liu und Jiahua Chen. 2022. "Zero-Shot Aspect-Based Sentiment Analysis." arXiv preprint arXiv:2202.01924.
- Takala, Pyry, Pekka Malo, Ankur Sinha und Oskar Ahlgren. 2014. "Gold-Standard for Topic-Specific Sentiment Analysis of Economic Texts." Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).
- Tetlock, Paul C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." The Journal of Finance 62 (3): 1139-68.
- van Atteveldt, Wouter, Mariken A. C. G. van der Velden und Mark Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms". Communication Methods and Measures, 15 (2): 121-140.
- Veeranna, Sappadla Prateek, Jinseok Nam, Eneldo Loza Mencía und Johannes Fürnkranz. 2016. "Using Semantic Similarity for Multi-Label Zero-Shot Classification of Text Documents." European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.

Xing, Franz Z., Lorenzo Malandri, Yue Zhang und Erik Cambria. 2020. "Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets." Proceedings of the 28th International Conference on Computational Linguistics, 978–87.

Yin, Wenpeng, Jamaal Hay und Dan Roth. 2019. "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach." EMNLP

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong und Qing He. (2020). "A comprehensive survey on transfer learning." Proceedings of the IEEE, 109(1), 43–76.

Bilder im Kontext: Die Entwicklung des Corpus Vitrearum vom Bildarchiv zu Born- Digitals

Pittroff, Sarah

sarah.pittroff@adwmainz.de

Akademie der Wissenschaften und der Literatur | Mainz,
Deutschland

Gerber, Anja

anja.gerber@bbaw.de

Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Steller, Jonatan

jonatan.steller@adwmainz.de

Akademie der Wissenschaften und der Literatur | Mainz,
Deutschland

Die großen geisteswissenschaftlichen Langzeitvorhaben im deutschen Akademienprogramm umfassen auf Jahrzehnte angelegte Forschungsprojekte, die in der Vergangenheit analog konzipiert wurden und in der Gegenwart vor der Herausforderung stehen, nicht nur ihre Methoden sondern auch ihre Forschungsergebnisse und Publikationsformen einer digitalen Transformation zu unterziehen. Als ursprünglich analoges Langzeitvorhaben, mit nachträglich ergänzter digitaler Komponente, ist das Corpus Vitrearum Medii Aevi (CVMA) Deutschland ein Paradebeispiel für die kontinuierliche Entwicklung von der Buchreihe über das ergänzende digitale Bildarchiv bis hin zum Aufbau von *born-digital* Publikationen. Unter dem Namen 'Glasmalereien im Kontext' verlinken Mitarbeitende des Forschungsvorhabens entkontextualisierte Einzelfotografien von Kirchenfenstern, machen die Fenster in ihrem räumlichen Kontext digital erfahrbar und bereichern sie mit Linked Open Data an. Der vorliegende Beitrag soll das Vorgehen bei der Entwicklung des Formats aufzeigen und argumentiert darüber hinaus, dass solche offenen digitalen Publikations-

formen ein zentraler Beitrag der Digital Humanities zur Weiterentwicklung traditioneller Vorhaben und zukunftsgerichtete Anbindung ihrer Forschungsergebnisse sein können.

Ausgangslage: Das entkontextualisierte Bildarchiv

Das CVMA Deutschland ist ein interakademisches Forschungsvorhaben mit Arbeitsstellen in Freiburg (Akademie der Wissenschaften und Literatur Mainz) sowie Potsdam (Berlin-Brandenburgische Akademie der Wissenschaften). Die Aufgabe des Corpus ist die Erfassung, Dokumentation und Katalogisierung von Glasmalereien des Mittelalters und der Frühen Neuzeit (bis etwa 1550) in Deutschland. Das CVMA Deutschland ist Mitglied des 1952 gegründeten Internationalen Corpus Vitrearum, welches sich aufgrund der erheblichen Verluste durch die beiden Weltkriege zum Ziel gesetzt hat, die Glasmalereien des Mittelalters vollständig fotografisch zu erfassen und nach festen Richtlinien zu dokumentieren. Die Forschungsergebnisse werden auf Grundlage der auf internationaler Ebene vereinbarten Richtlinien (Comité international d'histoire de l'art und Union académique internationale 2016) als Katalogtexte in den Corpusbänden publiziert. Dabei entsteht neben einem umfassenden schriftlichen Werk ein umfangreiches kuratiertes Bildarchiv, welches in Deutschland seit 2015 über eine gemeinsame Plattform strukturiert online verfügbar gemacht wird (CVMA Deutschland o. J. a). Mit Metadaten angereicherte Bilddaten dienen als nachnutzbares digitales Instrument für die wissenschaftliche Arbeit.

Die Rolle des Bildes als Repräsentation des Kunstwerkes ist je nach Publikationsform unterschiedlich. Im gedruckten Corpusband präsentiert sich die Abbildung als visuelle Argumentation mittels Gegenüberstellungen und vergleichenden Anbringungen inhaltlich gleichberechtigt zum Text. Beide Kommunikationsformen ergänzen sich zu einer umfassenden Darlegung der wissenschaftlichen Erkenntnisse. Digitale Bildarchive stehen im Vergleich zu klassischen Bucheditionen vor Herausforderungen auf zwei Ebenen: der des potentiellen Nutzer:innenkreises eines über das Internet frei zugänglichen und vernetzten Bildarchivs, der im Gegensatz zu dem oft engen und bildungsbürgerlichen Nutzer:innenkreis der gedruckten Corpusbände steht, wo traditionell Kontextwissen durch die Gesamtschau innerhalb eines Bandes abgedeckt wird. Online-Bildarchive sprechen in der Regel breitere Zielgruppen an, Webseiten bedienen veränderte Lese- und Betrachtungsmethoden (cf. Hyman, Moser und Segala 2014, 37). Herausforderungen stellen sich genauso auf der Ebene der nun vom Werk als auch vom Textkontext entfremdeten Abbildung, die ihre ästhetische wie wissenschaftliche Bedeutung im Online-Bildarchiv neu manifestieren muss.

Das Online-Bildarchiv des CVMA Deutschland (CVMA Deutschland o. J. a) versammelt bereits mehr als 7.000 Abbildungen aus der Buchreihe sowie zusätzliches, bisher unveröffentlichtes Material und wird von den beiden Arbeitsstellen konstant um weitere Bestände erweitert. Dabei stehen Detailaufnahmen einzelner Scheiben, zugehörige Fensterdarstellungen oder Montagen kom-

plexer Bilderzählungen nicht zwangsläufig nebeneinander. Die Sortierung der Galerieansicht orientiert sich zunächst an einer Datenlogik, nicht an ihrem Inhalt. Über diesen Zugang fehlen übergreifende Informationen zum Nachvollziehen der Zusammenhänge zwischen Einzelbildern und ihrer wissenschaftlichen Einordnung. Der Workflow einer Veröffentlichung über das Online-Bildarchiv lässt die Möglichkeit, Abbildungen semantisch zu Objekten zu gruppieren und sie innerhalb der Architektur bzw. der Räume zu verorten, in denen sie sich befinden, als Desiderat zurück.

Grundlage: Nachhaltiges und interoperables Metadatenmanagement

Grundlage für die Erweiterung und Weiterentwicklung des Online-Bildarchivs auf ein *born-digital* Publikationsformat ist die einheitliche Anlage von Metadaten entlang internationaler Standards. Die Grundsätze des nachhaltigen und interoperablen Metadatenmanagements müssen nach der Langzeitverfügbarkeit der Daten und damit auch nach der sie zugänglich machenden Software und Hardware fragen. Die Antwort zielt dabei auf Standards, die von vielen geteilt, verstanden und weiterentwickelt werden. Auch eine interoperable und nachhaltige Datenmodellierung, die unabhängig von ausführenden und alternden Softwarelösungen lesbar bleiben möchte, basiert auf standardisierten Ausdrücken und generischen Ansätzen der Formulierung von Beziehungen.

Die Integration von Metadaten in die Bilddateien ist ein maßgeblicher Bestandteil der Strategie ihrer Langzeitarchivierung und Auffindbarkeit. Das CVMA orientiert sich dabei am Standard für PDF-Dokumente, der *Extensible Metadata Platform* (XMP), seit 2012 ISO Standard ISO 16684-1 (ISO 2019). Die Metadaten werden hierbei in den Header der Bilddateien geschrieben, so dass die digitale Ressource und ihre Metadaten eine Einheit bilden. Die CVMA-XMP-Metadatenspezifikation (CVMA Deutschland 2016) vereint administrative, beschreibende und technische Metadaten. Sie nutzt weit verbreitete Standards nach, z. B. Dublin Core, IPTC und wurde um einen CVMA-spezifischen *namespace* erweitert, mit dem die genuin glasmalereispezifischen Informationen erfasst werden. Die Bilddateien enthalten auf diese Weise ihre relevante Metadaten und können durch Dritte heruntergeladen werden, wobei die Einheit der digitalen Fotografie und ihrer Metadaten eine umfassende Nachnutzbarkeit über die visuelle Information hinaus gewährleistet.

Die Auszeichnung der Abbildungen mit Normdaten nutzt nicht nur der Verlinkung von Informationen mit außenstehenden Repositorien, sondern dient auch der internen Strukturierung des Bildarchivs. So werden prospektiv alle Normdateneinträge, zur Zeit zumindest die IconClass Notationen sowie die Personendaten, neben den eigentlichen Bilddateien als Ressourcen des Archivs betrachtet und prozessiert. Jede Ressource erhält durch die automatisierte Vergabe von Uniform Resource Identifier (URI) eine CVMA-interne ID. Diese erlaubt eine eindeutige Referenzierung innerhalb des Bildarchivs und

auch nach außen. Über unterschiedliche Serialisierungen können diese Ressourcen für verschiedene Bedarfe ausgegeben werden. JSON-, RDF- und TTL-Formate erlauben eine Einbindung der Ressourcen in weitere Datenkontexte, die HTML-Serialisierung dient in erster Linie einer visuellen Erfassung der Ressourcen über den Browser. Für die Bilddaten ist dies die Einzelansicht der Abbildung im Online-Bildarchiv mit einer Auswahl an relevanten Metadaten, die über Kartenansicht und begleitender Textanzeige ein grundlegendes Informationsset zum dargestellten Werk bietet. Für Iconclass-Notationen umfasst dies die Darstellung aller Bildressourcen aus dem Online-Bildarchiv, die mit jener spezifischen Notation ausgezeichnet sind.

Erweiterung: Das Cultural Heritage Framework

Das Ziel der Strukturierung des Bildarchivs lautet aber nicht, weitere Sammlungen zu erzeugen, sondern die Daten durch eine komplexere semantische Verknüpfung zu modellieren. Dafür kommt das an der Digitalen Akademie in Mainz entwickelte Cultural Heritage Framework, kurz CHF, zum Einsatz (Schrade 2017).

Das Framework dient unterschiedlichen Projekten, die mit Objekterfassung im Bereich des kulturellen Erbes befasst sind, dazu, Digitalisate unterschiedlicher Provenienz zu einem Kulturerbeobjekt zu verknüpfen, mit (wissenschaftlichen) Texten anzureichern und als *born-digital* Onlinepublikation zu veröffentlichen.

Das Datenmodell besteht aus mehreren Komponenten, die auf den jeweiligen Erfassungsbedarf des Projektes angepasst werden: 'Artefakte', 'Entitäten', 'Ereignisse', 'Personen' und 'Orte'. Artefakte können dabei mit den digitalen institutionellen Sammlungen, zum Beispiel Ressourcen aus dem CVMA Online-Bildarchiv, verknüpft werden. Datierungen, geographische Angaben wie Koordinaten, Ortsangaben oder *Geonames Identifier* und Stichwörter werden zwischen den entsprechenden Datenbanken, die diese Informationen speichern, sowie dem Datenmodell ausgetauscht. Es organisiert textuelle und nicht-textuelle Forschungsdaten.

Für das CVMA wurde hieraus das auf der Webseite verfügbar gemachte Modul "Glasmalerei im Kontext" entwickelt. Das Publikationsmodell kann aus drei Perspektiven betrachtet werden.

Aus der Sicht des Datenmodells strukturiert es Forschungsdaten zu Bauwerken, meist Kirchen, hierarchisch: vom Standort zum Gebäude, zum Gebäudeteil, zum Fenster, zur Einzelscheibe bzw. Montage. Spezielle Entitäten können mit weiteren spezifischen Ressourcen verknüpft werden. Die oberste Hierarchieebene sieht die Einbindung eines Grundrisses vor, während die unterste Hierarchieebene eine direkte Verlinkung zu den Bildressourcen des Online-Bildarchivs erlaubt. Auf allen Ebenen der Fenster- und Scheibendarstellungen wird eine Gegenüberstellung der Glasmalereien und ihrer Erhaltungsschemata ermöglicht, um mittelalterlich erhaltene Bestandteile und die Anteile möglicher Restaurierungsmaßnahmen visuell nachzuvollziehen. Damit wird das Publikationsformat zu einem wertvollen Forschungsinstrument.

Aus der Sicht der Abbildungen bietet das Modul 'Glasmalereien im Kontext' die Möglichkeit zur semantischen Strukturierung und kontextualisierten Darstellung der im Online-Bildarchiv publizierten Bildressourcen. So bietet sich dem Nutzer des Bildarchivs bei jeder über das Modul verlinkten Ressource automatisch der Weg in die Standarddarstellung. Eine Datenmodellierung über das Content Management System und dem dort aufgesetzten CHF wirkt sich also strukturierend auf das Online-Bildarchiv aus.

Aus der Sicht der Nutzenden schließlich eröffnet sich ein niedrigschwelliger Zugang nicht nur zu einzelnen Forschungsressourcen, wie im Bildarchiv, sondern auch zu den Forschungsergebnissen, die in Form von Bildmaterial und kurzen Texten als Microsites veröffentlicht sind. Hier gibt es in übersichtlichen Schritten immer tiefergehende Informationen von der Baugeschichte des Kirchenstandortes bis zu den Vorbildern einzelner in den Glasmalereien dargestellter Szenen.

Ergebnis: Glasmalereien im Kontext – *Born-Digitals* als semantische Datenmodellierung, offene digitale Publikationsform und Open Educational Resource

'Glasmalereien im Kontext' ist eine Kombination aus semantischer Datenmodellierung und deren Ausspielung als digitale Publikationsform (CVMA Deutschland o. J. b). Aufgeschlüsselt nach Standorten können Nutzer:innen hier Kirchenräume und ihren mittelalterlichen Glasmalereibestand über Text- und Bildmaterial erkunden. Die Notwendigkeit, Struktur in die entkontextualisierte vorliegende Bildsammlung zu bringen, hat im Ergebnis eigenständige *born-digital* Publikationen aus vormals gedruckten Corpusbänden entstehen lassen, die sich in konservativer Modellierung an der Struktur der Printpublikationen orientieren und darüber hinaus auch innovative Modellierungen zulassen, die als niedrigschwellige aber hochwertige, aktualisierte und non-lineare Ergänzungen zu den gedruckten Werken veröffentlicht werden.

Für die nah am Corpusband formulierte Modellierung werden ausgewählte Standorte mit überarbeiteten und reduzierten Katalogtexten zu den jeweiligen Gesamtfenstern, Szenen und Einzelscheiben angeboten. Eine Kirche wird in Objektgruppen untergliedert, bei denen es sich um Gebäudeteile wie Chor oder Langhaus handelt. Die darunter liegenden Ebenen bilden die dort enthaltenen Objekte. Dies sind die Buntglasfenster, die in ihrer Gesamtdarstellung als Bildmontagen und im Detail als Einzelscheiben dargestellt sind. Beide Entitäten sind im Objekt enthaltene Objektteile.

Während eines mehrere Jahrzehnte umfassenden Forschungszeitraums ergeben sich unter Umständen für bereits bearbeitete Objekte neue Erkenntnisse oder Forschungsmeinungen, die der Revision bedürfen. Das born-digital Format 'Glasmalereien im Kontext' bietet die Möglichkeit, bereits gedruckte Katalogtexte und -abbil-

dungen in den Corpusbänden zu ergänzen und zu erweitern, auf die vorangegangenen Texte Bezug zu nehmen oder von ihnen Abstand zu gewinnen. So können bspw. neue Erkenntnisse zu bereits vor geraumer Zeit publizierten Forschungen einfließen und neu aufgedeckte Standorte veröffentlicht werden. Hier zeigt sich die einzigartige Möglichkeit, aktuelle Forschungen zitierfähig und nachnutzbar zu publizieren.

Darüber hinaus bietet das generische Datenmodell aber auch die Möglichkeit, alternative Rekonstruktionsvorschläge für einen architektonischen Teilraum, einer Objektgruppe im CHF, vorzulegen. Hier werden Abbildungen nicht in ihrer aktuellen Anbringung im Kirchenraum kontextualisiert, sondern ihre vermutete Anbringungen zu unterschiedlichen Zeitpunkten in der Vergangenheit für einen Ort modelliert. Diese Nebeneinanderstellung unterschiedlicher Zustände bzw. Rekonstruktionen stellt einen innovativen Ansatz in der Forschung der mittelalterlichen Glasmalerei dar, der insbesondere in der universitären Lehre in Seminaren Anwendung findet.

Dabei ist das CHF in seiner Form als 'Glasmalereien im Kontext' zugleich Modellierungsumgebung und semantisches Tool und wird mit der Verfügbarmachung der Ergebnisse über das Internet als digitale Publikationsform sofort zur Open Educational Resource.

Bibliographie

Comité international d'histoire de l'art und Union académique internationale. 2016. „Corpus Vitrearum: Richtlinien“. *CVMA Freiburg*. Vers. 4. Troyes. <http://cvi.cvma-freiburg.de/documents/CVRichtlinienEdition.pdf>.

CVMA Deutschland. 2016. „XMP Metadatenspezifikation“. *CVMA Deutschland*. Vers. 1.1, 8. Juni 2016. <https://corpusvitrearum.de/cvma-digital/spezifikationen/cvma-xmp/11.html>.

——. o. J. a. „Bildarchiv“. CVMA Deutschland. Letzter Zugriff am 2. August 2022. <https://corpusvitrearum.de/bildarchiv>.

——. o. J. b. „Mittelalterliche Glasmalereien im Kontext“. CVMA Deutschland. Letzter Zugriff am 2. August 2022. <https://corpusvitrearum.de/glasmalerei-im-kontext>.

Hyman, Jack A., Mary T. Moser und Laura N. Segala. 2014. „Electronic Reading and Digital Library Technologies: Understanding Learner Expectation and Usage Intent for Mobile Learning“. *Educational Technology Research and Development* 62 (1): 35–52. <https://doi.org/10.1007/s11423-013-9330-5>.

IconClass. o. J. „IconClass Illustrated Edition“. *IconClass*. Letzter Zugriff am 2. August 2022. <https://icon-class.org>.

ISO (International Organization for Standardization). 2019. *Graphic Technology – Extensible Metadata Platform (XMP) – Part 1: Data Model, Serialization and Core Properties*. ISO 16684-1:2019, 2, April 2019. Paris: ISO. <https://www.iso.org/standard/75163.html>.

Schrade, Torsten. 2017. „Sammlungs- und Editionsportale mit dem Cultural Heritage Framework der Digitalen Akademie: Ein Werkstattbericht“. Vortrag beim Workshop *Editionsportale*, Friedrich-Schiller-Universität Jena, 3. August 2017. <https://digicademy.github.io/2017-editionsportale-jena>.

The Getty. 2017. „Art & Architecture Thesaurus“. *The Getty Research Institute*. 7. März 2017. <https://www.getty.edu/research/tools/vocabularies/aat/>.

Bullingers Briefwechsel zugänglich machen: Stand der Handschriftenerkennung

Ströbel, Phillip

pstroebel@cl.uzh.ch
Computer Linguistik, Universität Zürich

Hodel, Tobias

tobias.hodel@unibe.ch
Walter Benjamin Kolleg, Universität Bern, Schweiz

Fischer, Andreas

andreas.fischer@unifr.ch
Institute of Complex Systems (iCoSys), Haute école d'ingénierie et d'architecture de Fribourg, Haute école spécialisée de Suisse occidentale

Scius, Anna

anna.scius-bertrand@hefr.ch
Institute of Complex Systems (iCoSys), Haute école d'ingénierie et d'architecture de Fribourg, Haute école spécialisée de Suisse occidentale

Wolf, Beat

beat.wolf@hefr.ch
Institute of Complex Systems (iCoSys), Haute école d'ingénierie et d'architecture de Fribourg, Haute école spécialisée de Suisse occidentale

Janka, Anna

anna.janka@unibe.ch
Walter Benjamin Kolleg, Universität Bern, Schweiz

Widmer, Jonas

jonas.widmer@unibe.ch
Walter Benjamin Kolleg, Universität Bern, Schweiz

Scheurer, Patricia

patricia.scheurer@uzh.ch
Computer Linguistik, Universität Zürich

Volk, Martin

volk@cl.uzh.ch
Computer Linguistik, Universität Zürich

Automatisierte Handschriftenerkennung fokussierte in der Vergangenheit vorwiegend auf Training und Erkennung einzelner Handschriften, die individuell trainiert wurden. Typischerweise finden sich in den Geisteswissenschaften jedoch Datensätze, die von mehreren Händen und häufig auch über längere Zeiträume verschriftlicht wurden. Aussagen zu Erkennungsalgorithmen müssen entsprechende Voraussetzungen ernst nehmen und nicht nur die Fähigkeit nachweisen, einzelne Hände mit hoher Qualität zu erkennen, sondern auch mit ähnlichen jedoch nicht identischen Handschriften umzugehen.

Ein umfangreicher frühneuzeitlicher Briefwechsel aus dem 16. Jahrhundert, der hauptsächlich in zwei Sprachen (Latein und Frühneuhochdeutsch) vorliegt und mehrere hundert Hände umfasst, ist diesbezüglich ein interessanter Vergleichsfall, um bestehende Plattformen mit neuen Möglichkeiten der *Data Augmentation* und dem Einbezug von Transformer-basierten neuronalen Netzen zu vergleichen. Wir nutzen dafür die im Rahmen des Projekts *Bullinger Digital*¹ erarbeiteten Daten, um Aussagen über Erkennungsqualität sowie Potenziale in der automatischen Erkennung zu machen. Für die Evaluation verwenden wir zwei spezifisch erstellte Testsets. Diese widerspiegeln die Tatsache, dass von einigen (wenigen) Personen eine Vielzahl von Briefen und von diversen Personen nur wenige Schreiben im Datensatz vorhanden sind (Abbildung 1). Die beiden Testsets sind zwar ähnlich groß (1'235 Zeilen) als auch Wenigschreiber (1'013 Zeilen), leisten aber Aussagen zu sehr unterschiedlichen Größenordnungen, da von den Vielschreibern eine weit umfangreichere Masse erkannt werden wird, aber auch mehr Material zum Training von Modellen zur Verfügung steht. Bei der Evaluation können wir aufgrund der Aufteilung aber präzisere Voraussagen machen, welche Datenmassen mit welcher Qualität erkannt werden. Weiter lässt sich abschätzen, welche Erkennungsform sich für welche (Trainings-)Datenmenge eignet.

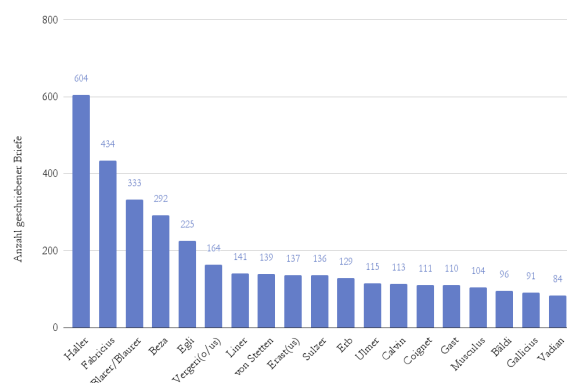


Abbildung 1. Verteilung der Top-20-Autoren, welche Briefe an Bullinger schrieben (Anzahl Briefe auf y-Achse).

Im Folgenden testen wir drei Ansätze, die Aufschlüsse zum Stand und möglichen Entwicklungen im Bereich der Handschriftenerkennung versprechen, am Korpus:

1. *Transkribus* mit seinen zwei Engines (HTR+ (Strauss et al. 2018) und PyLaia (Puigcerver 2017)) testet eine etablierte Methode am Datensatz.

2. *Data Augmentation* soll die Erkennung mit der State-of-the-Art-Engine HTR-Flor (de Sousa Neto et al. 2020) verbessern.
3. Neue Transformer-basierte Modelle (Li et al. 2021) sollen auf ihre Tauglichkeit für historische Daten geprüft werden.

Seit 1974 erscheinen in regelmäßigen Abständen Bände der Heinrich-Bullinger-Briefwechsel-Edition, erarbeitet am Institut für Schweizerische Reformationsgeschichte (IRG) der Universität Zürich. Der letzte Band stammt von 2022 und weitere sind aktuell in Arbeit (Bullinger 2022). Die Edition bedient diverse Nutzer*innengruppen, insbesondere Theolog*innen und die Geschichtswissenschaften, was sich etwa an den Editionsgrundsätzen (Bullinger 1973) zeigt, z. B. stillschweigende Auflösung von Abkürzungen, Identifikation von Personen und Orten etc.

Um die Vielzahl der nicht transkribierten Briefe schneller maschinell zu verarbeiten, arbeiten wir im Rahmen des Projekts *Bullinger Digital* an der automatisierten Aufbereitung der Dokumente. Die Texte der bereits edierten Briefe sowie der am IRG provisorisch transkribierten Briefe haben wir mittels des *Text-to-Image-Verfahrens* (Leifert, Labahn, and Sánchez 2020) mit den Bildvorlagen automatisiert aligniert. Damit steht nun ein Korpus von 1'297'908 Token für Training und Validierung zur Verfügung.²

Bei der Zuordnung der Zeilen wurde mit einem *Threshold* agiert, der dafür sorgte, dass nur transkribierte Textteile zugeordnet wurden, bei denen eine relativ hohe Wahrscheinlichkeit der Übereinstimmung errechnet wurde. Dadurch wurden zwar ca. 25-30% aller Zeilen nicht zugeordnet, Fehler in der Layouterkennung und nicht identifizierte Streichungen und Ähnliches führten aber nicht zu falschen Zuordnungen, die wiederum negativen Einfluss auf das Training von Texterkennungsmodellen hätte.

Texterkennung mit etabliertem Framework und Plattform: Transkribus

2015 wurde die Plattform *Transkribus* gelauncht und danach von 2016 bis 2019 *Deep-Learning*-basierte Erweiterungen, insbesondere Layout- und Texterkennungsengines (Muehlberger et al. 2019) implementiert. Die Plattform wird von einer Kooperative betrieben und für diverse Zwecke in der Forschung und durch Erinnerungsinstitutionen eingesetzt. Training von Handschriftenmodellen und Erkennung neuer Seiten erfolgen über ein GUI.

Basierend auf den oben erwähnten Trainings- und Testdaten erzeugten wir mehrere Handschriftenmodelle, dabei testeten wir jeweils beide verfügbaren Engines (HTR+ und PyLaia), um Unterschiede in der Qualität aufzuzeigen. Wir übernahmen die Layouterkennung unverändert.

Die in Transkribus trainierten Modelle (Tabelle 1) zeigen gute Ergebnisse. Gerade für Vielschreiber wird trotz vieler unterschiedlicher Hände insbesondere mit HTR+ eine Erkennung unter 8% *Character Error Rate* erreicht.³ Die

Erkennung der Wenigschreiber ist dagegen etwas fehlerbehafteter. Alle Modelle wurden "austraining", ohne dass ein problematisches *Overfitting* beobachtet wurde (Hodel 2020).

Eine signifikante Verbesserung erreichten wir durch die Nutzung von grossen Modellen als *Basemodels*. Mit einem vortrainierten Modell basierend auf 5'820'990 Wörtern (im Sinne von *Tokens*) in lateinischer Schrift (keine Kurrentschrift), kann die Erkennqualität weiter verbessert werden.

Tabelle 1. Resultate der sprachunabhängig trainierten Modelle in Transkribus mit HTR+ und PyLaia. HTR+: 500 Epochen, PyLaia: 250 Epochen.

		CER			
		Vielschreiber	Wenigschreiber		
Trainingsdaten	base model	HTR+	PyLaia	HTR+	PyLaia
multilingual	-	7.26	9.9	10.06	12.7
	Latin	6.79	-	9.51	-

Data Augmentation zur Erweiterung des Trainingsmaterials

Um für einen gegebenen Schreibstil ein zuverlässiges Erkennungssystem zu trainieren, braucht es eine große Anzahl annotierter Textzeilen. Es ist deshalb schwierig, seltene Hände automatisch zu transkribieren, da der entsprechende Schreibstil nicht oder nur ungenügend in den Trainingsdaten repräsentiert ist. *Data Augmentation* kann verwendet werden, um die Trainingsdaten automatisch mit weiteren Beispielen anzureichern und so den Lernprozess zu unterstützen. In unserer Arbeit verfolgen wir dazu einen vielversprechenden Ansatz: Wir lernen die Schreibstile von seltenen Händen mittels *Generative Adversarial Networks (GANs)*, um anschließend beliebige Texte zu synthetisieren und den Trainingsdaten als zusätzliche Lernbeispiele hinzuzufügen.

Für die Synthetisierung wird *lineGen* (Davis et al. 2020) eingesetzt, ein kürzlich vorgeschlagenes GAN-Netzwerk, welches auf ganzen Textzeilen arbeitet und drei Zielfunktionen integriert: den *Adversarial Loss* (Unterscheidung zwischen echten und synthetischen Bildern), den *Perceptual Loss* (visuelle Qualität der generierten Bilder) und den *CTC Loss* (Qualität der automatischen Transkription). Das trainierte lineGen-System erlaubt es, einen beliebigen Text mit dem gelernten Stil zu synthetisieren, d. h. aus dem Text ein Textzeilenbild zu generieren. Abbildung 2 zeigt Beispiele von generierten Textzeilenbildern, wenn lineGen auf rund 17'000 Textzeilenbildern unterschiedlich lange und unter Verwendung der Standard-Hyperparameter mit verschiedenen Schreibstilen trainiert wird und im Anschluss ein englischer Text synthetisiert wird. Das GAN konvergiert auf einen Schreibstil, der ähnlich leserlich ist wie echte Beispiele aus den Trainingsdaten.⁴

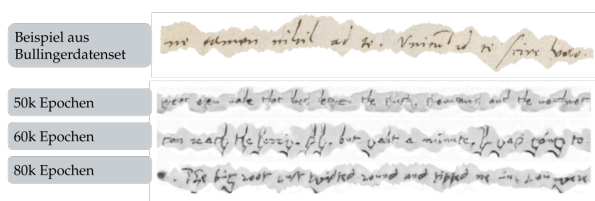


Abbildung 2. Beispiele von synthetischen Textzeilenbildern basierend auf lineGen über unterschiedliche Anzahl Epochen.

Der Einfluss der synthetischen Lernbeispiele auf die Erkennungsrate wurde in zwei Szenarien untersucht, welche eine kleine Trainingsmenge vorsehen, wie es für Wenigschreiber typisch ist: In Szenario A gehen wir von 1'000 Trainingszeilen aus und in Szenario B von 200 Trainingszeilen. Als Erkennungssystem wird HTR-Flor (de Sousa Neto et al. 2020) eingesetzt, eines der besten Systeme nach aktuellem Stand der Technik, und für die Auswertung werden rund 1'000 zufällig ausgewählte Textzeilen als Testdaten verwendet. Abbildung 3 zeigt den Trainingsverlauf für Szenario A unter Verwendung der Standard-Hyperparameter für HTR-Flor. Die CER auf den Testdaten erreicht 26.8% für das Training mit 1'000 echten Textzeilen. Wenn mit 1'000 synthetischen Textzeilen trainiert wird, scheitert das Lernen. Eine mögliche Interpretation ist, dass die synthetische Schrift nicht genügend natürliche Variation beinhaltet, welche fürs Trainieren nötig ist. Wenn aber die 1'000 echten mit den 1'000 synthetischen Textzeilen kombiniert werden, kann die CER signifikant auf 24.1% verbessert werden. Abbildung 4 zeigt den Trainingsverlauf für Szenario B. Hier scheitert das Lernen sowohl mit 200 echten als auch mit 200 synthetischen Textzeilen. Hingegen führt die Kombination von echten und synthetischen Trainingszeilen erneut zu einer signifikant verbesserten CER von 47.7%.

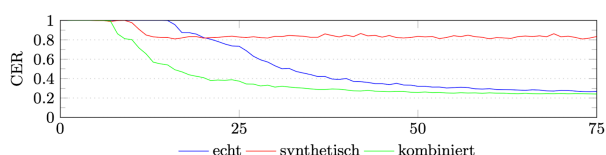


Abbildung 3. Training von HTR-Flor während 75 Epochen im Szenario A (1'000 Trainingszeilen).

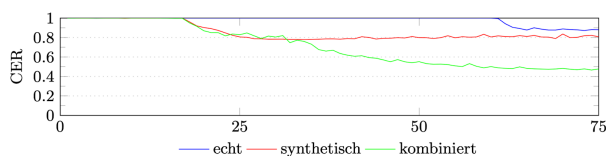


Abbildung 4. Training von HTR-Flor während 75 Epochen im Szenario B (200 Trainingszeilen).

Diese initialen experimentellen Resultate sind vielversprechend und legen die Vermutung nahe, dass GAN-basierte synthetische Lernbeispiele dabei helfen können, die Erkennung für Wenigschreiber zu verbessern. In einem nächsten Schritt planen wir umfangreichere Experimente, welche darauf abzielen, den Stil von Wenigschreibern zu lernen oder den Stil von Gruppen ähnlicher

Schriftbilder. Eine wichtige Fragestellung dabei wird sein, wie viele Lernbeispiele nötig sind, um einen Stil zuverlässig lernen zu können. Anschließend planen wir, ein *Transfer Learning* durchzuführen, ausgehend von einem gut trainierten Grundsystem für Vielschreiber, welches mit Hilfe der synthetischen Lernbeispiele an die Wenigschreiber angepasst wird.

Nutzung einer Transformer-basierten Architektur

Transformer-basierte Architekturen (Vaswani et al. 2017) haben das Feld der natürlichen Sprachverarbeitung in Sachen Sprachmodellierung revolutioniert und zu einem Paradigmenwechsel beim Trainieren von Systemen für unterschiedlichste Anwendungen geführt. Transformer-Modelle wie BERT (Devlin et al. 2018) und deren Weiterentwicklungen, welche auf großen Mengen an Sprachdaten trainiert wurden, sind als starke Transfer-Lerner (Ruder et al. 2019) bekannt und erlauben einen vielfältigen Einsatz beim Fine-Tuning für spezifische Aufgaben (*Natural Language Understanding*, Fragenbeantwortung, Wortarten-Klassifikation). Transformer können auch für Bildverarbeitung genutzt werden (Dosovitskiy et al. 2021; Touvron et al. 2021), was die Entwicklung von BERT-ähnlichen und auf großen Bildmengen vortrainierten Modellen zur Folge hatte (Bao, Dong, and Wei 2021).

Für unser Projekt nutzen wir als Grundlage TrOCR (Li et al. 2021), welchem ein *Encoder-Decoder*-Struktur zugrunde liegt. Bildseitig beruht der Encoder auf der BEiT-Architektur (Bao, Dong, and Wei 2021), welche für die *Feature*-Extraktion aus den Bildern verantwortlich ist, während der *Decoder* die Bildinformation mit Hilfe eines RoBERTa-Sprachmodells (Liu et al. 2019) in eine *Subword*-Folge "übersetzt". Li et al. benutzten in zwei *Pre-Training*-Phasen über 700 Millionen an synthetisch erzeugten und echten Textzeilenbildern mit dem dazugehörigen Text in englischer Sprache, die danach mit dem IAM-Datenset (Marti and Bunke 2002) einem *finetuning* unterzogen wurden. Die CER von 2.89% liegt nur 0.14 Prozentpunkte hinter einem klassischen Ansatz (Diaz et al. 2021).

Im Gegensatz zu den beiden anderen Projektteilen wurde für die TrOCR-Anwendung die Zeilen nach Sprachen unterschieden. Diese Sprachverteilung und einige weitere Kennzahlen können Tabelle 2 entnommen werden.

Tabelle 2. Kennzahlen der extrahierten Zeilen aus dem Material der Bullinger-Briefe.

	# Zeilen	%	# Worte	# Wörter / Zeile	# Buchstaben	# Buchstaben / Zeile
Latein	134'236	81.02	1'073'106	7.99	7'314'648	54.49
FNHD	31'437	18.98	253'372	8.06	1'547'725	49.23
Total	165'673		1'326'478	8.01	8'862'373	53.49

Wir untersuchen die Eignung von TrOCR für die Texterkennung auf historischen Daten, indem wir die Anzahl Epochen, die für das Fine-Tuning aufgewendet werden, variieren und multi- wie auch monolinguale Modelle trainieren. Alle Modelle führen während eines Drittels der Epochenzahl ein *Warm-Up* durch.⁵ Für die Evaluation ha-

ben wir auch die Testsets nach Sprachen aufgeteilt. Tabelle 3 fasst die Ergebnisse zusammen.

Tabelle 3. Resultate der Modelle trainiert auf verschiedenen Sprachzusammenstellungen und Anzahl Epochen (fett = beste Performance in einer Spalte).

		CER					
		Vielschreiber	Wenigschreiber				
Trainingsdaten	# Epochen	multiling.	Latein	FNHD	multiling.	Latein	FNHD
multilingual	1	9.53	9.49	9.7	11.58	11.07	12.79
	2	8.38	8.51	7.71	10.46	10.18	11.11
	3	7.81	8.07	6.5	9.95	9.61	10.74
	4	7.21	7.61	5.27	9.68	9.31	10.55
	5	7.31	7.85	5.25	9.69	9.54	10.25
	6	7.41	7.89	5.09	9.61	9.39	10.11
Latein	5	11.09	7.99	26.21	16.1	9.69	31.21
FNHD	5	28.41	33.02	6.06	27.89	34.83	11.43

Aufgrund der Resultate entschieden wir uns, die monolingualen Modelle auf 5 Epochen zu trainieren. Auf die frühneuhochdeutschen Zeilen der Vielschreiber angewandt liefert das monolinguale Modell mit einer CER von 6.06% entgegen der Erwartungen nicht das beste Resultat. Dieses erreichte ein multilinguales Modell (trainiert über 6 Epochen) mit einer CER von 5.09%, also knapp einem Prozentpunkt Unterschied. Die größere Menge an Bilddaten beeinflusst während des Fine-Tunings merklich die Performanz.

Unsere Experimente zeigen, dass Transformer-basierte HTR für historische Daten CERs in akzeptablen Bereichen liefert, ohne dass TrOCR bis zu unserem Fine-Latein oder FNHD gesehen hätte. Trotz der im Vergleich zum Pre-Training kleinen lateinischen und frühneuhochdeutschen Textmenge lernt TrOCR die Dekodierung von neuen Sprachen zuverlässig. In Zeilen, in welchen die Sprache hingegen ändert (Code-Switching), was im Bullinger-Briefwechsel relativ häufig passiert (Volk et al. 2022), reagiert TrOCR zu spät für den Sprachwechsel (Abbildung 5). Solche Phänomene werden wir mit flexibleren Modellen abzufangen versuchen.

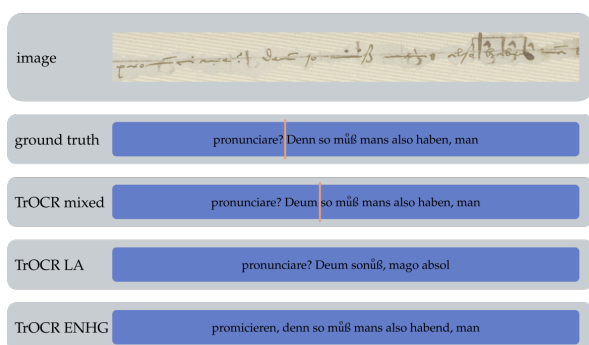


Abbildung 5. Performanz verschiedener TrOCR-Modelle auf einer Zeile, in welcher Sprachwechsel vorliegt. Der Trennstrich markiert die Sprachgrenze, die im multilinguales Modell zu spät und in monolingualen gar nicht erkannt wird.

Schlüsse

Die Arbeit mit dem Bullinger-Korpus ist aufschlussreich in unterschiedlicher Hinsicht. Zentral für diese Arbeit sind drei *Schlussfolgerungen*:

1. Wir stellen eine Harmonisierung unterschiedlicher (*Deep-Learning*-basierter) Systeme mit Bezug zur Erkennungsqualität von Handschriften fest. Unabhängig davon, ob etablierte Plattformen oder neue Systeme wie TrOCR genutzt werden.
2. Der Gebrauch von *Data-Augmentation*-Techniken verspricht einen Gewinn, der aktuell noch weiter auszuloten ist, im Grundsatz aber schon zu Verbesserungen führt.
3. Aktuell haben etablierte Plattformen den Vorteil, dass grosse *Basemodels* genutzt werden können, die noch leichte Vorteile mitbringen, bei grossen Projekten wie Bullinger Digital aber noch anhand der Fehlermuster ausgelotet werden müssen.

Fußnoten

1. <https://www.bullinger-digital.ch/>.
2. Stand 1.6.2022.
3. Character Error Rate wird mit CER abgekürzt und steht für Zeichenfehlerrate.
4. Für den Aufbau des Experiments: Spoto et al. 2022.
5. Warm-up bedeutet, dass zu Beginn des Trainings mit kleinen Lernraten gerechnet wird, um das Modell langsam an die neuen Daten zu gewöhnen.

Bibliographie

Bullinger, Heinrich. 1973. *Briefe der Jahre 1524-1531*. Edited by Ulrich Gäbler and Fritz Büsser. Vol. 1. Heinrich Bullinger Werke. Zweite Abteilung: Briefwechsel. Zürich: Theologischer Verlag.

———. 2022. *Briefe von April bis Dezember 1547: Anhang: Neue Briefe aus den Jahren 1523 bis 1546*. Edited by Reinhard Bodenmann, Yvonne Häfner, and Judith Steiniger. 1st ed. Vol. 20. Heinrich Bullinger Werke. Zweite Abteilung: Briefwechsel. Zürich: Theologischer Verlag Zürich.

Davis, Brian, Bryan Morse, Brian Price, Chris Tensmeyer, Curtis Wigington, and Rajiv Jain. 2020. "Text and Style Conditioned GAN for the Generation of Offline-Handwriting Lines." In *The 31st British Machine Vision (Virtual) Conference 2020*. Bath. https://www.bmvc2020-conference.com/conference/papers/paper_0815.html.

Díaz, Daniel Hernandez, Siyang Qin, Reeve Ingle, Yasuhisa Fujii, and Alessandro Bissacco. 2021. "Rethinking Text Line Recognition Models." arXiv. <http://arxiv.org/abs/2104.07787>.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv. <http://arxiv.org/abs/2010.11929>.

Hodel, Tobias. 2020. "Best-Practices Zur Erkennung Alter Drucke Und Handschriften – Die Nutzung von Transkribus Large- Und Small-Scale." In *DHd 2020. Spielräume Digital Humanities Zwischen Modellierung Und Interpretation*, edited by Christof Schöch, 84–87. Paderborn. <https://doi.org/10.5281/zenodo.3666689>.

Leifert, Gundram, Roger Labahn, and Joan Andreu Sánchez. 2020. "Two Semi-Supervised Training Approaches for Automated Text Recognition." In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 145–50. <https://doi.org/10.1109/ICFHR2020.2020.00036>.

Li, Minghao, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. "TrOCR: Transformer-Based Optical Character Recognition with Pre-Trained Models." arXiv. <http://arxiv.org/abs/2109.10282>.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv. <https://doi.org/10.48550/arXiv.1907.11692>.

Marti, U.-V., and H. Bunke. 2002. "The IAM-Database: An English Sentence Database for Offline Handwriting Recognition." *International Journal on Document Analysis and Recognition* 5 (1): 39–46. <https://doi.org/10.1007/s100320200071>.

Puigcerver, Joan. 2017. "Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?" In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:67–72. <https://doi.org/10.1109/ICDAR.2017.20>.

Ruder, Sebastian, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. "Transfer Learning in Natural Language Processing." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18. Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-5004>.

Sousa Neto, Arthur Flor de, Byron Leite Dantas Bezerra, Alejandro Héctor Toselli, and Estanislau Baptista Lima. 2020. "HTR-Flor: A Deep Learning System for Offline Handwritten Text Recognition." In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 54–61. <https://doi.org/10.1109/SIBGRAPI51738.2020.00016>.

Spoto, Martin, Beat Wolf, Andreas Fischer, and Anna Scius-Bertrand. 2022. "Improving Handwriting Recognition for Historical Documents Using Synthetic Text Lines." In *Proceedings 20th Conf. of the International Graphonomics Society (IGS)*. Las Palmas de Gran Canaria.

Strauss, Tobias, Max Weidemann, Johannes Michael, Gundram Leifert, Tobias Grüning, and Roger Labahn. 2018. "System Description of CITlab's Recognition & Retrieval Engine for ICDAR2017 Competition on Information Extraction in Historical Handwritten Records." *CoRR* abs/1804.09943. <http://arxiv.org/abs/1804.09943>.

Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. "Training Data-Efficient Image Transformers & Distillation through Attention." In *Proceedings of the 38th International Conference on Machine Learning*, 10347–57. PMLR. <https://proceedings.mlr.press/v139/touvron21a.html>.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 30:5998–6008. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.

Volk, Martin, Lukas Fischer, Patricia Scheurer, Raphael Schwitter, Phillip Ströbel, Benjamin Suter. 2022. "Nunc Profana Tractemus. Detecting Code-Switching in a Large Corpus of 16th Century Letters." In *Proceedings of LREC 2022*. Marseille.

Coding editions. Computational approaches to the editing of pre-modern texts.

Cugliana, Elisa

elisa.cugliana@uni-koeln.de

CCeH - Universität zu Köln, Deutschland

Introduction

The endlessness of the digital space allows scholarly editors to conceive edition projects on a grand scale. Indeed, the eternal mission of philology can be identified in the quest for the best way to represent and/or reconstruct primary sources, in order to save them from oblivion and grant the public the most informed access to them. Consequently, the lack of space boundaries characterising the digital environment is one of the first aspects making such an environment perfectly suitable for hosting ambitious edition projects. Considering then the possibilities disclosed by hypertextuality, multimodality and multimediality, one can hardly argue against the digital way of editing, despite the ongoing challenges it must face (cf. for instance Rosselli Del Turco 2016), of which sustainability is clearly in the forefront. However, there are reasons to believe that state-of-the-art digital editions have not yet overcome the limits of a still rather bookish paradigm, obeying a mostly representational logic (Van Zundert 2018, Cugliana and Van Zundert 2022).

In this contribution, I will argue that the key to the next level of digital scholarly editing is to be found in computation, that is, in the actual coding of the whole editorial workflow. I will present the theory, the advantages and the challenges of such a computational approach to the editing of pre-modern texts, backing up my claims with examples from the praxis and from my own scholarly work in the field of digital philology. Of course, the field of computational editing is still in its infancy, which means that fully computational editions have not been published yet.

However, there are projects¹ which can already prove the validity of the arguments presented in this paper or that at least seem to move in the same direction.

Theoretical background

While the digital paradigm has been widely described and commented upon (among others by Stella 2007 and Sahle 2016), there is a crucial aspect that has not been sufficiently underlined, which however could represent a turning point in the history of (digital) philology. It is the case of the use of programming code throughout the different phases of the editorial process, aiming at the realisation of what Barabucci and Fischer (2017) defined as the formalisation of textual criticism. The authors, in their conclusions, state that a “shared formalization would lead to the semi-automatization of the editorial process”, where “the responsibility of the editors would be to describe their choices and decisions”, while that of the computers would be to “deal with applying these rules and decisions in the best way”.

The act of formalising the competence of the editor is to be seen as an achievement *per se*, for it can lead to more accountability and verifiability of the scholarly processes, which can be shared in its complexity with the community. As a matter of fact, such processes are of fundamental importance not only from a methodological perspective, but also for the proper understanding and contextualisation of the presented results. In this respect, Van Zundert (2018) points out how the current digital methods used in scholarly editing, instead, still aim at merely recording the results of the philological work in a representational fashion,² without keeping track of the decisions, analyses and actions that led to those results. In a computational edition, on the other hand, the scholarly competence and actions are not only expressed in a formal language, but they are also operationalised. This, in turn, brings about new affordances, such as the replicability and reusability of the editorial work, as well as the introduction of a degree of indirection allowing for highly controlled modelling of the edition process.

If methods are there to realise theoretical principles in the best possible way, at the same time they can in turn influence the principles themselves, opening up new perspectives uncovering, for instance, implicit biases and illogicalities. Indeed, the use of computation for the modelling and operationalisation of editorial knowledge can contribute to perfecting the editorial workflow in the way envisioned by McCarty (2005), who referred to the “meaningful surprise” often arising from the process of modelling and computing.³

The praxis of the computational method

Some examples from the actual application of this approach will hopefully prove its potential for the field of Digital Scholarly Editing.

Achievements

During my PhD, which I completed in February 2022, I edited an Early High German version of Marco Polo’s travel account, also known as “Version DI” of the *Devisement dou Monde*. One of the main goals of my research was to automatise relevant parts of the editorial workflow, such as the normalisation of the texts. As a matter of fact, bearing in mind the concept of “text as a wheel” of different possible facets, as theorised by Sahle 2013, I wanted to produce (at least) three different levels of normalisation of the text (diplomatic, semi-diplomatic and interpretative).

Together with my colleague Gioele Barabucci (Norwegian University of Science and Technology), we developed a method based on a rule-and-exception principle, featuring three XProc pipelines, one for each level of normalisation (Cugliana and Barabucci 2022). Each pipeline consists of a series of XSLT stylesheets which deal with the different steps of the normalisation process, such as the levelling of allographs, the expansion of abbreviations, the regularisation of capitalisation etc.⁴ In particular, the choice of using a pipeline system instead of a single stylesheet was due to the fact that some actions needed to be taken before other ones could follow: for instance, inserting a capital letter after a full stop implies that the full stops are already in place. Each pipeline generates a new TEI XML file with a different level of normalization.

This proved to be a very suitable strategy for dealing with the complexity of the normalisation process. Concerning the choice of XProc pipelines, it has already been shown that small-step pipelines making use of a stateless language such as XSLT prove to be advantageous in that they reduce the complexity of computer programs, they can be easily shared with the peers and they improve the sustainability of the code (Barabucci and Schaeben 2021). Not only were the pipelines successful in generating three levels of normalisation of the texts, but they could also be applied to the transcriptions of all the witnesses edited, despite the fact that some were written in the East-Swabian dialect, and some in Bavarian. This was probably due to the geographical proximity and the contemporaneity of their production, which leads to hypothesise the possibility of creating “pools of rules” for the editing of witnesses written in specific areas and historical periods.

In my talk I will present the system at the basis of the normalisation of the texts featured in the edition of DI, giving some insights into the very development of the pipelines, both from a strictly philological and from a computational perspective. In particular, I will focus on some tricky aspects such as the cases of ambiguities and exceptions, which might represent a hurdle for the full systematisation of the editorial workflow.

Outlook

A case of reuse of the normalisation pipelines written for Version DI of the *Devisement dou Monde* is the project for a computational edition of the medieval German versions of the *sortes* text *Prenostica Socratis Basilei* (Alonso Guardo 2015). This project is still in its early stages,⁵ but a

taste of its complexity will be given in the talk. With a prototypical example I will show how the same or slightly adapted XSLT stylesheets can be used to normalise texts different from the ones which had served as a basis for their development, as well as how they can be embedded in the editorial workflow for the creation of a fully computational edition. This will open up the discussion on the reuse of program code for different editorial projects (which parts of the code can be reused, which need to be adapted, what rules can apply to different projects and what is instead determined by each individual edition).

Sortes texts (in German “Losbücher”) represent an extremely interesting genre, which unfortunately has not enjoyed much scholarly attention in the field of digital editing, although it has a long history dating back to antiquity.⁶ They were very interactive texts which were not meant to be read linearly (Heiles 2018) and which were used to predict the future with the help of sortition mechanisms, sometimes quite extravagant ones. The computational edition of the medieval German versions of the *Prenostica Socratis Basilei* aims at establishing a new methodology for the editing of this genre, finding strategies to weave together games and textual criticism, with the goal of representing the *sortes* in their genesis and reception throughout the centuries. Specifically, the whole editing process will find expression in an actionable formal language.

Conclusions

As can be evinced from this abstract, my contribution has a twofold purpose: on the one hand giving a short, but hopefully convincing introduction to the theoretical aspects justifying the scholarly value of a computational approach to editing and, on the other, presenting some meaningful results obtained in the praxis. In particular, the focus will mainly be on the success and challenges of the normalisation pipelines developed in collaboration with Gioele Barabucci, showing what it actually meant to translate into code the usually very analogue task of normalising medieval texts, what problems we encountered and how we solved them. Finally, as an outlook for the possible applications of the computational principle, I will briefly illustrate the work I am doing for my next project, the edition of the *Prenostica Socratis Basilei*, which was conceived from the start as a computational one. In my talk, I will show some prototypical examples concerning specific aspects of the computational workflow.

Fußnoten

1. An example of the application of an integrated computational approach for different phases of the editorial process is the digital critical edition of the Chronicle of Matthew of Edessa (Andrews, Safaryan and Atayan 2019).
2. Although there seems to be a need for further discussions also as far as the “edition as interface” is concerned. In this respect, Andrews and Van Zundert (2018) argue that the interface is “not just an argument about

the text, but also an argument about the ‘attitude’ of the editor, a window into his or her take on methodology and the digital edition itself. It is also a revelation of the technical skills available to the editor. The interface tells us something not only about the methodology but also about the import of the edition”. The authors observe, however, that many editors still underestimate the role of the interface, whose “language” is one of great complexity.

3. “Take, for example, knowledge one might have of a particular thematic concentration in a deeply familiar work of literature. In modelling one begins by privileging this knowledge, however wrong it might later turn out to be, then building a computational representation of it, e.g. by specifying a structured vocabulary of word-forms in a text-analysis tool. In the initial stages of use, this model would be almost certain to reveal trivial errors of omission and commission. Gradually, however, through perfective iteration trivial error is replaced by meaningful surprise” (McCarthy 2005).

4. The pipelines are available online on the pre-release website of the DI edition: <https://marcopolo.cceh.uni-koeln.de>.

5. In February 2022 it received an “Initial Funding” from the Faculty of Arts and Humanities at the University of Cologne (<https://phil-fak.uni-koeln.de/en/research/research-funding/initial-funding>).

6. The roots of the *sortes* have been traced down to the Mediterranean area (Abraham 1968), but their use soon spread all over Europe. Since the Renaissance these texts have been used as entertaining games to play in a group (Urbini 2006), but this was not the case in antiquity and for the greater part of the Middle Ages. Then, they were taken seriously and could have a strong influence on people’s decisions.

Bibliographie

Abraham, Werner. 1968. “Studien zu einem Wahrsage-text des späten Mittelalters.” *Hessische Blätter für Volkskunde* 59, 9-24.

Alonso Guardo, Alberto. 2015. *Prenostica Socratis Basilei: étude, édition critique et traduction*. Textes littéraires du Moyen âge 4. Paris: Classiques Garnier.

Andrews, Tara L. and Joris J. Van Zundert. 2018. “What Are You Trying to Say? The Interface as an Integral Element of Argument.” In *Digital Scholarly Editions as Interfaces*, edited by Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, Gerlinde Schneider, 3-33. Norderstedt: BoD.

Andrews, Tara Lee, Anahit Safaryan, and Tatevik Atayan. 2019. “Continuous Integration Systems for Critical Edition: The Chronicle of Matthew of Edessa”. *DataVerseNL*. <https://doi.org/10.34894/MZ7FBI>.

Barabucci, Gioele and Franz Fischer. 2017. “The Formalization of Textual Criticism. Bridging the Gap between Automated Collation and Edited Critical Texts.” In *Advances in Digital Scholarly Editing: Papers Presented at the Dixit Conferences in the Hague, Cologne, and Antwerp*, edited by Peter Boot, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Elena Spadini, and Dirk Van Hulle, 47-53. Leiden: Sidestone Press.

Schaeben, Marcel, and Gioele Barabucci. 2021. 'Small-Step Pipelines Reduce the Complexity of XSLT/XPath Programs'. In *Proceedings of the 21st ACM Symposium on Document Engineering*, 1–4. Limerick Ireland: ACM. <https://doi.org/10.1145/3469096.3474922>.

Cugliana, Elisa and Gioele Barabucci. 2022. "Signs of the Times: Medieval Punctuation, Diplomatic Encoding and Rendition." *Journal of the Text Encoding Initiative* 14. DOI: <https://doi.org/10.4000/jtei.3715>.

Cugliana, Elisa and Joris J. Van Zundert. 2022. "A Computational Turn in Digital Philology". *Filologia Germanica / Germanic Philology* 14, 43-72.

Heiles, Marco. 2018. *Das Losbuch. Manuskriptologie Einer Textsorte Des 14. Bis 16. Jahrhunderts*. Köln: Böhlau Verlag.

McCarty, Willard. 2005. *Humanities Computing*. Basingstoke, New York: Palgrave Macmillan.

Rosselli Del Turco, Roberto. 2016. "The Battle We Forgot to Fight: Should We Make a Case for Digital Editions?" In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 219-238. Cambridge: Open Book Publishers.

Sahle, Patrick. 2013. *Digitale Editionsformen*. 3 Voll. Schriften des Instituts für Dokumentologie und Editorik. Norderstedt: BoD.

Sahle, Patrick. 2016. "What is a scholarly digital edition (SDE)?" In *Digital Scholarly Editing. Theory, Practice and Future Perspectives*, edited by Matthew Driscoll and Elena Pierazzo 19-39. Cambridge: Open Book Publishers.

Stella, Francesco V. 2007. "Metodi e prospettive dell'edizione digitale di testi mediolatini." *Filologia mediolatina* XIV, 149-180.

Urbini, Silvia. 2006. *Il Libro delle sorti*. Modena: Francesco Cosimo Panini.

Van Zundert Joris J. 2018. "Why the Compact Disc Was Not a Revolution and Cityfish Will Change Textual Scholarship, or What Is a Computational Edition?" *Ecdotica*, 129-56. <https://doi.org/10.7385/99283>.

Datenaufbereitung und -kuration im Spannungsfeld von Reproduzierbarkeit und Wiedernutzung als Leitideen der Open Sciences. Eine Fallstudie aus der Kunstgeschichte

Niemann, Klara

klaraniemann@gmx.de
Universität zu Köln

Klammt, Anne

aklammt@hotmail.com

Deutsches Forum für Kunstgeschichte Paris

Einleitung

Zu den Anstrengungen der Europäischen Union für die Umsetzung von *Open Sciences* gehört auch eine Klärung und Förderung der Reproduzierbarkeit und der Sicherung von Integrität. Unter Reproduzierbarkeit wird dabei im *Scope Report* der Europäischen Kommission die Wiederholbarkeit des Forschungsprozesses mit denselben Daten und Methoden verstanden (Directorate-General for Research and Innovation 2020). Für die Autor*innen des Reports stellt die Reproduzierbarkeit einen speziellen Fall der Wiedernutzbarkeit (Re-Use) von Forschungsdaten dar. Aus unserer Sicht bilden jedoch die strikte Reproduzierbarkeit und eine offene Wiedernutzung Szenarien, die in der Praxis zu widersprüchlichen Anforderungen an die Kuration von Forschungsdaten sowie ihrer Ausgabe über graphische Nutzeroberflächen und APIs führen. Im Verhältnis mit der Datenintegrität entsteht dabei ein nicht vollständig aufzulösendes Dilemma, das Versuche der Aushandlung aber in eine produktive Spannung überführen können. Dies möchten wir verdeutlichen, indem wir einerseits unseren Gebrauch der Begriffe Re-Use, Neuausrichtung und Datenkuration präzisieren, und die Problematik andererseits an einem Fallbeispiel aus der Forschungsarbeit des Deutschen Forums für Kunstgeschichte Paris (DFK Paris) veranschaulichen, für das wir nach Wegen gesucht haben, zwischen den Polen Reproduzierbarkeit und Re-Use zu vermitteln. Es handelt sich um die Aufbereitung und neue Präsentation einer 20 Jahre alten Datenbank zur wechselseitigen Rezeption des Kunstgeschehens in Texten der Kunstkritik aus Deutschland und Frankreich zwischen 1870 und 1960 (DFKV) (DFK Paris 2022b).

Fallstudie

Von 1999 bis 2005 wurden in drei aufeinanderfolgenden Förderungen Quellenanthologien, Aufsätze, eine Monografie und drei bibliographische Datenbanken zur deutsch-französischen Kunstrezeption erstellt (DFK Paris 2022c), die den Blickwinkel der in den 1990er und 2000er Jahren sehr einflussreichen Kulturtransferforschung einnahmen (Espagne & Werner 1988; Espagne 1999; Gaethgens 2009). Die Datenbanken waren technologisch und im Datenmodell einheitlich angelegt, während sich die Verschlagwortung jeweils spezifisch entwickelte. Insgesamt beinhalten sie knapp 6800 mehrheitlich kommentierte und verschlagwortete Hinweise auf Artikel, Meldungen und Notizen in 314 verschiedenen Reihen deutscher, beziehungsweise französischer Kunstzeitschriften und wenigen zeitgenössischen Buchpublikationen. Ziel war es, Kunsthistoriker*innen ein Hilfsmittel zur wissenschaftlichen Arbeit anzubieten. Ab Winter 2004/2005 standen sie offen online zur Verfügung und sind bis 2016 über die Webseite des DFK Pa-

ris auffindbar gewesen. 2019 erfolgte ein erster Relaunch der zwischenzeitlich in eine MySQL Datenbank migrierten Datenbanken. Von 2021 bis 2022 wurden die Daten umfangreich kuratiert und im Juni 2022 neu veröffentlicht (DFK Paris 2022a). Ausgangspunkt war die mangelhafte Nutzbarkeit der öffentlichen Datenpräsentation, die aus der mehrfachen Migration hervorgegangen war und das Verständnis für die Zusammensetzung und Bedeutung der Daten minderte. Das Webangebot führt heute ein einstiges Werkzeug fort, dessen zugrundeliegende Forschungsfrage des Kulturtransfers inzwischen aufgrund der Weiterentwicklung hin zur Untersuchung von Mobilität und Migration nicht mehr in gleicher Weise gestellt wird.

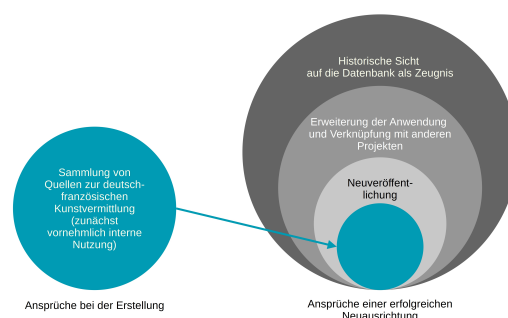
Begrifflichkeiten: Re-Use, Datenkuration und Neuausrichtung

Um die von uns verwendeten Begrifflichkeiten zu schärfen und Divergenzen zu anderen Definitionen vorzubeugen, sei eine Erläuterung des hier zugrunde gelegten Verständnisses der Begriffe *Re-Use*, *Neuausrichtung* und *Datenkuration* vorangestellt. *Re-Use* beschreibt die erneute, offen gedachte Nutzung von Daten in der Forschung zur Beantwortung einer neuen Fragestellung. Dies kann die Rekombination mit anderen Daten beinhalten und ist ein wesentliches Ziel der Bemühung um die Anwendung der FAIR Prinzipien (Huie et al. 2021).

Als Datenkuration möchten wir hier eine Gruppe von Aktivitäten verstehen, die an der Daten-haltenden Institution angesiedelt ist, und mit dem Ziel, eine verständige Nachnutzung durch Dritte zu ermöglichen, durchgeführt wird. Notwendige Voraussetzung sind dabei die Dokumentation und Kenntnis der Erzeugung, Prozessierung und Anzeige der Daten, wie es Flanders und Muñoz aus Perspektive der Geisteswissenschaften zusammengestellt haben (Flanders u. Muñoz 2015). Dabei weisen die von Kim und Koh (Kim u. Koh 2021) herausgegebenen Forschungen zur Geschichte von Digital Humanities-Projekten darauf hin, dass unter „Erzeugung“ wie „Prozessierung“ ein Zusammenspiel von theoretischer Ausrichtung, methodischer Vorgehensweise und organisatorischen wie institutionellen Faktoren zu betrachten ist. Die Dokumentation nach Flanders und Muñoz ist die Voraussetzung, um Daten überhaupt auf eine neue Verwendung hin aufzubereiten. Die Herausforderung liegt nach Woodall (2017) dann darin, die Eignung der Daten für diesen neuen Anwendungsfall bestimmen zu können und sie gezielt daraufhin zu entwickeln, um Fehlinterpretationen vorzubeugen. Im Idealfall sollte die Datenkuration darauf gerichtet sein, innovative Forschungen zu ermöglichen und die Daten entsprechend für möglichst viele Ansatzpunkte öffnen, beispielsweise durch Verknüpfung mit Normdaten.

Mit dem Begriff der Neuausrichtung schlagen wir eine spezifische Auslegung der Datenkuration vor, die sich diesen Herausforderungen auch auf Ebene der Benutzeroberfläche widmet, wie es die Medienwissenschaftlerin Drucker (2021, 78) für die Aufbereitung von Forschungsdaten in den Geisteswissenschaften angeregt

hat. Wir erweitern das Verständnis von der Aufbereitung damit inhaltlich gegenüber dem vorher in den *Data Sciences* auf die Daten gerichteten Fokus (Woodall u. Wainman 2015) und nähern es dem *Refashioning* nach Bolter und Grusin (1999, 45 f.) an. Ziel ist es also, einen (alten) bestehenden Datensatz visuell und textuell so zu vermitteln, dass dessen Beschaffenheit für den *Re-Use* verständlich wird (bspw. durch eine entsprechende Suchmaske). Indem die Neuausrichtung durch ihre spezifische kuratorische Ordnung und Anreicherung der Daten jedoch bereits Beispiele der Weiternutzung, Interpretation und Verknüpfung anbietet, nimmt sie ein ambivalentes Verhältnis sowohl zur Maxime der weitmöglichsten Öffnung der Daten zum *Re-Use* als auch zur historischen Gewachsenheit der Daten ein. Diese gilt es wiederum zu kommunizieren.



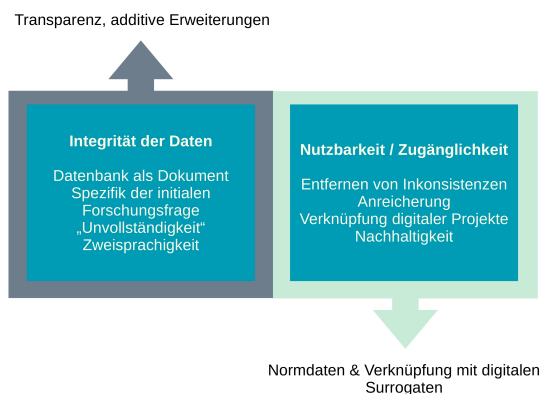
Die verschiedenen Nutzungsansprüche einer republizierten Datenbank im Vergleich zur ursprünglichen Zielsetzung am Beispiel der Datenbanken DFKV (Grafik: Klara Niemann, CC BY 4.0).

Vom Dilemma zwischen Neuausrichtung und historischer Integrität

Als Maßnahmen der Datenkuration möchten wir aus unserem Fallbeispiel die semantische Anreicherung von Daten durch die Referenzierung mit Normdaten und die Einbeziehung von digitalen Surrogaten des Quellmaterials für Datenbanken der DFKV heranziehen. Auf diese Weise wurden von 9076 in den Datenbanken DFKV genannten Personen (Autor*innen, Künstler*innen, Kurator*innen und weitere) 4686 (52 %) mit Normdaten (Getty Union List of Artists Names (ULAN), Gemeinsame Normdatei (GND), *notices d'autorité* der Bibliothèque nationale de France (BnF)) oder Wikidata referenziert. Dies hat zur Entdeckung von insgesamt 603 Personennamen geführt, die entweder in mehr als einer der Datenbanken auftreten oder auch unbemerkt jeweils mit verschiedenen Namensvarianten eingetragen wurden. Diese Varianten sind meist in den Quellen angelegt, wenn etwa Vornamen und Adelstitel in Französisch oder Deutsch übersetzt wurden. Die Gewohnheit, in den Zeitschriften Namen zu übersetzen oder auch Umlaute und Ak-

zente anzupassen, ist ein Hinweis zur ursprünglichen Rezeption. Dass diese verschiedenen Schreibweisen bei der Datenerfassung in den 2000er Jahren übernommen, aber nicht mit übereinstimmenden Personen assoziiert wurden, ist wiederum eine wichtige Information zur Einschätzung der Qualität der Daten. Darüber hinaus konnten einige Aliasnamen aufgedeckt werden, die den Erstbearbeiter*innen nicht bekannt waren.¹ Zusammengefasst beträgt der Anteil an Dubletten damit rund 9 %. Für 2548 der 5735 in der Datenbank beschriebenen Zeitschriftenbeiträge (Zeitraum vor 1940) konnte eine Verlinkung auf ein Digitalisat in den Angeboten der Universitätsbibliothek Heidelberg oder der BnF erstellt werden. Wo ein Dateneintrag zusammenfassend auf mehrere verschiedene Zeitschriftenbeiträge verweist, musste die Datengrundlage erweitert werden, sodass die Gesamtanzahl bibliografischer Attribute in der Datenbank durch die Kuration von 5948 auf 6194 angestiegen ist.

Im Sinne der Neuausrichtung sollten die Mehrfachnennungen der Personen zusammengefasst werden und die Qualität der Daten einschätzbar sein, aber zugleich die historisch bedingte Ambiguität erhalten bleiben. Dieses Dilemma zwischen der Neuausrichtung und dem Erhalt der ursprünglichen Datenbanken als Artefakt hat uns zur Frage geführt, was die historische Integrität dieser Daten ausmacht. Medienarchäologische Studien und Datenzentren haben von unterschiedlichen Ausgangspunkten ausgehend wahlweise die vollständige Emulation oder die Erhaltung der Präsentationsschichten bei einem Wechsel der zugrundeliegenden Technik vorgeschlagen (Waelder 2017; Steiner et al. 2022). Im Falle der Datenbanken der DFKV sind jedoch zum einen bereits unterschiedliche Softwares und Ansichten zur Dateneingabe und für die Ausgabe im Internet verwendet worden, sodass mehrere Versionen emuliert werden mussten. Zum anderen hat sich aus den Befragungen von ehemaligen Mitarbeiter*innen zum Gebrauch der Datenbanken in den 2000er Jahren ergeben, dass die Software als solche kaum wahrgenommen wurde und weder Funktionen zur Filterung noch des Exports später beschrieben werden konnten.² In dieser Situation haben wir uns entschlossen, nicht die historische Integrität der Datenbanken als Ganzes zu betrachten, sondern auf die Ebene der einzelnen Datensätze zu gehen und sie abstrakter als einen definierten Zustand der Daten anzusehen.³ Dadurch sind wir dazu gelangt, die Datenbanken orientiert an CIDOC CRM als Konvolute von Dokumenten (E31; Bekiari et al. 2022, 83 f.) zu verstehen, deren Erzeugung wie auch Anreicherung diskrete Ereignisse (E5; Bekiari et al. 2022, 63 f.) ihrer Objektgeschichte bilden. Mit den Ereignissen sind ein Zeitraum (1999–2004 und 2021–2022), die ausführenden Personen, die Art der Aktivitäten und damit die Zusätze und Streichungen beschreibbar.



Das Dilemma zwischen Integrität und Nutzbarkeit der Daten am Beispiel der Datenbanken DFKV (Grafik: Klara Niemann, CC BY 4.0).

Gestaltung des GUI

Ausgehend von der Idee, die Zustände und Anreicherungen der Daten selbst erfahrbar zu machen, haben wir das GUI gestaltet (Niemann 2021). Es ist in drei Funktionsbereiche aufgeteilt: die Übersichtsseite zur Suche und Auswahl der in verkürzter Ansicht gezeigten Dateneinträge, die vollständige Ansicht der einzelnen Dateneinträge und eine Merkliste mit individuell von den Nutzer*innen ausgewählten Einträgen.

In der vollständigen Ansicht der einzelnen Dateneinträge wurde mit Farbhintergründen und Schichten gearbeitet, die die Inhalte den verschiedenen Phasen der Datenbanken und ihrer Bearbeitung zuordnen.⁴ Auf neutralem Grund sind bibliografische Angaben und Textauszüge angelegt, die faktische Informationen zum aufgenommenen Quelltext liefern. Farblich hinterlegt sind die Schlagworte, Kommentierungen und weitere Anreicherungen der Datenerfassung und somit des ersten objektgeschichtlichen Ereignisses. Als Widgets können für die Autor*innen und die als genannte Personen angegebenen Namen die Referenzierungen auf Normdaten oder Wikidata und Hinweise auf weitere in der Datenbank vorhandene Schreibweisen aufgerufen werden. Sie bilden somit das Ereignis der Kuration ab. In gleicher Weise sind Informationen zur Zeitschrift bei der BnF oder der GND aufrufbar. Als Fly-in schiebt sich von rechts über die Grundebene ein Widget mit Erläuterungen zur Nutzung, die anhand von Symbolerklärungen die Hintergründe der objektgeschichtlichen Stationen und die Entscheidungen der Kuration transparent vermitteln. Als äußerste Schicht kann jeder Dateneintrag als JSON in einem weiteren Widget aufgerufen werden, um die Anreicherungen und Verknüpfungen der Informationen per IDs (als Spuren der ursprünglichen Erfassung in einer relationalen Datenbank) nachzuvollziehen. Insgesamt stellen diese Gestaltungsentscheidungen eine Reaktion auf den Anspruch der Nachvollziehbarkeit und damit der Reproduzierbarkeit dar.

Die Suchfunktionen auf der Übersichtsseite sind hingegen auf den Re-Use ausgerichtet. Die verschiedenen Optionen, über die Datenbankzugehörigkeit, mit dyna-

mischen Filtern in vorgegebenen Kategorien, dem interaktiven Zeitstrahl oder der Freitextsuche zu suchen, regen eine Entdeckungstour durch die Daten an, bei der die Nutzer*innen weniger über eine spezifische Suche als ein Schweifen in das Material einsteigen. Um dies zu fördern, kann man sich auf der bereits beschriebenen Vollanzeige der Dateneinträge außerdem horizontal per Klick von einem zum nächsten bewegen. Die digitalen Surrogate werden schließlich über ein Icon aufgerufen und öffnen sich in einem IIIF-Viewer, der sich in einem neuen Browsertab öffnet. Indem die Manifeste der Zeitschriftenbände verknüpft sind, können die referenzierten Artikel und Beiträge vollständig gelesen werden, in dem Band geblättert werden und weitere von den Bibliotheken erstellte Metadaten aufgerufen werden.

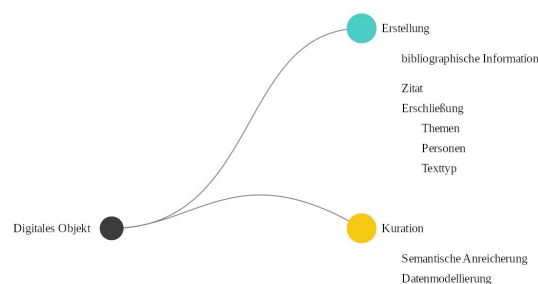
Umsetzung in der Datenmodellierung

Nicht implementiert haben wir eine vorerst prototypische Umsetzung dieser Schichtung in einem Datenmodell, das wir mit dem Linked Art Data Model erstellt haben (Klammt 2022). Das Linked Art Data Model (LADM) ist eine Anwendung des CIDOC CRM (Linked Art), das den Re-Use von Kulturdaten unterstützen möchte. Anders als LIDO XML geht es dabei nicht um ein Transferformat, mit dem Metadaten verschiedener Quellen zusammengeführt werden können, sondern darum, Kulturdaten zu einfach nutzbaren Linked Open Data zu machen. Dieser Fokus spiegelt sich auch in der Wahl von JSON als Dokumentformat. In einem ersten Prototyp haben wir die einzelnen Dateneinträge als Informationsobjekt modelliert, das auf meist einen Zeitschriftenbeitrag referenziert, dessen bibliografische Angaben mit dem LADM ausgedrückt werden können. Jeder Dateneintrag hat als Ereignisse seine Erstellung und die Kuration eingetragen, mit den Zeiträumen und den jeweils verantwortlichen Projektleitern. Über verschiedene definierte Eigenschaften sind die Verlinkung zum IIIF-Manifest, die Kommentierungen und Verschlagwortung genauso wie die Referenzierung auf Normdaten im Zuge der Neuausrichtung als LOD erklärt. Die Verwendung des Modells für die Beschreibung von Forschungsdaten liegt außerhalb seiner ursprünglichen Intention. Sie erlaubt aber auf Datenebene transparent zu dokumentieren, welche Maßnahmen zur Ausrichtung der Daten auf den Re-Use ausgeführt wurden. Gleichzeitig können diese Veränderungen reversibel eingeschrieben werden.

Resümee

Im Prozess der Neuausrichtung der DFKV-Datenbanken durch verschiedene Anreicherungsprozesse und das Einbetten in ein neues GUI sahen wir uns in der Praxis mit der Aufgabe einer offen gedachten Reproduzierbarkeit gegenübergestellt. Reproduzierbarkeit ist dabei sowohl der Weiterverwendung als auch der Integrität der Daten verpflichtet. Wenn es darum geht, geisteswissenschaftliche Datenbanken langfristig zu erhalten, heißt das, diese auch jenseits der Langzeitarchivierung

im Sinne der dynamischen Entwicklung der Forschungsfragen anschlussfähig an die wissenschaftliche Community zu halten. Entscheidungen der Datenkuration müssen in diesem Prozess individuell entsprechend des Einzelfalls, aber immer respektive der Geschichte und konkreten Beschaffenheit der Daten getroffen werden. Versteht man alle Eingriffe und Veränderungen als spezifische Ereignisse der Datenhistorie, gilt es, diese für den Re-Use transparent zu kommunizieren. Im Angesicht der ersten, nicht mehr abrufbaren Datenbanken aus den späten 1990ern und frühen 2000ern wird der Handlungsbedarf in diesem Bereich deutlich.



Datensätze als digitales Objekt, das durch Erstellung und Kuration geformt wird (Grafik: Anne Klammt; Lizenz: CC BY 4.0).

Fußnoten

1. Ein Beispiel ist die Kunsthistorikerin Lina Boelsche, die unter dem männlichen Synonym Hermann Billung Zeitschriftenartikel veröffentlichte und dementsprechend unter diesem Namen in den Datenbanken geführt wurde. Die Verknüpfung mit Wikidata (<https://www.wikidata.org/wiki/Q95196696>) ermöglichte die nachträgliche Identifikation.
2. Die Befragung per Interviews und Fragebögen wurde von Deborah Schlauch, DFK Paris, von Januar bis Mai 2022 durchgeführt.
3. Unbemerkt haben wir uns damit der Frage nach den „signifikanten Eigenschaften“ nach Giaretti et al. (2009) und Recker (2021) angenähert. Für den freundlichen Hinweis danken wir T. Staecker, Darmstadt.
4. Als Beispiel eines ausführlichen Datenbankeintrags siehe: https://dfk-paris.org/de/page/deutsch-franzoesische-kunstvermittlung-1870_1940-und-1945_1960-datenbank-2391.html#/records/10720.

Bibliographie

- Bekiari, Chryssoula et al. 2022. "Definition of the CIDOC Conceptual Reference Model Version 7.2.1. Volume A: Definition of the CIDOC Conceptual Reference Model. Version 7.2.1".
- Bolter, Jay David und Richard Grusin. 1999. *Remediation. Understanding New Media*. Cambridge: MIT Press.
- DFK Paris. 2022. "Datenkuration am Beispiel der Datenbank Deutsch-Französische Kunstvermittlung 1870-1940 und 1945-1960". <https://dfk-paris.org/de/re->

search-project/datenkuration-am-beispiel-der-datenbank-deutsch-franz%C3%B6sische-kunstvermittlung-1871 (zugegriffen 2. August 2022).

DFK Paris. 2022. "Deutsch-Französische Kunstvermittlung 1870-1940 und 1945-1960". https://dfk-paris.org/de/page/deutsch-franzoesische_kunstvermittlung_g_1870%E2%80%931940_und_1945%E2%80%931960-2389.html (zugegriffen 2. August 2022).

DFK Paris. 2022. "Publikationen des Projekts 'Deutsch-Französische Kunstvermittlung'." https://dfk-paris.org/de/page/dfkv_publikationen-3307.html (zugegriffen 2. August 2022).

Flanders, Julia, und Trevor Muñoz. 2015. „An Introduction to Humanities Data Curation | DH Curation Guide“. <https://web.archive.org/web/20150822055930/http://guide.dhcurator.org/contents/intro/> (zugegriffen 2. August 2022).

Gaehtgens, Thomas W. 2009. "Introduction: De la réception de l'art moderne français en Allemagne entre 1870 et 1945". In *Perspectives croisées. La critique d'art franco-allemande 1870-1945*, hg. von Thomas W. Gaehtgens, Mathilde Arnoux und Friederike Kitschen, 3-26. Paris: Éd. de la Maison des sciences de l'homme.

Huie, J. Russell et al. 2021. "FAIR Data Reuse in Traumatic Brain Injury: Exploring Inflammation and Age as Moderators of Recovery in the TRACK-TBI Pilot." In *Frontiers in neurology* 10.3389/fneur.2021.768735.

Kim, Dorothy und Adeline Koh. 2021. *Alternative Historiographies of the Digital Humanities*. punctum Books 10.53288/0274.1.00.

Klammt, Anne. 2022. "DFKV - Data Model". https://github.com/archaeoklammt/DFKV_data_model (zugegriffen 2. August 2022).

Linked Art. <https://linked.art/> (2. August 2022).

Niemann, Klara. 2021. "Die Aufbereitung der Datenbank Deutsch-Französische Kunstvermittlung 1870-1940 und 1945-1960 und ihre zukünftigen Nutzungsmöglichkeiten". In *Jahresbericht des DFK Paris 2020/2021*, 110-111.

Steiner, Elisabeth, Gunter Vasold und Martina Scholger. 2022. "Repositorien als digitale Gedächtnisträger zwischen Evolution und Langzeitplanung". In *DHd2022: Kulturen des digitalen Gedächtnisses*, hg. von Michaela Geierhos et al. 10.5281/zenodo.6304590.

Waelder, Paul. 2017. "Summary: Media Archaeologies Evening. First December 2017, Barcelona". <http://catedratelefonica.uoc.edu/wp-content/uploads/2018/01/Media-Archeologies-BCN.pdf> (zugegriffen 2. August 2022).

Woodall, Philipp. 2017. "The Data Repurposing Challenge: New Pressures from Data Analytics". In *Journal of Data and Information Quality* 8 10.1145/3022698.

Woodall, Philip, und Anthony Wainman. 2015. "Data quality in analytics: key problems arising from the repurposing of manufacturing data". In *Proceeding of the International Conference on Information Quality (ICIQ'15)*, 174-184.

Die besonderen Herausforderungen multimodaler heterogener Daten- und Quellentypen an die Datenverwaltung. Ist eine Forschungsdateninfrastruktur ohne eine Datenbank umsetzbar?

Gerber, Anja

anja.gerber@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Das Forschungsvorhaben

In einem Kooperationsprojekt zwischen der Potsdamer Arbeitsstelle des Corpus Vitrearum Medii Aevi (CVMA) Deutschland an der Berlin-Brandenburgischen Akademie der Wissenschaften sowie dem Institut für Kunstwissenschaft der Technischen Universität Berlin wird die digitale Erschließung und Repräsentation der Nikolaikirche in Bad Wilsnack sowie der Quellen zur als "Wilsnackfahrt" bekannten Wallfahrt umgesetzt. Hierbei handelt es sich um ein interdisziplinäres Forschungsvorhaben aus den Bereichen Kunstgeschichte, Mediävistik, Digital Humanities (DH) und Informationswissenschaften. Verschiedene Forschende haben unterschiedliche Bedarfe an die Infrastruktur sowie die erzeugten digitalen Informationen und Dateien.



Abbildung 1: Wilsnack, Kirche St. Nikolai, Außenansicht, Blick von Westen, Foto: Berenike Rensinghoff.

Die Kirche St. Nikolai in Bad Wilsnack ist auch als Wunderblutkirche bekannt und war seit Auffindung der drei „Wunderbluthostien“ im Jahr 1383 ein zentraler Wallfahrtsort (Bednarz et al. 2010, 87–165; Kühne und Ziesack 2005, 9–18). Zahlreiche Quellen verschiedener Art belegen dies, als Stiftungen in Form von Glasmalereien und Ausstattungsstücken des Bauwerks oder als Überlieferungen wie Pilgerzeichen, Testamente und Ablass. Schriftquellen liegen teilweise nur gedruckt oder nicht erschlossen in Archiven verschiedener nord- und mitteleuropäischer Länder vor und benötigen eine systematische Aufarbeitung. Überlieferungen sind zum Teil nicht sicher belegt und nur durch Verweise auf nicht mehr vorhandene Quellen bekannt. Das Bauwerk war aufgrund der politischen und religiösen Geschehnisse im Laufe der Jahrhunderte verschiedenen baulichen Veränderungen ausgesetzt. Neben umfangreichen Restaurierungsmaßnahmen wurden Ausstattungsstücke verändert, bewegt, entfernt oder gar – insbesondere während der Zeit der Reformation und durch verschiedene Eigentümerwechsel – zerstört (Bednarz et al. 2010, 88–92).



Abbildung 2: Wilsnack, Kirche St. Nikolai, Blick in den Chor mit Hochaltarretabel, Foto: Berenike Rensinghoff.

Datenbasis

Die Quellen der Wallfahrt bilden die Basis für die Datenerfassung und eine daraus erfolgende Repräsentation, z. B. auf Karten oder als 3D-Modell des Bauwerks sowie ausgewählter Ausstattung. Die Forschungsdateninfrastruktur, die aktuell entwickelt wird, muss einerseits mit heterogenen Inhalten der Quellen, andererseits auch mit verschiedenen Dateitypen umgehen. Neben Bilddateien als TIFF oder JPG sowie Regesten der Quellen, die in proprietären Office-Formaten erstellt und in ein menschen- und maschinenlesbares Format wie XML oder JSON konvertiert werden, entstehen ebenfalls 3D-Daten (Punktwolken, digitale Rekonstruktionen) in unterschiedlichen Formaten, z. B. OBJ oder PLY, und Koordinaten als GeoJSON für die Kartendarstellungen. Heterogene Dateitypen und verschiedene Informationsschichten in z. T. ungenauem bzw. ungesichertem Überlieferungsstatus werden zu einem multimodalen Datenmodell zusammengeführt. Das Bauwerk betreffende Informationen werden mit Ereignissen in Verbindung gebracht, um zu erfassen und abzufragen, was bspw. Akteure der damaligen Zeit gesehen haben könnten, als sie die Kirche St. Nikolai zu einem bestimmten Zeitpunkt betraten, welche Ausstattungsstücke wann sichtbar waren oder welche Routen Pilger zurückgelegt haben. Die digitalen Ressourcen werden auf einer von der Software getrennten Dateiebene wiederum mit Metadaten angereichert.

Forschungsdateninfrastruktur

Im Bereich des digitalen Kulturerbes gibt es nur wenige standardisierten Lösungen für digitale Infrastrukturen, auf die zurückgegriffen werden kann. Bekannte virtuelle Forschungsumgebungen sind MonArch und WissKI. MonArch ist jedoch für die Zwecke des Vorhabens nicht ausreichend, da es hier primär um *Building Information Modeling* für die Erfassung und semantische Annotation bauwerksbezogener Inhalte geht (AriInfoware 2022). WissKI ist eine Forschungsumgebung, die über das CMS Drupal Zugriff auf verschiedene Dateitypen bietet und mittels Datenmodell organisiert ist. Eine Repräsentation von Kartendarstellungen und eine Einbindung von 3D-Viewern ist möglich (Germanisches Nationalmuseum o. J.). Im Bereich des Digitalen Kulturerbes findet vor allem das *CIDOC Conceptual Reference Model* (CRM) für die Datenmodellierung Anwendung, da hiermit objektbezogene und ereigniszentrierte Informationen und ihre Eigenschaften modelliert werden können (Bekiar et al. 2022). So kann eine Nachnutzbarkeit bereits vorhandener aber auch der eigenen, zum Großteil noch zu erstellenden Daten gewährleistet werden. Eine Besonderheit stellt im Bereich der Kunstgeschichte und des Kulturerbes die Vermittlung von Beziehungen zwischen Informationen dar, die getroffene Aussagen über Objekte und Orte in Zusammenhang mit Akteuren und Ereignissen in einen Kontext setzen (Burricher et al. 2021, 111f.).

Eine Speicherung der Dateien erfolgt in einer durch das Projekt vorgegebenen Ordnerstruktur (Kontinent/Land/Bundesland/Stadt/Gebäude/Datentyp). Verschiedene Tools greifen auf die digitalen Ressourcen zu und die Datenschicht ist von der Software getrennt (Gerber 2022, 16–28). Für ihre Erschließung mit Metadaten aber auch für die Erfassung des Datenmodells wird derzeit der bereits in den beiden Arbeitsstellen des CVMA Deutschland verwendete CVMA Digitaler Ressourcen Manager (CVMA DRM) (Gerber und Fischer 2021) weiterentwickelt. In erster Linie wurde er für die Erschließung von digitalen Bilddateien konzipiert. Die Inhalte werden online über ein webbasiertes GUI auf Basis von Javascript aufgerufen und verarbeitet. Der Zugriff auf die Daten erfolgt über einen Webserver. Metadaten werden in einer der jeweiligen digitalen Ressource zugehörigen maschinen- und menschenlesbaren JSON-Datei gespeichert und perspektivisch in GitLab versioniert. Das zu annotierende Datenformat ist unerheblich, da sich die JSON-Dateien mit jedem Datentyp verknüpfen lassen, neben Bilddaten z. B. auch Textdateien, Daten der Punktwolken oder Koordinaten. Eine Zuordnung erfolgt über den Dateinamen und den persistenten Identifikator (PID). Die Metadateninformationen haben denselben Dateinamen wie die jeweils verschlagwortete Ressource, die Dateierweiterung *.meta* und sie werden in derselben Ordnerstruktur wie die durch sie beschriebenen Dateien abgelegt. Durch die Speicherung als JSON-Dateien werden zunächst nur die Metainformationen und stark herunter skalierte Bilder bzw. Vorschauansichten der Modelle geladen. Die großen, hochauflösten Daten sind für eine Anzeige explizit auszuwählen, so dass sich die Ladezeiten erheblich verkürzen. Die Konfiguration der Ansicht und des Metadatenschemas erfolgt ebenfalls über JSON-Dateien.

The screenshot shows a complex web interface with a table of research objects. The table has columns for object ID, name, location, date, and various metadata fields. The interface includes search filters and a sidebar with navigation options.

Abbildung 3: Ansicht Graphical User Interface (GUI) des CVMA Digitaler Ressourcen Manager für die Erschließung der Forschungsobjekte, Anja Gerber 2022.

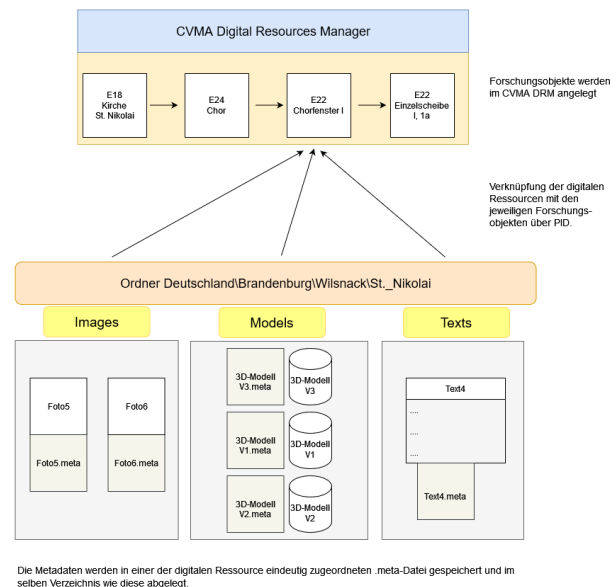


Abbildung 4: Verknüpfung der Forschungsobjekte mit Ressourcen und Metadattendateien, Anja Gerber 2022.

Die erschlossenen Inhalte der Quellen werden durch das Datenmodell organisiert und mittels verschiedener Viewer, u. a. Kompakt für 3D-Modelle (Universität zu Köln 2018–2022), Digilib für Bilder (digilib Community 2001–2022) und das Chronotopische Tool für Kartendarstellungen (Fischer und Thomas 2021), repräsentiert. In diese Viewer können zudem weitere digitale Ressourcen, z. B. Bilder oder beschreibende Texte, eingebunden und mit den dort dargestellten Daten angezeigt, sowie die Inhalte mit zusätzlichen Informationen außerhalb der Metadaten annotiert werden. Ein einzelnes System müsste aufwendig angepasst oder es muss auf verschiedene Datenbanken zurückgegriffen werden, um die verschiedenen Dateitypen speichern zu können. Daher ist die Entwicklung einer Weboberfläche geplant, über die der Ressourcenmanager zur Erfassung und Recherche, aber auch die Tools zur Anzeige der Präsentationsschicht eingebunden und angezeigt werden. Bisher erfolgt ein Aufruf der jeweiligen Tools über verschiedene Links. Sie befinden sich bereits auf demselben Webserver und greifen auf denselben Datenbestand zu. Für die Regesten und

Texte der Quellen muss zum Zeitpunkt des Abstracts noch eine Lösung gefunden werden.

Datenmodellierung

Über das Datenmodell werden die sehr heterogenen Informationen und Dateien inhaltlich organisiert und mittels automatisiert vergebener PID miteinander verknüpft. Dieser Prozess erfolgt manuell. Aus den Quellentexten entnommene Angaben zu Akteuren, Datierungen, Ereignissen, Objekten und Orten erhalten konkrete Bezeichnungen und bilden die Gruppe der Forschungsobjekte. Diese werden dann verschiedenen Kategorien des Datenmodells zugewiesen, z. B. St. Nikolai als Bauwerk der Klasse *E18 Physical Object*, Wilsnack als Stadt der Klasse *E53 Place* oder Bischöfe und Pilger als Personen der Klasse *E39 Actor*. Die Quelle liegt als digitale Ressource vor und wird mit Metadaten angereichert sowie den entsprechenden Forschungsobjekten zugewiesen. Der Quellentyp, wie Ablass, Testament, Urkunde, wird zusammen mit ihrem Fundort in den Metadaten erfasst. Es ist nicht ausreichend, das Datenmodell nur am Bauwerk oder an geographischen Informationen auszurichten, da heterogene Daten miteinander in Verbindung gebracht werden müssen und Informationen zu Bauwerk und Ereignissen beinhalten.

Derzeit orientiert sich das Modell an der Basisontologie von CIDOC CRM. Im Laufe der folgenden Projektphase wird in verschiedenen Unterontologien geprüft, welche Klassen und Eigenschaften zutreffend sein könnten, um bestimmte Angaben zu präzisieren. Hier erscheinen insbesondere CIDOC CRMba für archäologische Bauwerke (Ronzino et al. 2016), CRMdig für Herkunftsmetadaten (Doerr, Stead und Theodoridou 2016) sowie CRMgeo für räumlich-zeitliche Daten (Hiebel et al. 2015) als geeignet. Die Kategorien für die Erfassung der Forschungsobjekte entsprechen zu diesem Zeitpunkt hauptsächlich den *High Level Classes*, für die Hierarchisierung des Bauwerks werden bereits untergeordnete Klassen verwendet. Die Modellierung der Beziehungen zwischen den Forschungsobjekten dient deren Kontextualisierung und erfolgt unter Verwendung der den jeweiligen Klassen zugewiesenen Eigenschaften. Aus Gründen der Datenlogik lässt sich nicht jede Klasse mit jeder Eigenschaft verbinden. Eine feinere Granulierung erfolgt während des Fortschreitens des Projekts, da die Entwicklung des Datenmodells im Dialog mit den Erkenntnissen der Forschung erfolgt. Eine zu grobe oder kleinteilige Erfassung führt zu ungenauen oder zu geringen Informationen, die miteinander in Verbindung gebracht werden. Anpassungen werden im Projekt dokumentiert. Es besteht ebenfalls die Möglichkeit, ab Herbst 2022 das in der Weiterentwicklung befindliche Datenmodell für die semantische Annotation von 3D Artefakten der NFDI4Culture (Rossenova 2021) zu testen und für das Wilsnackprojekt anzupassen.

Die sehr heterogenen Forschungsobjekte erhalten bereits während des Erfassungsprozesses konkrete Bezeichnungen wie "Kirche St. Nikolai", "Chor", "Chorfenster I", "Hostienwunder", "1383", "Hostienverbrennung", "Johann Ellefeldt", "1552", u. s. w., so dass eine Zuordnung eindeutig ist. Nicht jedes Ausstattungsstück oder jeder Akteur der Zeit werden erfasst. Gemeinsam in der Projektgruppe erfolgt eine Priorisierung. Gleich benannte

Forschungsobjekte, wie Säulen, nicht näher bezeichnete Altäre oder die Wunderbluthostien, werden nummeriert. Eine Dokumentation erfolgt in den Metadaten der zugehörigen digitalen Ressourcen und im Datenmodell. Das Bauwerk wird durch die Kategorien Bauwerk – Gebäudeteile – immobile und mobile Ausstattung hierarchisiert. Den Forschungsobjekten werden alle zugehörigen Informationen und Dateien – egal welchen Typs – zugewiesen, z. B. werden alle Fotografien, Bauberichte und entstehenden Modelle des Chors mit dem Forschungsobjekt "Chor" verknüpft. Jede erfasste Information, egal ob Forschungsobjekt oder Datei, erhält einen PID, so dass eine Eindeutigkeit gewährleistet ist. Die Zuweisung erfolgt händisch und kann korrigiert werden. Einzelscheiben der Montagen oder des Gesamtfensters können gruppiert werden, wenn sie dieselben Bewegungen im Raum durchlaufen. Eine Erfassung erfolgt jedoch für jedes Forschungsobjekt einzeln, da sie verschiedene Ereignisse durchlaufen haben und noch können.



Abbildung 5: Wilsnack, Kirche St. Nikolai, Nördliches Querhaus, Fenster n VIII, CVMA Deutschland Potsdam/Berlin-Brandenburgische Akademie der Wissenschaften, Foto: Holger Kupfer.

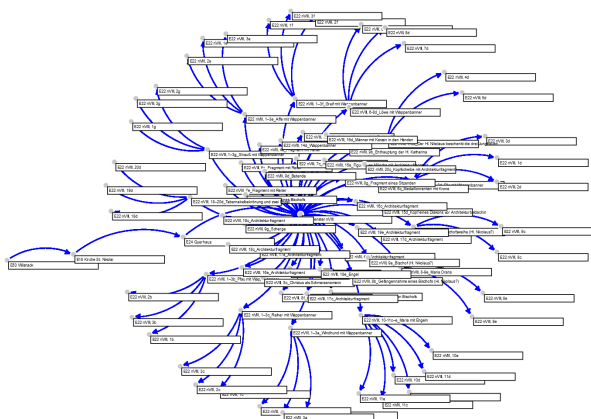


Abbildung 6: Wilsnack, Kirche St. Nikolai, Modellierung des Fenster nVIII anhand CIDOC CRM, Visualisierung: Gordon Fischer / Daten: Anja Gerber 2022.

Erschließung der Daten

Eine Annotation von digitalen Ressourcen mit Metadaten erfolgt auf der Dateiebene mit einer projekteigenen und sich in der Entwicklung befindlichen Spezifikation, die sich aus Gründen der Nachnutzbarkeit an bekannten und etablierten Standards orientiert. Der Erstentwurf verwendet das Austauschschema LIDO 1.1 (ICOM-CIDOC LIDO Working Group 2021), welches sich u. a. an CIDOC CRM, CDWA Lite (J. Paul Getty Trust 2006) und Spectrum (Collections Trust 2017) orientiert, sowie Elemente aus Dublin Core (DCMI 1995–2022, DCMI 2020). Bei LIDO handelt es sich um einen etablierten Standard für die (Meta-)Datenerfassung im Bereich des Digitalen Kulturerbes (Knaus, Stein und Kailus 2019, 12–15). Dieses Schema ist neben der Erfassung von Grafiken und Fotografien auch für Ausstattungsstücke geeignet (Knaus, Kailus und Stein 2022, 18–21). Die Spezifikation des Projekts wird im Hinblick auf die Annotation der zu erstellenden 3D-Modelle um für diese erforderliche Angaben erweitert. Hier erscheint der Standard CARARE 2.0 als geeignet, da technische Angaben (digitale Herkunftsmetadaten) und beschreibende Informationen für Gebäude, Artefakte und *born-digital* Objekte möglich sind (Ferne, Gavriliis und Angelis o. J.). Mit LIDO können Metadaten desselben Typs unter Zuweisung verschiedener Kategorien und Rollen erfasst werden, z. B. ob jemand an der Erstellung einer Ressource beteiligt war oder lediglich auf dieser abgebildet ist. Administrative, beschreibende und technische Metadaten werden abgedeckt. So können die Datei selbst aber auch ihr Erstellungsprozess beschrieben werden. Die Konfiguration des Metadaten-schemas erfolgt ebenfalls über eine JSON-Datei. Anpassungen werden dokumentiert und versioniert. Um eine eindeutige Zuordnung zu Akteuren, Ereignissen, Orten oder auch eine zweifelsfreie Klassifizierung der Objekte zu ermöglichen, werden bereits etablierte Normdatenvokabulare, wie GND, Wikidata, Iconclass, Getty Art & Architecture Thesaurus oder Geonames, genutzt und die Identifier erfasst. Zudem gibt es Überlegungen, Normdaten im Rahmen der NFDI4Culture über Wikibase (Rosse-

nova u. a. 2021) zu erfassen und als Linked Open Data in den Wissensgraphen (NFDI4Culture o. J.) einzuspeisen.

Bibliographie

AriInfoWare GmbH. "Aktuelle Informationen – MonArch". In *MonArch*. 2022. Zugriffen 28. Juli 2022. <https://openmonarch.org/informationen/>.

Bednarz, Ute, Eva Fitz, Peter Knüvener, Frank Martin, Markus Mock, Götz J. Pfeiffer und Martina Voigt. "Die mittelalterlichen Glasmalereien in Berlin und Brandenburg." *Corpus Vitrearum Medii Aevi Deutschland XXII*. Berlin: Akademie Verlag, 2010.

Bekiari, Chrysoula, George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, und Athanasios Velios. "Volume A: Definition of the CIDOC Conceptual Reference Model". CIDOC CRM Special Interest Group. Juni 2022. <https://cidoc-crm.org/versions-of-the-cidoc-crm>.

Burricher, Brigitte, Björn Gebert, Christoph Mackert, und Gabriel Viehhauser. "Digitale Mediävistik". *Das Mittelalter. Perspektiven mediävistischer Forschung* 26, Nr. 1, (2021): 101–117. <https://doi.org/10.17885/heup.mla.2021.1.24312>.

Collections Trust. "Spectrum 5.0. All Procedures". 2017. Zugriffen 28. Juli 2022. <https://collectionstrust.org.uk/spectrum/procedures/>.

digilib Community. "digilib - The Digital Image Library". 2001–2022. Zugriffen 28. Juli 2022. <https://robcast.github.io/digilib/>.

Doerr, Martin, Stephen Stead und Maria Theodoridou. "Definition of the CRMdig. An Extension of CIDOC-CRM to support provenance metadata." Proposal for approval by CIDOC CRM-SIG. Version 3.2.1. April 2016. https://cidoc-crm.org/crm-dig/sites/default/files/CRMdig_v3.2.1.pdf.

Dublin Core Metadata Initiative (DCMI). "Specifications". 1995–2022. Zugriffen 28. Juli 2022. <http://dublin-core.org/specifications/dublin-core/>.

Dublin Core Metadata Initiative (DCMI). "DCMI Metadata Terms." 2020. Zugriffen 28. Juli 2022. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>.

Ferne, Kate, Dimitris Gavriliis, und Stavros Angelis. "The CARARE Metadata Schema, v2.0". o. J. Zugriffen 28. Juli 2022. <http://3dicons-project.eu/wp-content/uploads/2018/08/The-CARARE-metadata-schema2.pdf>.

Fischer, Gordon und Christian Thomas. "Alexander von Humboldt auf Reisen: Chronotopische Zugänge zur edition humboldt digital." *vDHd2021 – Experimente*. 2021. Zugriffen 28. Juli 2022. <https://vdhd2021.hypotheses.org/292>.

Germanisches Nationalmuseum. „Features | wisski.eu“. *WissKI – a scientific communication infrastructure*. o. J. Zugriffen 28. Juli 2022. <https://wisski.eu/features>.

Gerber, Anja. "Forschungsdateninfrastruktur für multimodale digitale Daten- und Quellentypen am Beispiel des Standortes Wilsnack, St. Nikolai." Masterarbeit, Fachhochschule Potsdam und Humboldt Universität zu Berlin, 2022.

Gerber, Anja und Gordon Fischer. "CVMA Foto Manager - ein Open-Source-Metadateneditor für die Erschließung von Bilddaten." *NFDI4Culture Community Plenary*. 17.-19.11.2021. Zugriffen 28. Juli 2022. <https://nfdi4culture.de/de/aktuelles/nachrichten/second-culture-community-plenary-abstracts-of-lightning-talks-and-presentations.html>.

ICOM-CIDOC LIDO Working Group. "LIDO - Lightweight Information Describing Objects Version 1.1." *LIDO v1.1 Documentation*. 20. Dezember 2021. <https://lido-schema.org/schema/v1.1/lido-v1.1.html>.

Institut für Museumsforschung der Staatlichen Museen zu Berlin - Preußischer Kulturbesitz. "SPECTRUM 3.1. The UK Museum Documentation Standard, deutsche erweiterte Fassung." *Materialien aus dem Institut für Museumsforschung - Sonderheft 5*. Berlin, 2013. https://www.smb.museum/fileadmin/website/Institute/Institut_fuer_Museumsforschung/Publikationen/Materialien/Sonderhefte/mat-Sonderheft_5-SPECTRUM_3.1.pdf.

J. Paul Getty Trust. "CDWA Lite. Specification for an XML Schema for Contributing Records via the OAI Harvesting Protocol 1.1". 2006. Zugriffen 28. Juli 2022. https://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.pdf.

Hiebel, Gerald, Martin Doerr, Øyvind Eide, und Maria Theodoridou. "CRMgeo: a Spatiotemporal Model. An Extension of CIDOC-CRM to link the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement." Proposal for approval by CIDOC CRM-SIG. Version 1.2. September 2015. https://www.cidoc-crm.org/crmgeo/sites/default/files/CRMgeo1_2.pdf.

Knaus, Gudrun, Regine Stein und Angela Kailus. "LIDO-Handbuch für die Erfassung und Publikation von Metadaten zu kulturellen Objekten. Band 1: Graphik." Heidelberg: arthistoricum.net, 2019. <https://doi.org/10.11588/arthistoricum.382.544>.

Knaus, Gudrun, Angela Kailus und Regine Stein. "LIDO-Handbuch für die Erfassung und Publikation von Metadaten zu kulturellen Objekten. Band 2: Malerei und Skulptur". Heidelberg: arthistoricum.net, 2022. <https://doi.org/10.11588/arthistoricum.1026>.

Kühne, Hartmut und Anne-Katrin Ziesak. "Wunder - Wallfahrt - Widersacher. Die Wilsnackfahrt". Regensburg: Pustet Verlag, 2005.

NFDI4Culture. "Aufgabenbereich 5. Technologien, Recht & Ethik. Übergreifende technische, ethische und rechtliche Aktivitäten". o. J. Zugriffen 28. Juli 2022. <https://nfdi4culture.de/de/aufgaben/aufgabenbereiche/aufgabenbereich-5.html>.

Ronzino, Paola, Franco Niccolucci, Achille Felicetti, und Martin Doerr. "Definition of the CRMba. An extension of CIDOC CRM to support buildings archaeology documentation." Proposal for approval by CIDOC CRM-SIG. Editorial Status: Under Revision since 3.12.2016. Version 1.4. Dezember 2016. https://cidoc-crm.org/crmba/sites/default/files/2016-12-3%23CRMba_v1.4.1_UR.pdf.

Rossenova, Lozana. "Semantic Annotation for 3D Cultural Artefacts: MVP." 29. Oktober 2021. <https://doi.org/10.5281/zenodo.5628847>.

Rossenova, Lozana, Zoe Schubert, Richard Vock, und Ina Blümel. "Beyond the render silo - Semantically anno-

tating 3D data within an integrated knowledge graph and 3D-rendering toolchain". Potsdam, 7. März 2022. <https://doi.org/10.5281/zenodo.6328155>.

Universität zu Köln, Department of Digital Humanities. "Kompakkt - 'cause the world is multidimensional." 2018-2022. Zugriffen 28. Juli 2022. <https://kompakkt.de/home>.

»Die Greta Garbo der Leichtathletik« – Eine systematische Analyse der Modifizervossianischer Antonomasien mithilfe von Word Embeddings

Schwab, Michel

michel.schwab@hu-berlin.de
Humboldt-Universität zu Berlin

Fischer, Frank

fr.fischer@fu-berlin.de
Freie Universität Berlin

Einführung und Forschungsstand

Die vossianische Antonomasie (VA) ist ein rhetorisches Stilmittel aus der Familie der Antonomasien, eng verwandt mit Metonymie und Metapher. Während bei der klassischen Antonomasie ein Eigenname durch eine typische Eigenschaft ersetzt wird (wenn etwa Michael Schumacher als »der Kerpener« bezeichnet wird), funktioniert die vossianische Antonomasie genau umgekehrt. Hier wird ein typisches Merkmal einer Person durch den Eigennamen einer anderen Person evoziert.

Wenn ein Journalist zum Beispiel Wilson Kipketer, den dänischen Mittelstreckenläufer kenianischer Herkunft, als »Greta Garbo der Leichtathletik« bezeichnet, wird eine typische Eigenschaft der Filmschauspielerin aufgerufen, in diesem Fall ihre distanzierte, zurückhaltende Art, wie dieses Zitat aus der New York Times zeigt: »Kipketer is as guarded [zurückhaltend] as he is fast; some reporters have labeled him the Greta Garbo of track and field.« (NYT, 8. August 1997).

Eine vossianische Antonomasie setzt sich im Normalfall aus drei Teilen zusammen: dem Target (Wilson Kipketer), der Source (Greta Garbo) und dem Modifier (Leichtathletik) (vgl. Bergien 2013). Der Modifier verschiebt eines oder mehrere Merkmale der Source in das Umfeld des

Targets. In dieser Arbeit konzentrieren wir uns auf die systematische Analyse des Modifiers.

Die automatisierte Erkennung und Extraktion vossianischer Antonomasien hat sich in den letzten fünf Jahren rasch ausdifferenziert. Während Jäschke et al. 2017 und Fischer et al. 2019 semi-automatisierte Verfahren nutzten, um VA-Ausdrücke in großen Zeitungskorpora ausfindig zu machen, setzten Schwab et al. (2019, 2022) automatisierte Verfahren ein, die meist auf neuronalen Netzen basierten.

Da wir mit einem großen Korpus und Word Embeddings arbeiten, ist unser Forschungsbeitrag der erste, der eine quantitative Untersuchung dieses Phänomens mit einer thematischen Gruppierung der verschiedenen Modifier verbinden kann. Unsere Forschungsergebnisse stellen wir auch über eine interaktive Visualisierung bereit (<https://vossanto.weltliteratur.net/dhd2023/modifier.html>).

Datensatz

Wir nutzen den VA-Datensatz aus Schwab et al. 2019, welcher mittels eines semi-automatisierten Verfahrens aus dem New York Times -Korpus (Sandhaus 2008) generiert wurde. Das NYT-Korpus besteht aus mehr als 1,8 Mio. Zeitungsartikeln der NYT aus den Jahren 1987–2007. Mit Hilfe des regulären Ausdrucks

```
\\ b(the|an?)\\s+([\\w.,'-]+\\s+){1,5}? (of|for|among)\\b
```

wurden Kandidatensätze ermittelt, d.h. alle Sätze, die Phrasen enthalten, welche mit »the«, »a« oder »an« anfangen und mit »of«, »for« oder »among« enden, wobei zwischen diesen beiden Polen ein bis fünf Wörter platziert sein können. Die Wörter zwischen Anfangs- und Endsignal stellten die potenzielle Source-Phrase dar und wurden mit einer Wikidata-Liste abgeglichen, die alle Entitätennamen aus Wikidata (inkl. Aliasse) enthielten, die die Eigenschaft »instanceOf« »human« aufweisen. Die Source-Kandidaten wurden also auf Menschen beschränkt, die in Wikidata verzeichnet sind (dabei handelt es sich um eine bewusste Beschränkung bei der Untersuchung des Phänomens – VA können auch mit Orten, Markennamen, Comicfiguren usw. operieren). Anschließend wurden diese Kandidaten mit einer manuell erstellten Sperrliste abgeglichen, um falsche Kandidaten auszuschließen.

Dieser Datensatz wurde in Schwab et al. 2022 verfeinert. Alle VA-Phrasen (Target, Source, Modifier) wurden auf Wortebene innerhalb der Sätze annotiert. Insgesamt enthält der Datensatz 5.995 Sätze, davon enthalten 3.066 einen VA-Ausdruck und 2.929 enthalten keinen, sind aber syntaktisch ähnlich aufgrund des genutzten regulären Ausdrucks.

In Tabelle 1 sind die zehn häufigsten Modifier des Datensatzes aufgelistet. Die häufigsten Ausdrücke sind temporale Ausdrücke (»his day«, »his time«, »the 90s«), geografische Angaben (»Japan«, »China«) und Sportarten (»tennis«, »baseball«, »ballet«).

Tabelle 1: Die zehn häufigsten Modifier im Datensatz inklusive ihrer Häufigkeit.

Modifier	Anzahl
his day	56
his time	35
Japan	32
the 90s	21
China	17
our time	17
tennis	16
his generation	16
baseball	16
her time	14

Methode

Wir nutzen kontextabhängige Word Embeddings, um die Modifier-Phrasen in hochdimensionale Vektoren zu transformieren, die die Semantik des Textes wiedergeben sollen.

Mit Hilfe von Word Embeddings wurden in den letzten Jahren viele Benchmarks im Bereich Natural Language Processing erstellt. Insbesondere kontextabhängige Word Embeddings, d.h. die numerische Repräsentation von Wörtern und Tokens in Abhängigkeit ihres Kontexts, haben viel Aufmerksamkeit auf sich gezogen. Der Vorteil dieser Word Embeddings im Gegensatz zu kontextunabhängigen Word Embeddings ist die Möglichkeit, Homonyme korrekt darzustellen. Wir benötigen die numerische Repräsentation der Phrasen, um anschließend ein Clustering-Verfahren durchführen zu können, welches die Modifier in Themenbereiche gruppieren soll.

Wir greifen auf Sentence-Transformers zurück, welches aus Sentence-BERT (Reimers et al. 2019) hervorgegangen ist. Das Modell basiert auf transformerbasierten Sprachmodellen wie BERT (Devlin et al. 2019). Im Gegensatz zu BERT wird S-BERT allerdings mittels einer siamesischen Netzwerkstruktur trainiert, der Output wird durch eine Pooling Operation in einen hochdimensionalen Vektor transformiert. Dadurch kann das trainierte Modell effizient semantische Ähnlichkeiten zwischen Texten errechnen. Wir nutzen das Modell »all-mpnet-base-v2«, welches die besten Resultate in der Anwendung auf verschiedene Datensätze zeigte (siehe https://www.sbert.net/docs/pretrained_models.html).

Dies wenden wir auf die einzelnen Modifier an. Das Netzwerk liefert für jeden Modifier einen 768-dimensionalen Vektor. Diese numerischen Vektoren lassen sich nun durch ein Clustering-Verfahren gruppieren.

Wir entscheiden uns für den k-means-Algorithmus (MacQueen 1967), um die Vektoren in Cluster einzuteilen. Wir nutzen k-means aufgrund verschiedener Annahmen. Einmal gehen wir davon aus, dass es relativ wenige Ausreißer gibt, da die VA-Ausdrücke aus dem Datensatz häufig in ähnlichen Themengebieten in der New York Times vorkommen (vgl. Fischer et al. 2019). Außerdem können wir die Anzahl der Cluster angeben und diese während der Analyse variieren, um zu beobachten, wie sich die Gruppierungen in Abhängigkeit davon verhalten. Dies funktioniert mit dichtebasierten Clustering-Algorithmen nicht so einfach. Da k-means in der Berechnung der Cluster die quadrierte euklidische Distanz nutzt, nor-

malisieren wir die Output-Vektoren, da die normalisierte quadratische euklidische Distanz proportional zur Kosinus-Distanz ist, welche in Reimers et al. 2019 genutzt wird, um die Ähnlichkeit zwischen zwei Vektoren zu berechnen.

Im Anschluss an das Clustering möchten wir den einzelnen Clustern Themen zuordnen, durch ein an das »Topic Modeling« angelehntes Verfahren. Stark vereinfacht basieren klassische Topic-Modeling-Modelle auf der Annahme, dass Wörter, die besonders häufig gemeinsam in Sequenzen vorkommen, ein abstraktes Thema bilden. Meist wird Topic Modeling auf längere Dokumente angewandt, bei denen von signifikanten Wort-Überschneidungen ausgegangen werden kann. 97 Prozent der Modifier-Phrasen bestehen jedoch aus einem bis vier Wörtern und weisen dadurch kaum Überschneidungen auf. Somit sind sie für klassisches Topic Modeling ungeeignet. Selbst beim sogenannten Short Text Topic Modeling wird mit Textsorten wie Tweets oder Rezensionen trainiert, welche immer noch bedeutend länger sind als unsere Phrasen.

Wir nutzen stattdessen den Vorteil, dass viele der Formulierungen Nominalphrasen oder Nomen sind. Dadurch sind sie unter anderem im WordNet (Fellbaum 1998) zu finden, einer lexikalischen Datenbank, die Wortbedeutungen, Synonyme und viele andere Features bereitstellt. Das Projekt WordNet Domains (Bentivogli et al. 2004) hat zusätzlich jedem Wort bzw. jedem Synset (Gruppe ähnlicher Wörter) in WordNet semi-automatisch ein oder mehrere Domains zugeordnet, die für uns als Themengebiete genutzt werden können. Diese Domains sind hierarchisch gegliedert. Dies nutzen wir aus und weisen jeder Modifier-Phrase, soweit vorhanden, ihre Domains zu. Vorher nutzen wir noch das NLTK Toolkit (Bird et al. 2009), um alle Stoppwörter zu entfernen und die Ausdrücke dadurch auf die Nomen zu reduzieren, zum Beispiel »her time« zu »time« oder »the harmonica« zu »harmonica«. Sollte die übrigbleibende Phrase im WordNet nicht vorhanden sein, teilen wir sie in ihre einzelnen Wörter auf und verfahren wie oben beschrieben für jedes Wort der Phrase. Zum Schluss weisen wir die am häufigsten vorkommende Domain der Phrasen je Cluster dem jeweiligen Cluster als Themengebiet zu. In der Web-App kann man sich zusätzlich die zehn hochfrequentesten Domains anschauen.

Anschließend können wir die Cluster visualisieren. Da die Vektoren hochdimensional sind, nutzen wir verschiedene Dimensionsreduktionsalgorithmen, um sie auf zwei Dimensionen zu reduzieren. Wir vergleichen mehrere Algorithmen – PCA (Hauptkomponentenanalyse, Pearson 1901), t-SNE (t-distributed stochastic neighbor embedding Methode, van der Maaten 2008), UMAP (Uniform Manifold Approximation and Projection, McInnes et al. 2018), IVIS (Szubert et al. 2019) –, welche in der Web-App ausgewählt werden können. Nach einigen Durchläufen hat sich die Kombination von PCA und t-SNE als bestes Verfahren herausgestellt, welches wir kurz vorstellen. Wir wenden zuerst PCA an und reduzieren die Vektoren auf eine Länge von 50. Die Hauptkomponentenanalyse vereinfacht Daten, indem die Einträge der Vektoren durch eine geringere Zahl möglichst aussagekräftiger Linearkombinationen (die Hauptkomponenten) genähert werden. Zusätzlich nutzen wir t-SNE, um die 50-dimensionalen Vektoren auf zweidimensionale Vektoren

zu reduzieren. Der Vorteil von t-SNE im Gegensatz zu PCA liegt in der Möglichkeit, nichtlineare Abhängigkeiten darzustellen. t-SNE reduziert die Vektoren außerdem so, dass Vektoren, die in der höheren Dimension eine kurze Distanz haben, auch in der reduzierten Dimension eine kurze Distanz zueinander haben. Dadurch wird die lokale wie auch globale Struktur bewahrt.

Erkenntnisse

Die Modifier für vossianische Antonomastien fallen im New York Times -Korpus sehr vielgestaltig aus und sind nicht auf bestimmte Phrasen oder Wortgruppen limitiert. Abbildung 1 zeigt beispielhaft die Visualisierung mit neun Clustern, wobei die Themen von uns zunächst manuell zugeordnet wurden. Einige der Cluster lassen sich bis auf wenige Ausnahmen eindeutig bestimmten Themen zuordnen, wie Sport, Musik und Tanz, Kunst, Film und Literatur, Naturwissenschaft, Geografie, Politik, Wirtschaft oder temporale Ausdrücke.

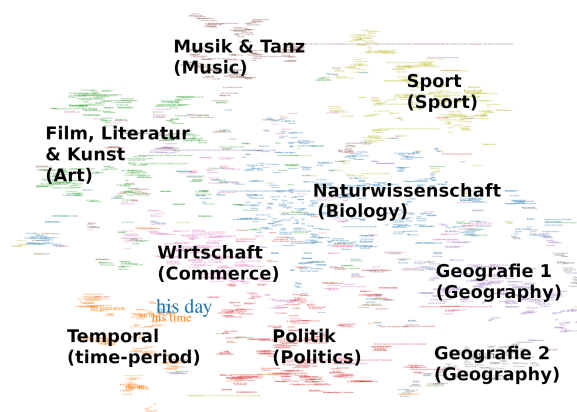


Abbildung 1: Die Abbildung zeigt die Visualisierung des Clusterings nach Dimensionreduktion mit neun Clustern. Den Clustern wurde manuell ein Themengebiet zugeordnet.

Die automatisch gefundenen Themen durch WordNet Domains stimmen mit den von uns manuell zugewiesenen Themen fast vollständig überein, wie Abbildung 1 zeigt. Den von uns manuell annotierten Themen sind in Klammern die automatisch gefundenen Themen beige-sellt.

Der blaue Cluster in der Mitte ist allerdings sehr divers und lässt sich nicht genau einer Kategorie zuordnen. Hier tauchen Flora und Fauna auf (»the pumpkins«, »Rottweilers«), was zu dem automatisch gefundenen Themengebiet Biologie passen würde. Allerdings kommen auch »space wear«, »Buddhism« sowie »soft drinks« und »the physics world« vor. Wir haben uns für das Rubrum Naturwissenschaft entschieden, welches auf einen Großteil der Phrasen zutrifft.

Einer der Gründe für die Zusammensetzung dieses diversen Clusters ist der Umstand, dass k-means keine Ausreißer zulässt und daher jeder Punkt genau einem Cluster zugeordnet wird. Dadurch finden sich auch Phrasen, die eigentlich nicht in ein bestimmtes Themengebiet gehören oder für die es eigentlich zu wenige ähnliche Phrasen gibt, in einem Cluster wieder.

Ein anderer Grund ist die Diversität der Modifier. Viele Modifier bestehen nicht nur aus einem, sondern aus mehreren Wörtern. Diese Phrasen könnte man verschiedenen Themengebieten oder Subgenres zuordnen, z.B. »ancient Alexandria« (Temporal, Geografie), »Korean radio« (Geografie, Technologie) oder »food writing« (Speisen und Getränke, Literatur). Dies sind auch häufig Phrasen, die durch die Dimensionsreduktion nicht in der Nähe der anderen Phrasen des Clusters liegen, weil zum Beispiel »food writing« in die Nähe von anderen kulinarischen Phrasen verortet wurde, obwohl es ein Subgenre der Literatur ist. An diesem Beispiel sieht man, dass das Clustering-Verfahren das Wort richtig zugeordnet hat (»food writing« gehört zum kulturellen Cluster), aber falsch visualisiert wurde.

Abhängig von der Anzahl der Cluster unterteilt sich zum Beispiel die Kultur nach und nach in Subgenres wie Literatur, Musik, Tanz, Film/TV oder Kunst. Dies kann man in Abbildung 2 gut beobachten. Der linke Teil von Abbildung 2 zeigt einen Ausschnitt der Visualisierung, in der sechs Cluster gebildet wurden. Hier sind die meisten kulturbezogenen Phrasen in einem einzigen Cluster (grün) gruppiert. Im rechten Teil ist der gleiche Ausschnitt zu sehen, allerdings mit zwölf Clustern. Man kann gut erkennen, dass sich der kulturelle Cluster fast vollständig in drei neue Cluster (grau, orange, blau) aufgeteilt hat, nämlich »Kunst«, »Literatur und Film/TV« und »Musik und Tanz«. Auch hier gibt es Grenzfälle wie zum Beispiel »musicals«. Das Clustering hat die Phrase zu »Musik und Tanz« gruppiert, wohingegen in der Visualisierung das Wort in die Nähe des Film-Clusters gerückt wurde, in dem auch theaterbezogene Themen auftauchen.

Auch die Geografie teilt sich mit wachsender Anzahl an Clustern in zwei Hälften, wobei in der einen ein Großteil der US-amerikanischen Geografika angesiedelt sind, allerdings nicht ausschließlich.

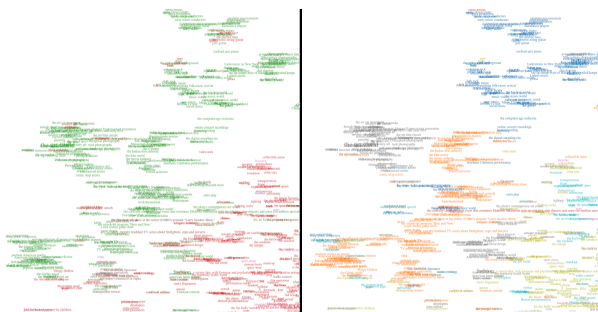


Abbildung 2: Die Abbildung zeigt einen Ausschnitt der Visualisierung mit sechs Clustern auf der linken Seite und zwölf Clustern auf der rechten Seite. Der Ausschnitt zeigt den Großteil der kulturellen Phrasen.

Wie oben bereits angemerkt, stellen wir eine interaktive Visualisierung zur Datenexploration zur Verfügung, in der die oben beschriebenen Fälle nachvollzogen werden können. Die verschiedenen Dimensionsreduktionsverfahren können selbst ausprobiert und die Anzahl der Cluster variiert werden (1–15). Außerdem werden die zehn am häufigsten vorkommenden Domains je Cluster gezeigt, um einen Überblick über die Themen zu bekommen. Die Größe der Labels spiegelt die Anzahl der Vorkommen im Datensatz wider. Zudem kann durch

eine Bereichsauswahl gezoomt werden (<https://vossanto.weltliteratur.net/dhd2023/modifier.html>).

Fazit und Ausblick

Unser Ansatz lenkt den Blick von den Eigennamen in Source und Target einer vossianischen Antonomasie auf den Modifier. Wir konnten zeigen, dass bestimmte Themenfelder besonders häufig sind, also eine besondere Neigung aufweisen, in einer vossianischen Antonomasie Verwendung zu finden. Die Themen wurden in Clustern gruppiert und zweidimensional projiziert. Durch verschiedene Verfahren kann das Modell noch verfeinert werden, z.B. durch den Einsatz anderer Cluster- oder Reduktionsverfahren.

Mit Hilfe von Entity Embeddings kann man in Zukunft ähnliche Analysen der Source und des Targets durchführen, um etwa auf Zusammenhänge zwischen den einzelnen Teilen einer vossianischen Antonomasie zu fokussieren. So würde sich zum Beispiel erforschen lassen, in welchen semantischen Abhängigkeiten Source, Target und Modifier eines VA-Ausdrucks zueinander stehen und welche Entitäten signifikant häufig mit welchen Modifier-Gruppen genutzt werden.

Mit Hilfe der Web-App kann man die Daten und Ergebnisse interaktiv explorieren und somit weitere Erkenntnisse erlangen, welche für die automatische Erkennung, aber auch für die automatische Generierung sinnvoller vossianischer Antonomasien eine wichtige Rolle spielen können. Beide Aufgaben verfolgen wir in Zukunft.

Bibliographie

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini und Emanuele Pianta. 2004. "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing." In: COLING 2004 Workshop on "Multilingual Linguistic Resources", Geneva, Switzerland, S. 101–108.

Bergien, Angelika. 2013. "Names as frames in current-day media discourse." In: Name and Naming. Proceedings of the Second International Conference on Onomastics. Cluj-Napoca: Editura Mega. S. 19–27.

Bird, Steven, Edward Loper und Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. 2019. "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, Volume 1 (Long and Short Papers), Minneapolis, Minnesota. <https://doi.org/10.48550/arXiv.1810.04805>

Fellbaum, Christiane (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fischer, Frank und Robert Jäschke. 2019. "'The Michael Jordan of greatness' – Extracting Vossian antonomasia from two decades of The New York Times, 1987–2007." In: Digital Scholarship in the Humanities 35, no. 1. S. 34–42. <https://doi.org/10.1093/llc/fqy087>

Jäschke, Robert, Jannik Strötgen, Elena Krotova und Frank Fischer. 2017. "'Der Helmut Kohl unter den

Brotaufstrichen'. Zur Extraktion vossianischer Antonomasiën aus großen Zeitungskorpora." In: Proceedings of DHd 2017 . Universität Bern. <https://doi.org/10.5281/zenodo.4646126>

MacQueen, J. 1967. "Classification and analysis of multivariate observations." In: 5th Berkeley Symp. Math. Statist. Probability . S. 281–297.

McInnes, Leland, John Healy und James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv preprint arXiv:1802.03426.

Pearson, Karl. "LIII. 1901. On lines and planes of closest fit to systems of points in space." In: The London, Edinburgh, and Dublin philosophical magazine and journal of science 2, no. 11. S. 559–572. <https://doi.org/10.1080/14786440109462720>

Reimers, Nils und Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , Hong Kong, China.

Sandhaus, Evan. 2008. "The New York Times Annotated Corpus." LDC2008T19. Philadelphia: Linguistic Data Consortium. <https://doi.org/10.35111/77ba-9x74>

Schwab, Michel, Robert Jäschke, Frank Fischer und Jannik Strötgen. 2019. "A Buster Keaton of Linguistics: First Automated Approaches for the Extraction of Vossian Antonomasia." In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , Hong Kong, China.

Schwab, Michel, Robert Jäschke und Frank Fischer. 2022. "'The Rodney Dangerfield of Stylistic Devices' – End-to-End Detection and Extraction of Vossian Antonomasia Using Neural Networks." In: Frontiers in Artificial Intelligence 5. <https://doi.org/10.3389/frai.2022.868249>

Szubert, Benjamin, Jennifer E. Cole, Claudia Monaco und Ignat Drozdov. 2019. "Structure-preserving visualisation of high dimensional single-cell datasets." In: Scientific reports , 9 (1), 1–10. <https://doi.org/10.1038/s41598-019-45301-0>

van der Maaten, Laurens und Geoffrey Hinton. 2008. "Visualizing Data using t-SNE." In: Journal of Machine Learning Research 9: 2579–2605.

Die historische Konfliktsimulation als wissenschaftliches Modellierungsproblem in der Lehre

Wintjes, Jorit

gorit.wintjes@uni-wuerzburg.de
Julius-Maximilians-Universität, Deutschland

Pielström, Steffen

pielstroem@biozentrum.uni-wuerzburg.de
Julius-Maximilians-Universität, Deutschland

Bock, Sina

sina.bock@uni-wuerzburg.de
Julius-Maximilians-Universität, Deutschland

Modellierung und Simulation in den Digital Humanities

Die Modellierung von Forschungsgegenständen sowie die Operationalisierung von Forschungsfragen sind zentrale Bestandteile geisteswissenschaftlicher Forschung (Beynon et al. 2006, Flanders and Jannidis 2015, Thaller 2017, Piotrowski, 2019). Mit Einzug der Digitalisierung können Forschungsgegenstände in strukturierte Forschungsdaten transformiert und durch diese formalisierte Modellierung mit informationstheoretischen Methoden untersucht werden. Die Kompetenzen, die für eine adäquate konzeptionelle und formale Modellierung sowie für eine quantitative Analyse geisteswissenschaftlicher Forschungsinteressen erforderlich sind, bilden den "fächerübergreifende[n] Kern der Digital Humanities" (Thaller 2017, S. 16).

Die Fähigkeit zur kritischen Auseinandersetzung mit den Vorannahmen und Auswirkungen der Reduktion, die für die formale Beschreibung von Forschungsgegenständen notwendigerweise einhergeht, ist mehreren geisteswissenschaftlichen Disziplinen gemein und von zentraler Bedeutung (Thaller 2017, Ciula et al, 2018, Piotrowski 2019).

In den Geschichtswissenschaften findet sich diese Auseinandersetzung einerseits in Diskussionen über die Anwendung formaler Methoden zur Beschreibung und Analyse historischer Phänomene,¹ andererseits aber auch in Reflektionen über Möglichkeiten der Vermittlung dieser Methodik in der Lehre. Ein Beispiel hierfür bilden analoge und digitale Simulationen, die in den Geschichtswissenschaften seit Langem als didaktische Methode eingesetzt werden, um die Modellierung von historischen Konflikten zu erlernen (Sabin 2011, Sabin, 2012, Sabin 2016).

Konfliktsimulationen als didaktisches Instrument im Unterricht

Konfliktsimulationen sind ein äußerst leistungsfähiges didaktisches Instrument für den akademischen und nicht-akademischen Unterricht; auf diese Weise eingesetzt können sie zu den *serious games* gezählt werden (Jones 1995; Michael and Chen 2006). Sie stellen ein Element partizipatorischen Unterrichts dar und eröffnen beispielsweise im Fach Geschichte den Teilnehmenden einen Zugang zu den für die Vorbereitung der Simula-

tion notwendigen Quellenmaterialien, der sich stark von einer eher rezeptiven Wissensaufnahme unterscheidet: für die Teilnahme an einer Konfliktsimulation, in deren Mittelpunkt immer die Verarbeitung von Information und das Entscheiden auf der Basis dieser prozessierten Information steht, ist eine auf die Ziele der Simulation hin ausgerichtete Aufbereitung des jeweils relevanten Quellenmaterials notwendig. Hierdurch bieten sich den Teilnehmenden Perspektiven, die ohne eine derartige Aufbereitung kaum deutlich würden.

Ein weiteres wichtiges Einsatzgebiet ist die Auseinandersetzung mit den Problemen, die bei der Prozessierung von Information, der darauf basierenden Entscheidungsfindung und anschließenden Kommunikationsprozessen entstehen. Hier können Konfliktsimulationen eindrücklicher aufzeigen, als dies in anderen Unterrichtsformen möglich wäre, wie durch dysfunktionale Entscheidungsprozesse Handlungsmöglichkeiten von Akteuren eingeschränkt werden.

Neben der Teilnahme führt schließlich die Erstellung einer Konfliktsimulation zu einer deutlich vertieften Auseinandersetzung mit dem gegebenen Phänomen (Sabin 2012, Sabin 2016). Bei einem erfolgreichen Simulationsdesign handelt es sich immer um eine stark abstrahierende Modellierung einer sehr komplexen Realität; der Versuch einer solchen Modellierung beinhaltet daher immer auch eine Auseinandersetzung mit der Frage, welche Faktoren Eingang in die Simulation finden sollen, und welche unberücksichtigt bleiben können.

Über die Frage nach Perspektivbildung und Modellierungsleistung der Teilnehmenden hinaus führt die bei der Durchführung von Simulationen schließlich immer wieder zu beobachtender Immersion zu einer sinnvollen Auflockerung des Unterrichts.

Konfliktsimulationen und das Problem der Zugänglichkeit im Unterricht

“Sheldon: I am here to sit with you and keep you company. - Bernadette: Oh, that's nice. - Sheldon: Yeah, by playing the most complicated board game ever invented: Campaign for North Africa. I bought it off eBay. It smells a little like chili, but all the pieces are there.”²

Ungeachtet ihrer Vorteile ist der Einsatz von Konfliktsimulationen im Unterricht mit einem zentralen Problem verbunden: der Zugänglichkeit für die Teilnehmenden. Jede Simulation stellt eine verregelte Reduktion der Realität dar – die vollständige Abbildung eines realweltlichen Phänomens in einer Simulation ist unmöglich –, bei der komplexe Geschehen in der Regel auf wenige Faktoren reduziert werden. Das Ineinandergreifen dieser Faktoren wird dann durch ein Regelwerk abgebildet, das der Simulation zugrunde liegt. Als Konsequenz ist eine erfolgreiche Teilnahme an einer Simulation nur auf der Grundlage belastbarer Regelkenntnis möglich. Konkret gesprochen bedeutet der Einsatz beispielsweise des Schachspiels, dass alle Teilnehmenden dessen Regelwerk beherrschen müssen.

Dieser Umstand hat drei wichtige Konsequenzen. Zum ersten wirken Simulationen umso abschreckender, je komplexer die Regelwerke sind. Aus diesem Grund ist der weitaus größte Teil der kommerziell erhältlichen Konfliktsimulationen für eine Unterrichtssituation im Normalfall denkbar ungeeignet (Sabin 2012); das im voranstehenden Zitat genannte *The Campaign for North Africa*³ nimmt zwar bis heute eine Ausnahmestellung ein, die allermeisten kommerziell erhältlichen Konfliktsimulationen sind aber ohne eine längere Einweisung nicht durchführbar. Eine zweite wichtige Konsequenz liegt in der Auswirkung, die das Regelwerk auf die Teilnehmenden an der Simulation hat. In den meisten Fällen interagieren diese aus dem Bemühen heraus, keine Fehler zu machen, mehr mit dem Regelwerk als mit der eigentlichen Simulationssituation. Dabei besteht die Gefahr, dass die Teilnehmer sich stärker auf das „korrekte Bedienen“ des Regelwerks konzentrieren als auf den eigentlichen Inhalt der Simulation. Zum dritten führen die sich aus dem Einsatz komplexer Regelwerke ergebenden Notwendigkeiten häufig dazu, dass Konfliktsimulationen – wenn überhaupt – nur in sehr einfacher Form im Unterricht Einsatz finden (Wintjes und Pielström 2018).

Das grundsätzliche Dilemma der Unvereinbarkeit von leichter Zugänglichkeit und komplexer Simulation lässt sich auch durch eine digitale Simulation nicht lösen, da auch hier eine Zunahme an Komplexität der Simulation immer mit einer Zunahme an Komplexität der Bedienung verbunden sein wird; in dieser Hinsicht verhalten sich digitale Simulationen nicht anders als analoge Simulationen.

facilitator-based simulations (FBS) im Unterricht

Einen Ausweg aus dem oben beschriebenen Dilemma bieten sogenannte *facilitator-based-simulations* (FBS), bei denen es sich um die älteste Form edukativer Konfliktsimulationen handelt (Wintjes und Pielström 2019, Wintjes 2022). Bei diesen ist eine Regelkenntnis der Teilnehmer nicht nötig; diese interagieren nicht direkt mit dem Regelwerk, vielmehr geben sie Entscheidungen, Aufträge oder Handlungsanweisungen an das Leitungsgremium der Simulation, die *facilitator*, heraus, die diese dann gemäß dem der Simulation zugrunde liegenden Regelwerk umsetzen und über die Ergebnisse den Teilnehmenden wiederum Bericht erstatten (Jones 1995). Das Regelwerk ist somit von den Teilnehmenden abgeschildert, die sich ganz auf ihre Aufgaben der Informationsprozessierung und des Entscheidens konzentrieren können. Durch den Wegfall der Notwendigkeit, das Regelwerk zu beherrschen, erweisen sich FBS als ausgesprochen zugänglich auch für diejenigen, die keinerlei Vorerfahrungen mit Konfliktsimulationen aufweisen können.

Für den Einsatz im Unterricht stellen FBS eine nahezu ideale Lösung dar, ermöglichen sie es doch mit einem Minimum an Aufwand – der leitende *facilitator* muss das Regelwerk beherrschen, im Idealfall wird er je nach Umfang der Simulation durch weitere *facilitator* unterstützt – eine Simulation durchzuführen, bei der sich die Teilneh-

menden ganz auf ihre Rolle in der Simulation konzentrieren können.

Diese stehen dabei vor der Herausforderung, sich nicht nur auf analytische Art und Weise mit den einzelnen Gegenständen der Simulation auseinanderzusetzen, sondern ihr Zusammenwirken innerhalb eines komplexen, multifaktoriellen Systems nachzuvollziehen. Das hierbei entstehende ganzheitliche Verständnis für beispielsweise historische Phänomene kann dann die Voraussetzung für eine adäquate multiperspektivische Modellierung komplexer Phänomene bilden; im Idealfall schließt sich daher an die Teilnahme an einer Konfliktsimulation in einem zweiten Schritt die Erstellung einer solchen an.

agent-based simulations (ABS) im Unterricht

Insbesondere in Veranstaltungen mit fortgeschrittenen Studierenden, die bereits grundsätzlich mit dem Instrument der Konfliktsimulation vertraut sind, bietet es sich an, die eigenständige Modellierung und kollaborative Implementierung von Konfliktsimulationen zu ausgesuchten Teilproblemen zu einer zentralen Aufgabe von Kleingruppen zu machen, um anschließend die – zunächst oft unbewusst getroffenen – Modellierungsentscheidungen der einzelnen Gruppen zu präsentieren und diskutieren.

Für die Implementierung formaler Modelle in imperative Programmiersprachen eignet sich dabei besonders das Konzept der objekt-orientierten Programmierung (OOP). Bei diesem Ansatz werden Forschungsdaten als Objekte mit Attributen repräsentiert, die über Objekten spezifisch zugeordnete Funktionen miteinander interagieren. Generell handelt es sich bei den für eine derartige Umsetzung notwendigen Fähigkeiten zur Abstraktion und Modularisierung um zentrale Kompetenzen für die imperative Programmierung (Jannidis 2017, S. 88).

Als Beispiel sei hier ein Seminar aus dem Sommersemester 2022 an der Julius-Maximilians-Universität Würzburg kurz vorgestellt: In diesem Seminar wurden Studierende des Masterstudienganges Digital Humanities an Thema der mathematischen Modellierung von Abnutzung in militärischen Konflikten herangeführt. Die Teilnehmenden wurden zunächst, nach einer kurzen Einführung in den historischen Kontext, vor die Aufgabe gestellt, in zwei kleinen Gruppen mit Hilfe von Würfeln und Markern Regeln zu entwickeln, mit deren Hilfe das beschriebene Phänomen simuliert werden kann. Gleichzeitig sollten die Studierenden experimentell untersuchen, welche Erkenntnisse über das Phänomen sich aus ihren Modell ableiten lassen. In einem zweiten Schritt wurde der Kurs mit dem wichtigsten deterministisch-mathematischen Modell konfrontiert, das bis heute für die Modellierung dieses Problems genutzt wird (Lanchester 1916). Zu diesem Zeitpunkt war den Teilnehmenden der Vergleich verschiedener Modellierungskonzepte für dasselbe Problem bereits möglich. In den folgenden Seminarstunden wurde das Konzept des *agent-based modeling* (vgl. Gavin 2014 und Romanovska et al. 2021) eingeführt und ein einfaches Python-Framework⁴ für eine *agent-based simulation* (ABS) von Abnutzung vorgestellt. Dieses

Framework dient einerseits zur Umsetzung vergleichender Experimente, aber auch als Grundlage für die Implementierung neuer Aspekte und der Setzung eigener Schwerpunkte durch die Studierenden.

Dieses Unterrichtskonzept erweist sich als ein geeigneter Rahmen für die generelle Diskussion über die Eignung verschiedener Modellierungsansätze im Hinblick auf unterschiedliche Forschungsprobleme, eine Einführung in die Erstellung und die Anwendung einer ABS und darüber hinaus ein technisch niederschwelliges Framework aus der Forschungspraxis für das objektorientierte Programmieren.

Ausblick

“Come, Watson, Come! The Game is afoot!”⁵

Bei Konfliktsimulationen – und insbesondere bei ABS – handelt es sich um leistungsfähige Werkzeuge, deren didaktischer Wert in drei Bereichen zu finden ist: zum ersten bieten sie den Teilnehmenden einen partizipatorischen Zugang zu dem jeweils behandelten Gegenstand und damit eine alternative Lernerfahrung. Zum zweiten erfahren die Teilnehmenden direkt die mit der Prozessierung komplexer Information sowie der darauf basierenden Entscheidung verbundenen Schwierigkeiten und gewinnen so Einsichten in ihre eigenen Entscheidungsprozesse. Die Auseinandersetzung mit dem Problem der Erstellung einer Simulation stellt schließlich einen niederschweligen Einstieg in das Problem der Modellierung von Forschungsgegenständen sowie der Operationalisierung von Forschungsfragen dar; im Rahmen der Erstellung einer Simulation können die Teilnehmenden anhand konkreter Beispiele den Aufbau möglicher Forschungsvorhaben diskutieren und kritisch reflektieren.

Über den konkreten Nutzen im Rahmen des historischen bzw. Gesellschaftswissenschaftlichen Unterrichts hinaus kann die Auseinandersetzung mit Konfliktsimulationen gerade aufgrund des niederschweligen Zugangs einen wichtigen Beitrag zur allgemeinen Auseinandersetzung mit Modellierungsfragen, mit dem Erwerb weiterer Kompetenzen (*Digital Literacy*) und damit einer erfolgreichen Partizipation in einer vernetzten, hoch technologisierten Gesellschaft leisten. Der Einsatz von Konfliktsimulationen, ihre Fähigkeiten und Begrenzungen verdienen daher auch in den Digital Humanities mehr Beachtung, als sie momentan erfahren.

Fußnoten

1. Beispielsweise im Rahmen der „Historical Social Research (HSR)“, einer international begutachteten, wissenschaftlichen Fachzeitschrift für die Anwendung formaler Methoden in der Geschichte. <https://www.ge-sis.org/hsr> oder in Jeremiah McCall's Artikel „Historical Simulations as Problem Spaces: Criticism and Classroom Use“ (McCall, 2012).
2. Linvill, Gay, dir. “The Neonatal Nomenclature.” *Big Bang Theory*, season 11, episode 16, CBS, 2018.
3. Das von Richard Berg entwickelte Spiel wurde 1978 von Simulations Publications Inc. veröffentlicht.

4. Der in Python implementierte *Attrition Simulator* (Pielström, Steffen 2022) ist frei zugänglich unter: <https://github.com/cosimg/attritionsim>.
5. Doyle, Arthur Conan. 1904. *The Adventure of the Abbey Grange*.

Bibliographie

- Beynon, Meurig, Steve Russ und Willard McCarty. 2006. „Human Computing—Modelling with Meaning.“ *Literary and Linguistic Computing* 21 (2): 141–57. doi:10.1093/llc/fql015.
- Ciula, Arianna, Oyvind Fide, Cristina Marras and Patrick Sahle. 2018. „Models and Modelling between Digital and Humanities: Remarks from a Multidisciplinary Perspective.“ *Hist. Soc. Res.* 43 (4): 343–362.
- Davis, Paul K, und Paul Bracken. 2022. „Artificial Intelligence for Wargaming and Modeling.“ *The Journal of Defense Modeling and Simulation*. doi:10.1177/15485129211073126.
- Flanders, Julia und Fotis Jannidis. 2015. „Data Modeling.“ In *A New Companion to Digital Humanities*, 229–37. John Wiley & Sons, Ltd. doi:10.1002/9781118680605.ch16
- Gavin, Michael. 2014. „Agent-Based Modeling and Historical Simulation.“ *Digital Humanities Quarterly* 8 (4). Zugriffen 29. Juli 2022. <http://www.digitalhumanities.org/dhq/vol/8/4/000195/000195.html>.
- Jannidis, Fotis. 2017. „Grundbegriffe des Programmierens.“ In *Digital Humanities : Eine Einführung*, hg. v. Fotis Jannidis, Hubertus Kohle und Malte Rehbein, 68–95. Stuttgart: J.B. Metzler. doi:10.1007/978-3-476-05446-3_6.
- Jones, Ken. 1995. *Simulations: A Handbook for Teachers and Trainers*. London³: Kogan Page Ltd.
- Lanchester, Frederick W. 1916. *Aircraft in Warfare: the Dawn of the Fourth Arm*. London: Constable and Company Limited.
- McCall, Jeremia. 2012. „Historical Simulations as Problem Spaces: Criticism and Classroom Use“ In *Journal of Digital Humanities*. Zugriffen 26. Juli 2022. <http://journalofdigitalhumanities.org/1-2/historical-simulations-as-problem-spaces-by-jeremiah-mccall/>.
- Michael, David und Chen Sande. 2006. *Serious Games: Games that Educate, Train and Inform*. Mason: Course Technology.
- Pielström, Steffen. 2022. „Attrition Simulator“. Conflict Simulation Group, GitHub repository. <https://github.com/cosimg/attritionsim>.
- Piotrowski, Michael. 2019. „Accepting and Modeling Uncertainty“. In *Zeitschrift Für Digitale Geisteswissenschaften*. doi:10.17175/sb004_006a.
- Romanowska, Isa, Colin D. Wren and Stefani A. Crabtree. 2021. „Agent-Based Modeling for Archaeology“. SFI Press, Santa Fe. doi:10.37911/9781947864382
- Sabin, Philip. 2011. „The benefits and limits of computerization in conflict simulation“. *Literary and Linguistic Computing* 26 (3): 323–28. doi:10.1093/llc/fqr024.
- Sabin, Philip. 2012. *Simulating War: Studying Conflict through Simulation Games*. London: Continuum, 2012.
- Sabin, Philip. 2016. „Wargames as an academic instrument“. In *Zones of Control. Perspectives on Wargaming* hg. von Pat Harrigan und Matthew Kirschenbaum, 421–438. Cambridge, MA: MIT Press.

Thaller, Manfred. 2017. „Digital Humanities als Wissenschaft“. In *Digital Humanities : Eine Einführung*, hrsg. v. Fotis Jannidis, Hubertus Kohle und Malte Rehbein, 13–18. Stuttgart: J.B. Metzler. doi:10.1007/978-3-476-05446-3_2.

Wintjes, Jorit und Steffen Pielström. 2019. „Pluie de Balles; Complex Wargames in the Classroom“. In *Analog Game Studies V (III)*. Zugriffen 29. Juli 2022. <https://analoggamestudies.org/2018/09/pluie-de-balles-complex-wargames-in-the-classroom/>.

Wintjes, Jorit und Steffen Pielström. 2019. „»Preußisches Kriegsspiel«: Ein Projekt an der Julius-Maximilians-Universität Würzburg“ *Militärgeschichtliche Zeitschrift* 78 (1): 86–98. doi:10.1515/mgzs-2019-0004.

Wintjes, Jorit. 2022. „A School for War – A Brief History of the Prussian Kriegsspiel.“ In *Simulation and Wargaming*, hg. von C. Turnitsa/C. Blais/A. Tolk (eds.), 25–64. Hoboken: Wiley. <https://doi.org/10.1002/9781119604815.ch2>

Die offene Edition. Vernetzung, Datenpublikation und Transparenz in der edition humboldt digital

Dumont, Stefan

dumont@bbaw.de
Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Kraft, Tobias

kraft@bbaw.de
Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Seifert, Sabine

sabine.seifert@uni-potsdam.de
Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Thomas, Christian

thomas@bbaw.de
Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Wierzoch, Jan

jan.wierzoch@bbaw.de
Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Das an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) angesiedelte Akademienvorhaben *Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung* erschließt und ediert die amerikanischen, russisch-sibirischen und europäischen Reise-tagebücher des preußischen Naturforschers und Entdeckers Alexander von Humboldt.¹ Begleitet werden die Tagebücher von thematisch zugehörigen Briefen aus seinem weltumspannenden Korrespondentennetz sowie von Manuskripten aus seinem umfangreichen Nachlass, von denen viele bis dato nie veröffentlicht wurden. Ergänzt werden diese edierten Texte durch Forschungsbeiträge, eine Chronologie zu Humboldts Leben und umfangreiche Register. Die Publikationsstrategie ist "digital first", d. h. die veröffentlichten Dokumente erscheinen zuerst online ohne Verzögerung durch 'Moving Walls' oder sonstige verlagsseitige Einschränkungen, und die gesamte Ausgabe ist unter einer Creative-Commons-Lizenz² vollständig frei zugänglich. Mit der Veröffentlichung der ersten Bände der Print-Edition im Verlag Springer Nature/J. B. Metzler³ wurde die Hybridstrategie des Projekts umgesetzt. Realisiert wird die digitale Edition mit ediarum⁴, das u. a. auf X-Technologien, der freien XML-Datenbank eXistdb und der Software Oxygen XML Author aufsetzt.

Fünf Jahre nach dem Launch der *edition humboldt digital* (*ehd*) haben wir acht Versionen dieser digitalen, textkritisch-dokumentarischen Edition⁵ vorgelegt. Mit der aktuellen Version 8 der *ehd* (veröffentlicht im Mai 2022)⁶ lösen wir nun das Versprechen ein, Humboldts komplexe handschriftliche und schwer zu entziffernde Texte auch *als Daten* bereitzustellen, indem wir (1) die kommentierten Texttranskriptionen von mehr als 500 Dokumenten (ca. 2.800 Seiten), (2) die umfassende Alexander von Humboldt-Chronologie mit ca. 1.600 datierten Ereignissen aus Humboldts fast 90-jährigem Leben und (3) ca. 18.000 Indexeinträge (z. B. Personen, Orte, Institutionen, bibliographische Einträge) auf GitHub (Ette et al. 2022) und (ab Winter 2022/23) auf Zenodo zur Verfügung stellen. Alle Datensätze liegen im TEI-XML-Format vor. Dem Single-Source-Prinzip folgend, basieren sowohl die digitale als auch die gedruckte Komponente (Buch, PDF und eBook-Derivate) vollständig auf denselben TEI-XML-kodierten Daten. Das TEI-XML-Subset der *ehd* wurde durch Übernahme etablierter TEI-Spezifikationen, v. a. des Basisformats für Manuskripte des Deutschen Textarchivs (DTABf-M⁷; Thomas/Haaf 2016-2019), entwickelt, um ein Höchstmaß an Standardisierung, Nutzbarkeit und Interoperabilität der Daten zu gewährleisten (Dumont/Haaf/Kraft/Czmiel/Thomas/Boenig 2016). Eine umfassende Dokumentation der Transkriptions- und Kodierungsrichtlinien steht zur Verfügung⁸ und kann von anderen, ähnlich gelagerten Projekten nachgenutzt werden⁹. Gleichzeitig bringen sich die Projektmitarbeiter:innen aktiv in die Verbesserung von bestehenden Richtlinien ein¹⁰. Dadurch fließen auf zweierlei Wegen Erfahrungen aus der editorischen Praxis in die Community zurück.

Seit Abschluss der Betaphase im Mai 2017 wird die *edition humboldt digital* versioniert publiziert, d. h. die Daten werden nicht einfach aktualisiert und überschrieben, sondern durch jedes Update (mittlerweile einmal im Jahr)

wird eine neue, zusätzliche Datenschicht hinzugefügt. Gleichzeitig werden alle vorangehenden Versionen weiterhin bereitgehalten und lassen sich über die Web-Oberfläche aufrufen – bis hin zu den Registereinträgen mit ihren dynamischen Verlinkungen. Die Datensätze werden in Zukunft immer als neue, zusätzliche Version publiziert. Ergänzt wird diese Versionierung seit 2022 durch die Einführung von "Editionsstufen", die die unterschiedlichen Bearbeitungszustände systematisch abbilden.¹¹ Zusammen mit der umfangreichen Dokumentation wird damit der gesamte Forschungs- und Editionsprozess offen gelegt und analysierbar. Einen ersten summarischen Überblick über den Fortschritt der *ehd* gibt die mit Version 8 neu hinzugekommene "Versionsgeschichte", die die acht publizierten Versionen auch quantitativ auswertet.¹²

Das nun zur Verfügung gestellte Datenset wird nicht direkt aus der eXistdb-Datenbank exportiert, sondern über die öffentlich zugängliche API abgerufen.¹³ Das ermöglicht diverse Optimierungen am Datenbestand für die externe Nachnutzung. So werden z. B. alle projektinternen IDs durch URLs aus Normdateien ersetzt – sofern eine solche in den einschlägigen Normdaten-Beständen vorhanden ist. Ist dies nicht der Fall, werden die projektinternen IDs als vollständige URLs ausgegeben und so immerhin technische Interoperabilität gewährleistet.

Die Registereinträge werden, wo immer verfügbar, mit URLs aus Normdateien versehen. Insbesondere das Personenregister weist dabei großes Potenzial für die unmittelbare Nachnutzung auf, da zahlreiche historische Personen in Humboldts Texten noch nicht in der wichtigsten Normdatei für die deutschsprachige Forschungsgemeinschaft, der GND¹⁴ der Deutschen Nationalbibliothek, dokumentiert sind.¹⁵ Diese ergänzen den Daten können dazu beitragen, die Normdateien der Community, wie GND, Wikidata etc. zu verbessern. Dabei stellen sich unterschiedlich große Hürden für die Zuarbeit: während Wikidata von Prinzip her ein offenes Communityprojekt ist, ist die GND institutionell angesiedelt und wird redaktionell betreut. Die Zuarbeit zur GND wurde im Projekt GND4C grundsätzlich für nicht-bibliothekarische Bereiche geöffnet, in naher Zukunft sollen die Ergänzungsmöglichkeiten seitens des Editionsprojekts ausgelotet werden. Leider stehen solche Zuarbeiten immer unter dem Vorbehalt der Projektkapazitäten, da sie bei der *ehd* – ebenso wie in den meisten anderen Projekten – eigentlich nicht vorgesehen sind.

Die API selbst bietet nicht 'nur' den Volltext an, sondern ebenfalls diverse Metadaten, um eine umfassende Vernetzung der Edition zu gewährleisten. So bietet eine Schnittstelle die Metadaten zu den edierten Texten und Forschungsbeiträgen unter anderem im Dublin Core-Format via OAI-PMH an. Dadurch werden alle diese Texte in der Open Access-Suchmaschine BASE nachgewiesen. Daneben werden eine BEACON-Schnittstelle und eine CMIF-Schnittstelle für *correspSearch* (Dumont, Grabsch und Müller-Laackman 2021) angeboten, die mittlerweile zu den 'klassischen' Ausstattungen einer digitalen Edition zählen dürfen. Eine Schnittstelle ins Semantic Web gibt es derzeit noch nicht.¹⁶ Grund dafür ist v.a., dass andere Schnittstellen und vor allem Funktionen der *ehd* bisher im Fokus der Entwicklungsarbeit standen. Das Vorhaben läuft bis 2032 und lässt es daher grundsätzlich zu, in Zukunft auch diesen Bereich anzugehen.

Denkbar wäre es, z.B. die Einträge der Chronologie oder der Register noch stärker zu schematisieren und tiefer zu kodieren, um dieses Wissen als Linked Open Data bereitzustellen. Das würde aber nicht nur Entwicklungsaufwand bedeuten, sondern auch erhebliche redaktionelle Arbeit, für die entsprechende Ressourcen geschaffen werden müssten. Die edierten Texte an sich werden für solch eine zusätzliche Aufbereitung - im Sinne einer "assertive edition" (Vogeler 2019) - wohl leider nicht in Frage kommen, dafür wäre der Aufwand (gegenüber dem Projektplan) zu hoch.

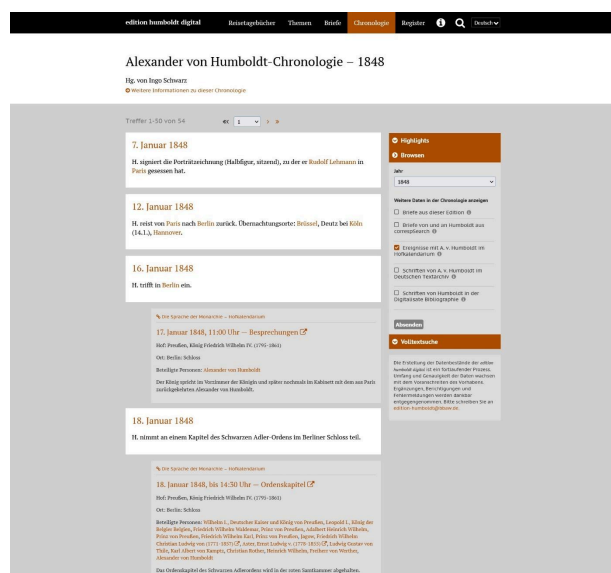


Abb. 1: Screenshot der Alexander von Humboldt-Chronologie in der *edition humboldt digital* mit eingebundenen Daten aus dem Hofkalendarium 1848.

Die *edition humboldt digital* stellt nicht nur ihre eigenen Daten zur Verfügung, sondern nutzt auch andere Daten nach und verwendet externe Webservices zur Anreicherung oder zur Vergrößerung des eigenen Funktionsumfangs. So werden die TEI-XML-Dateien zur Lemmatisierung an den Webservice DTA::CAB¹⁷ (Jurish 2011) geschickt, dort analysiert und angereichert und wieder zurück in die Datenbank gespeichert. Dadurch wird eine lemmabasierte Suche ermöglicht, die die unterschiedlichen Schreibweisen und Flexionsformen bei der Suche abfangen kann.¹⁸

Abgesehen von der Suche ist vor allem die Chronologie zu Alexander von Humboldts Leben ein zentraler Vernetzungs- und Integrationspunkt externer Ressourcen. Diese Chronologie wurde bereits in den 1960er Jahren an der damaligen Deutschen Akademie der Wissenschaften zu Berlin begonnen, in den 2000ern überarbeitet als HTML-Version im Web veröffentlicht und 2015/16 schließlich in TEI-XML überführt und in die *edition humboldt digital* integriert. Dort wird sie fortlaufend gepflegt, erweitert, mit externen Quellen sowie mit den edierten Texten und Registereinträgen verlinkt. Darüber hinaus integriert sie automatisiert verschiedene externe Angebote und Dienste in die Edition, wie z. B. die Metadaten der publizierten Korrespondenz Alexander von Humboldts aus correspSearch¹⁹, Schriften Humboldts aus

dem Deutschen Textarchiv²⁰ und Einträge aus dem Hofkalendarium der preußischen Monarchie²¹ mit Bezug zu Humboldts Leben (siehe Abb. 1). Dadurch öffnet die Chronologie die *edition humboldt digital* nach außen hin zu zahlreichen extern vorliegenden Materialien und Informationen.

Auch in den Registern werden externe Daten nachgenutzt. Zum einen werden die BEACON-Schnittstellen anderer ausgewählter digitaler Publikationen, wie die Kosmos-Vorlesungen Alexander von Humboldts im Deutschen Textarchiv, abgerufen und automatisiert verknüpft.²² Darüber hinaus werden ganze Datensätze der GND (also nicht nur die bloßen URIs) nachgenutzt. Mit ihrer Hilfe werden die Registereinträge der *ehd* automatisiert *untereinander* verlinkt – nämlich anhand der in der GND notierten familiären und freundschaftlichen Beziehungen. Außerdem können so Porträts von Wikimedia eingebunden werden. Ein Register, das in besonders großem Maße auf externe Dienste zurückgreift, ist das Pflanzenregister. Hier liegt in der Edition die Besonderheit vor, dass die Pflanzen nicht mit eigenen Registereinträgen versehen werden, sondern dieses Register ausschließlich automatisch über die taxonomischen Namen generiert wird. Dazu werden alle Pflanzennamen in den Transkriptionen von den Editor:innen auf ihren regulären Namen ergänzt oder korrigiert (natürlich nachverfolgbar). Anhand der taxonomischen Namen werden dann verschiedene Webservices abgefragt und automatisiert verknüpft. Damit öffnet die Edition v. a. die Tagebücher Humboldts für verschiedene Disziplinen wie die Biodiversitätsforschung.

Mit der intensiven Nachnutzung externer Webservices und Daten erhöht sich der Nutzen einer digitalen Edition signifikant. Gleichzeitig stellt diese Nachnutzung neue Probleme und Herausforderungen an die Entwicklung und den Betrieb digitaler Editionen, da externe Dienste sich grundsätzlich ändern können (aktualisierte Schnittstellen, Änderungen im Format etc.). Das - und Fragen der Performance - führt dazu, dass diese externen Daten auch in der *edition humboldt digital* vorgehalten werden müssen. Fraglich ist dann aber weiter, ob und in welchem Rahmen diese externen Daten auch in der Datenpublikation mitveröffentlicht werden können und müssen. Darüber hinaus kann man nicht garantieren, dass ein externer Dienst auch in Zukunft verfügbar sein wird. Das ist insbesondere ein Problem vor dem Hintergrund der relativ langen Laufzeit des Projekts: Werden die externen Services zur Anreicherung von Texten und Daten, die erst in den nächsten Jahren hinzukommen, noch vorhanden sein? Der Vortrag möchte am Beispiel der *edition humboldt digital* diese und weitere Herausforderungen und Chancen einer 'offenen Edition' vorstellen und die damit zusammenhängenden, skizzierten Themenfelder Datenpublikation, Bereitstellung und Nachnutzung von APIs und externen Daten sowie Transparenz im Editions- und Forschungsprozess diskutieren.

Fußnoten

1. Projektbeschreibung auf den Seiten der BBAW: <http://www.bbaw.de/forschung/avh-r/uebersicht> (Zugriff für alle im Abstract angegebenen Links: 03. August 2022);

edition humboldt digital : <https://edition-humboldt.de/> . Vgl. dazu Kraft/Dumont 2020; zu dem auch im vorliegenden Abstract zentralen Aspekt der Vernetzung siehe die Zusammenfassung und Illustration dieses Ansatzes in Kraft/Dumont 2017.

2. CC-BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>) für die TEI-XML-Daten; CC-0 für die Register-Einträge und Metadaten zu den Dokumenten.

3. Buchreihe *edition humboldt print* , <https://www.springer.com/series/16345> . 2020 erschien der erste Band über die *Geographie der Pflanzen* , herausgegeben von Ulrich Päßler. 2022 wurde Band 1 der Amerikanischen Reisetagebücher von Carmen Götz 2022 herausgegeben, gefolgt von Band 1 der Russisch-Sibirischen Reisetagebücher, herausgegeben von Tobias Kraft und Florian Schnee, in 2023.

4. <https://www.ediarum.org/> ; zur Erfassungssoftware siehe Dumont et. al 2021.

5. Siehe für eine Orientierung zum Editionsmodell der *ehd* Sahle 2016 sowie insbesondere zum Konzept der 'documentary edition' beispielsweise Pierazzo 2011.

6. Vgl. den Überblick zur Version 8 sowie den vorhergehenden Versionen der *ehd* unter <https://edition-humboldt.de/H0020382> ; API: <https://edition-humboldt.de/about/index.xml?id=api> ; TEI-XML (der jeweils aktuellen Version der *ehd*) <https://edition-humboldt.de/api/v1.1/tei-xml.xml> .

7. <https://www.deutschestextarchiv.de/doku/basisformat/> .

8. Editionsrichtlinien der *ehd* , v. 1.1.2 (9.5.2022), <https://edition-humboldt.de/richtlinien/index.html> .

9. So orientiert sich beispielsweise das Akademienvorhaben *Propyläen: Goethes Biographica* (<https://goethe-biographica.de/>) im Zuge seiner Entwicklung eines TEI-XML-Datenmodells für diese Hybrid-Edition v. a. für die Briefe von und an Goethe sowie dessen Tagebücher an den Richtlinien der *ehd* .

10. Z. B. das TEIC/TEI Issue #2028 "@calendar should allow multiple values", <https://github.com/TEIC/TEI/issues/2028> , das aufgrund der Diskussion auf der TEI-Mailingliste (August 2020) entstand und zur Implementierung in die TEI P5 Guidelines v. 4.3.0 (2021-08-31) führte.

11. Mit Version 8 wurde dieses Feature erstmals umgesetzt, zunächst nur bei den Tagebüchern. Zu den Editionsstufen siehe <https://edition-humboldt.de/richtlinien/ediarum.AVHR/editionsstufen.html> .

12. <https://edition-humboldt.de/H0020382> .

13. Von Axelle Lecroq wurde dafür die bereits seit Version 1 vorhandene API optimiert und ein entsprechendes Skript zum Abruf der Daten entwickelt.

14. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html .

15. Von derzeit 9022 vorliegenden Personenregistereinträgen sind 5581 mit einer GND-URI versehen, das entspricht rund 61% aller Personen. 663 Datensätze verfügen nur über eine VIAF-URI. Damit verfügen rund 40% über gar keine Norm-ID – i. d. R., weil kein Normdatensatz vorhanden ist.

16. Ein sehr kleiner Anfang konnte dennoch schon gemacht werden: In Wikidata wurden seitens der Freiwilligen dort bei den entsprechenden Personen die PermaIDs der *ehd* eingetragen, siehe z.B. <https://www.wikidata.org/wiki/Q132197> . Das war möglich, weil die *ehd* kon-

sequent GND-IDs zu den Personen einträgt, falls vorhanden.

17. <https://www.deutschestextarchiv.de/cab/> .

18. Die Suchfunktionalität auf der *ehd* -Seite wurde im Zuge der Version 8 (2022) grundlegend überarbeitet; siehe zur Einführung "ehd – explained. Kapitel 4: Die Suche" von Tobias Kraft, aufgenommen beim Humboldt-Tag am 16. September 2022 in der BBAW, verfügbar unter <https://youtu.be/l1D0zGd7osA> .

19. <https://correspsearch.net/de/suche.html?s=http://d-nb.info/gnd/118554700> .

20. <https://edition-humboldt.de/chronologie/index.xml?jahr=1827&dta=on> .

21. Vgl. Einleitung Hofkalendarium, <https://actaborussica.bbaw.de/v5/P0006298> , in Akademienvorhaben 2021.

22. Siehe z. B. den Registereintrag zu August Böckh: <https://edition-humboldt.de/H0003413> .

Bibliographie

Akademienvorhaben Anpassungsstrategien der späten mitteleuropäischen Monarchie am preußischen Beispiel (1786-1918) (Hg.). 2021. *Die Sprache der Monarchie (Version 5)*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. URL: <https://actaborussica.bbaw.de/>

Dumont, Stefan und Susanne Haaf, Tobias Kraft, Alexander Czmiel, Christian Thomas, Matthias Boenig. 2016. "Applying Standard Formats and Tools: 'Alexander von Humboldt auf Reisen' as an Example for the Collective Subsequent Use of DTABf and ediarum". Vortrag, *TEI Conference and Members' Meeting*, Vienna. Abstract (PDF): https://www.tei-c.org/Vault/MembersMeetings/2016/sites/default/files/TEIconf2016_BookOfAbstracts.pdf , 69-70 (zugegriffen: 03. August 2022).

Dumont, Stefan und Nadine Arndt, Sascha Grabsch, Lou Klappenbach. 2021. *ediarum.BASE.edition (Version 2.0.0)* [Computer software]. <https://doi.org/10.5281/zenodo.5897100> (zugegriffen: 03. August 2022).

Stefan Dumont, Sascha Grabsch und Jonas Müller-Laackman (Hg.). 2021. *correspSearch – Briefeditionen vernetzen (2.0.0)* [Webservice]. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. <https://correspsearch.net> (zugegriffen: 03. August 2022).

Ette, Ottmar (Hg.). 2022. *edition humboldt digital (Version 8)*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. <https://edition-humboldt.de/> (zugegriffen: 03. August 2022).

Ette, Ottmar und Stefan Dumont, Annika Geiser, Carmen Götz, Tobias Kraft, Ulrike Leitner, Ulrich Päßler, Florian Schnee, Christian Thomas (Hg.). 2022. *TEI-XML-Datenset der Tagebücher, Briefe, Dokumente, Forschungsbeiträge, Chronologieeinträge und Register der edition humboldt digital (Version 8)*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. URL: <https://github.com/telota/edition-humboldt-digital>

Jurish, Bryan. 2011. *Finite-State Canonicalization Techniques for Historical German*. Potsdam: Universität Potsdam. urn:nbn:de:kobv:517-opus-55789 (zugegriffen: 03. August 2022).

Kraft, Tobias und Stefan Dumont. 2017. *Edition humboldt digital vernetzt*, Poster. Zenodo. <http://doi.org/10.5281/zenodo.1035134> (zugegriffen: 03. August 2022).

Kraft, Tobias und Stefan Dumont. 2020. "The Humboldt Code. On creating a hybrid digital scholarly edition of a 19th century globetrotter." In *Wiener Digitale Revue* 1. <https://doi.org/10.25365/wdr-01-03-02> (zugegriffen: 03. August 2022).

Pierazzo, Elena. 2011. "A Rationale of Digital Documentary editions". In *Literary and Linguistic Computing*, 26,4, 463-477. <https://doi.org/10.1093/lc/fqr033> (zugegriffen: 03. August 2022).

Sahle, Patrick. 2016. "What is a Scholarly Digital Edition?" In *Digital Scholarly Editing: Theories and Practices*, hg. von Matthew James Driscoll und Elena Pierazzo, 19-39. Cambridge, UK: Open Book Publishers. <https://books.openedition.org/obp/3397> (zugegriffen: 03. August 2022).

Thomas, Christian und Susanne Haaf. 2016-2019. "Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format". In *Journal of the Text Encoding Initiative* [Online], Issue 10 | December 2016 - July 2019. <https://doi.org/10.4000/jtei.1650> (zugegriffen: 03. August 2022).

Vogeler, Georg. 2019. „The ‘Assertive Edition’. On the Consequences of Digital Methods in Scholarly Editing for Historians". *International Journal of Digital Humanities* 1 (2): 309-22. <https://doi.org/10.1007/s42803-019-00025-5>.

Disko: Zur Einbindung von Citizen Humanities beim Aufbau eines Diversitäts-Korpus

Schumacher, Mareike

schumacher@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Marie, Flüh

marie.flueh@uni-hamburg.de
Universität Hamburg

Peter, Leinen

P.Leinen@dnb.de
Deutsche Nationalbibliothek

Abstract

Für die Korpuskonstituierung im Projekt Disko (Diversitäts-Korpus) haben wir ein Konzept entwickelt, das sowohl auf Ansätzen aus den Digital als auch den

Public Humanities aufbaut und diese zu einer Citizen-Humanities-Komponente zusammenfügt (vgl. Heinisch 2020). Dieses Konzept dient dazu, ein Erzähltextkorpus aufzubauen, in dem Gender nicht nur binär dargestellt wird. Disko wird zur Grundlage eines Gender-Classifiers 2.0 (aufbauend auf dem Gender-Classifer 1.0 von Schumacher und Flüh 2021), den wir mithilfe von Verfahren des überwachten maschinellen Lernens so trainieren, dass diverse Gender-Zuschreibungen automatisch klassifiziert werden. Den Grundsätzen von Offenheit, Transparenz und Empowerment folgend, die für Citizen-Humanities-Projekte zentral sind (vgl. Heinisch 2020, Dunn und Hedges 2012: 19), werden Vertreter:innen unterschiedlicher Communities angesprochen und in die Korpuskonstituierung einbezogen. Dabei steht das Community-Building zwischen Offenheit und Zielgruppenspezifität.

Disko (Diversitäts-Korpus)

Disko ist ein Kooperationsprojekt zwischen der Deutschen Nationalbibliothek (DNB), der Technischen Universität Darmstadt und der Universität Hamburg. Die DNB sammelt im gesetzlichen Auftrag seit 1913 u.a. alle in Deutschland veröffentlichten Medienwerke; seit 2006 schließt dies auch die sogenannten Netzpublikationen, also genuin digitale Werke, mit ein. Die Digitalisierungsstrategie zielt auf eine systematische und auch projekt- und anlassbezogene Digitalisierung der physischen Bestände sowie die Vernetzung zur Wissenschaft ab und schreibt seit 2020 mit dem jährlichen DH-Call ein Unterstützungsangebot für Forschende aus, die mit den Daten der DNB arbeiten möchten. Grundlage dieser Aktivität bildet die Reform des Urheberrechts-Wissensgesellschafts-Gesetzes (UrhWissG) im Jahr 2018. Die konkrete Förderung durch die DNB besteht in der Bereitstellung von Metadaten, digitalen bzw. digitalisierten Objekten und einer passenden Infrastruktur zur Bearbeitung und Analyse der teilweise sensiblen Daten. Das Projekt Disko ist seit Frühjahr 2022 Teil dieser Förderlinie und kann darum nicht nur auf genuin digitale Medien zugreifen, sondern auch Texte retrodigitalisieren lassen, die bisher von keinem Digitalisierungsprojekt erfasst wurden. Darüber hinaus bietet die Zusammenarbeit mit der DNB die Möglichkeit, bei der Korpusbildung eine quantitative Einschätzung der Grundgesamtheit (Calvo Tello 2021: 96-97, Schöch 2017: 225) aller in einer definierten Zeitspanne in Deutschland erschienenen Romane zu berücksichtigen, da jedes mit einer deutschen ISBN erscheinende Buch hier gemeldet und in doppelter Ausführung abgegeben oder als digitales Objekt eingereicht werden muss.

Im Projekt m*w werden seit 2019 Genderrollen und -stereotype erforscht. Der Gender-Classifer 1.0 erkennt und klassifiziert weibliche, männliche und neutrale Genderrollen durchschnittlich zu 78% (F1-Score) in Romanen, Dramen und Dramen (vgl. Flüh/Lemke/Schumacher 2022). Bei der Annotation des Trainingskorpus und Fallstudien mit diesem Classifier (vgl. Schumacher und Flüh 2020; Flüh und Schumacher 2021/2022; Flüh/Horstmann/Schumacher 2022) zeigt sich, dass Brüche mit stereotypen Genderzuschreibungen in älteren Texten selten, in zeitgenössischen aber häufiger vorkommen.

Aus dem Desiderat, einen Classifier zu trainieren, der diverse, nicht nur binäre Genderrollen erkennen und klassifizieren kann, resultierte das Projekt DisKo: die Erschließung eines Trainingskorpus aus den Beständen der DNB aus zeitgenössischen Romanen mit nicht-binären Genderdarstellungen.

Erfasst wird ein Zeitrahmen der letzten rund 70 Jahre; in diesem Fall beinhaltet die Grundgesamtheit also alle in Deutschland zwischen 1950 und 2022 erschienenen belletristischen Werke. Übertragen auf den Gesamtbestand der DNB bedeutet das, dass prinzipiell ca. 450.000 physische Objekte und ca. 435.000 digitale Objekte mit dem Erschließungsmerkmal "Belletristik" in Frage kommen. Die Überschneidungsmenge der Bestände ist leider nicht bekannt, kann jedoch über Algorithmen des Werkclustering durch die DNB eingegrenzt werden. Angesichts des Umfangs des als Basis für das maschinelle Lernen potentiell geeigneten Datensatzes sind wir mit grundlegenden Herausforderungen der Korpuskonstituierung konfrontiert, wie sie auch Gius et al. (2019) beschreiben: Die Menge (digital) vorliegender Texte ist so groß, dass der naheliegendste Weg der Korpusbildung darin bestünde, sich dabei auf eine Auswertung der Metadaten zu beschränken. Weil es sich bei "Figurengender" um einen textimmanenten Aspekt handelt, der in den Metadaten nicht erfasst wird, ist diese Vorgehensweise hier nicht möglich. Für den Aufbau des Korpus kommen standardisierte Methoden wie *Random Sampling* und *Stratified Sampling* (Calvo Tello 2021: 107; Schöch 2017: 226) ebenfalls nur bedingt in Frage. *Random Sampling* ist ungeeignet, weil wir Texte benötigen, in denen zuverlässig Figuren diverser Gender-Kategorien vorkommen. Beim Aufbau einer balancierten Sammlung (*Stratified Sampling*), in der "für alle Kombinationen wesentlicher Merkmale eine Mindestanzahl von Datensätzen" (Schöch 2017: 226) vorkommen müsste, ist problematisch, dass es keine endliche Liste wesentlicher Merkmale der Genderthematik gibt. Ob eine Figur im Hinblick auf Gender stereotyp oder ungewöhnlich ist und welche Kategorien es jenseits der traditionellen Einteilung in "männlich" und "weiblich" gibt, ist nicht klar definiert. Darum nutzen wir eine dritte Methode: die Verkleinerung der Population durch ergänzende Kriterien (Calvo Tello 2021: 108–109), mit Zügen einer opportunistischen Auswahl (Schöch 2017: 226) unter Einbezug von geisteswissenschaftlichem Crowd-Sourcing (vgl. Dunn und Hedges 2012). Vor dem Hintergrund, dass große Teile der Grundgesamtheit aktuell ausschließlich physisch vorliegen und somit für digitale Analysemethoden vorerst nicht in Frage kommen, scheint dies der einzig mögliche Weg zu sein. Die drei Kriterien, die wir bei der Korpuszusammenstellung berücksichtigen, sind:

1. Kriterium der Ausbalanciertheit
 - aus jedem Jahr wird zunächst nur ein Roman übernommen
 - von jedem Autor/jeder Autorin wird nur ein Roman übernommen
2. Kriterium der Heterogenität in Bezug auf
 - literarische Genres
 - Autor*innengender
3. Kriterium der thematischen Relevanz: Nur Romane werden übernommen, in denen Figuren vorkommen, die

- mit stereotypen Genderrollen brechen
- sich nicht klar in ein binäres Gendersystem fügen

Während Parameter der Kriterien I. und II. aus den in der DNB erfassten Metadaten abgeleitet werden können, handelt es sich bei III. um ein Kriterium, das nicht erfasst wird. Erschwerend kommt hinzu, dass die Parameter der Kategorie III. nicht klar definiert sind, sondern von Interpretationen und (unbewussten) Vorannahmen abhängen. Um sowohl im Hinblick auf die Interpretation von Figurengender als auch auf versteckte Vorannahmen eine möglichst große Heterogenität zu erreichen und auf diese Weise den Aspekt des Representation-Bias (Suresh und Gutttag 2019: 4) mit einzubeziehen, haben wir für die Korpuskonstituierung eine dreiteilige Citizen-Humanities-Komponente konzipiert.

Citizen Humanities in DisKo

Der Aufbau eines Diversitäts-Korpus bringt drei Herausforderungen mit sich. Erstens können die Texte in der Grundgesamtheit nicht algorithmisch erschlossen werden, da große Datenmengen nur physisch vorliegen. Zweitens ist die Auswahl von Texten, in denen Gender nicht (nur) binär dargestellt wird, bereits eine interpretatorische Leistung. Darüber hinaus muss drittens der sog. Representation-Bias mit einbezogen werden, der personengebunden funktioniert (vgl. D'Ignazio und Klein 2020: 53; Suresh und Gutttag 2019). Im Falle von DisKo ist einerseits die Frage zentral, was genau unter nicht-binären Genderdarstellungen zu verstehen ist. Reicht es aus, wenn eine literarische Figur einer (binären) Genderkategorie mit einer Reihe von Eigenschaften beschrieben wird, die traditionell eher der anderen (binären) Genderkategorie zugeschrieben werden? Oder muss eine Figur explizit mit Begriffen wie "queer" oder "gender-fluid" charakterisiert werden? Wie explizit oder implizit müssen nicht-binäre Genderdarstellungen angelegt sein, damit sie als solche erkannt werden? Zu Beginn der Korpusgestaltung steht also eine interpretatorische Leistung, deren Ziel die Auslegung des Verständnisses von nicht-binären Genderkategorisierungen ist. Hinzu kommt, dass bei dieser Thematik Aspekte von Macht und Suppression eine Rolle spielen. D'Ignazio und Klein weisen in *Data Feminism* darauf hin, dass es wichtig ist, Daten zu sammeln, die marginalisierte Gruppen sichtbar machen (D'Ignazio und Klein 2020: 119). Darüber hinaus sollte die betreffende Community beim Sammeln der Daten einbezogen werden, um einen Empowerment-Effekt zu erreichen (vgl. D'Ignazio und Klein 2020: 120; Heinisch 2020: 164). Außerdem ist für den Aufbau des Diversitäts-Korpus *Embodied Knowledge* (Christie et al. 2020) von Bedeutung; gerade bei Projekten, die Aspekte des Feminismus und der queer Community umfassen, ist die körperliche soziale Erfahrung Grundbestandteil einer Verstehensleistung, die zu einem kulturellen Wissen beiträgt (vgl. Christie et al. 2020). Im Sinne eines geisteswissenschaftlichen Crowd-Sourcing (Dunn und Hedges 2012) betrachten wir es darüber hinaus als Gelingensbedingung des Projektes DisKo, möglichst viele Personen an der Korpuszusammenstellung zu beteiligen, die im Hinblick auf ihre Genderzugehörigkeit und ihren beruflichen Hintergrund divers sind. Über einen eigens ent-

wickelten Fragebogen werden laufend Vorschläge für DisKo eingereicht. Fakultativ können gleichzeitig Daten zu Genderzugehörigkeit und beruflichem bzw. intersektuellem Hintergrund angegeben werden. Von Beginn an werden auf der Webseite des m*w-Projektes auf einer eigenen Seite die Liste der für DisKo eingereichten Buchtitel sowie auch die dabei angegebenen Metadaten zu den Einreichenden offen einsehbar zugänglich gemacht (vgl. Schumacher und Flüh 2022). So werden die ethischen Grundsätze von Citizen Humanities *Transparenz* in Bezug auf beteiligte Interessengruppen (vgl. Heinisch 2020: 15), *Offenheit* der Ergebnisse des Crowdsourcing (vgl. Dunn und Hedges 2012: 19) sowie ein informativer *Mehrwert für Beteiligte* (vgl. Heinisch 2020: 15) gewährleistet.

Der Fragebogen wird in drei Phasen in unterschiedlichen Communities verbreitet, deren Auswahl auf Basis der drei Dimensionen der partizipatorischen Wissenschaft – wissenschaftlicher Impact und Output, Lernen, Involviertheit und Empowerment der Teilnehmenden und gesellschaftlicher Impact und Awareness in Bezug auf die Thematik – getroffen wurde (vgl. Heinisch 2020: 155). Der Idee Wodwards folgend, dass Communities sich um verbindende Fragestellungen herum bilden, wird jede Phase von einer Frage geleitet (Wodward 2007: 117). Ergänzend wird eine Disseminationsstrategie mit digitalen und analogen Anteilen umgesetzt, die auf Erkenntnisse aus dem Disseminationsprojekt forTEXT zurückgreift (Gius et al. 2021, Schumacher und Gius 2022, Schumacher und Horstmann 2019). Die drei Phasen sind non-exklusiv, d.h., wenn z.B. Mitglieder der primären Zielgruppe aus Phase II schon in Phase I in den Community-Diskurs eintreten, so sind sie willkommen. Seit Beginn des Projektes und über alle Phasen hinweg wird der Blog des m*w-Projektes als Herzstück der Kommunikation genutzt.

Phase I: Wie wird ein literaturwissenschaftliches Korpus aufgebaut, das zur Basis automatisierter Gender-Klassifikation verwendet werden soll?

Kernzielgruppe dieser Phase ist die Digital-Humanities-Community, die hauptsächlich aufgrund der Methodik Interesse an dem Projekt zeigt. Als etabliertes Medium der Wissenschaftskommunikation innerhalb dieser Community, das darüber hinaus auch erhebliches Potential für geisteswissenschaftliche Wissenschaftskommunikation allgemein bietet (vgl. Geier und Gottschling 2019), wird ein Twitter-Account aufgebaut. Dabei setzen wir auf ein organisches Wachstum, bei dem Interesse an dem Projekt über die getwitterten Inhalte ausgelöst wird. Außerdem werden etablierte Informationskanäle genutzt, wie z.B. der Discord-Server DHall, die DHd-Mailingliste, der Fachinformationsdienst für Allgemeine und Vergleichende Literaturwissenschaft oder das Projektschaufenster der Webseite des DHd-Verbandes. Auch die Teilnahme an Fachkonferenzen stellt einen wichtigen Teil der ersten Phase dar.

Phase II: Was macht non-binäre Genderdarstellung aus?

In dieser Phase geht es darum, den ersten Outreach des Projektes zu generieren, indem Mitglieder der LGB-TIQ+-Community und deren sogenannte Allies, die Interesse an der Genderthematik aufweisen, angesprochen werden. Zentral ist dabei die zielgruppengenaue Ausrichtung. Auch Kenntnis literarischer Texte ist vonnöten, sodass eine Community an der Schnittstelle zwischen LGBTIQ+-Themen und Interesse für Literatur gefunden werden muss. Diese Phase ist eng mit der ersten Phase des Community-Buildings verzahnt. Beteiligte der DH-Community dienen als Multiplikator*innen, indem sie z.B. im Rahmen ihrer Lehre auf das Projekt DisKo aufmerksam machen. Seitens des Projektes werden Gastvorträge in universitären Lehrveranstaltungen durchgeführt. Außerdem werden Flyer in Bibliotheken ausgelegt, die mittels QR-Code auf die DisKo-Umfrage verweisen. Darüber hinaus wird DisKo bei der Plattform *Bürger schaffen Wissen* eingereicht.

Phase III: Wie bedeutsam ist non-binäre Genderdarstellung für unsere Gesellschaft?

Das Thema Gender und insbesondere (non-)Binarität wird aktuell in unterschiedlichen Bereichen des öffentlichen Lebens diskutiert. Die gesellschaftliche Brisanz der Gender-Thematik und die Relevanz für den alltäglichen Umgang miteinander ist aber nicht nur derzeit ein wichtiges Thema. Mit dem Projekt m*w, in dessen Rahmen DisKo aufgebaut wird, möchten wir offenlegen, dass auch in der Literaturgeschichte immer wieder Figuren eine Rolle spielen, die sich nicht in ein binäres Gender-System fügen lassen. Wir möchten zeigen, dass die aktuelle Debatte also nicht neu ist, sondern Gender-Diversität schon lange ein gesellschaftliches Thema ist, das u.a. in Kulturprodukten wie literarischen Texten eine wichtige Position einnimmt. In dieser Phase setzen wir die in Citizen-Science-Ansätzen häufig sehr stark verankerte Idee einer *Third Mission* von Forschungsprojekten (vgl. Heinisch 2020: 152–153) um, indem wir Ergebnisse mit einer möglichst breiten Öffentlichkeit teilen und in die aktuellen Debatten einfließen lassen. Darum suchen wir in dieser Phase weitere Wege der Wissenschaftskommunikation, wie z.B. über die Videoplattform TikTok oder den Bilder-Sharing-Dienst Instagram.

Erste Ergebnisse

Zum jetzigen Zeitpunkt (Stand Dezember 2022) befinden wir uns in Phase I des Citizen-Humanities-Projektes DisKo. Maßnahmen zur Verbreitung der Umfrage wurden auf Twitter, Mastodon, der DHd-Webseite, internen Kanälen der DNB und über Flyer umgesetzt. Außerdem wurde ein Gastvortrag im Rahmen der Ringvorlesung „Einführung in die Digital Humanities“ an der Universität Hamburg gehalten. Der Rücklauf ist noch nicht sehr hoch. Insgesamt gab es 17 Teilnehmende der Umfrage. Allerdings wurden von diesen insgesamt 31 Titel angegeben, was bedeutet, dass jede*r Teilnehmende t durchschnittlich 1,8 Titel eingereicht hat. Die tatsächliche Verteilung ist recht heterogen, es wurden bis zu sechs Titel

pro Person angegeben. Die freiwilligen Angaben zum eigenen Hintergrund wurden meist beantwortet, wie aus Abb. 1 ersichtlich wird.

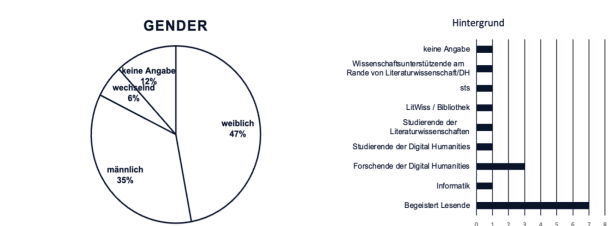


Abb. 1: Metadaten der Einreichenden der DisKo-Umfrage

Die eingereichten Titel umfassen derzeit eine Zeitspanne, die von 1928–2022 reicht, also 94 Jahre abdeckt. Der Schwerpunkt liegt mit 22 Erzähltexten auf Titeln, die nach 2000 erschienen sind. Nur acht der für DisKo vorgeschlagenen Texte sind vor dem Jahr 2000 erschienen. Es handelt sich bei den Texten sowohl um ursprünglich deutschsprachige Erzähltexte als auch um Übersetzungen. Zwar birgt die Integration von Übersetzungen für das Machine-Learning-Training die Gefahr, eine unkontrollierbare Variable einzubauen, da unklar ist, inwiefern die Ergebnisse der automatischen Erkennung davon beeinflusst werden. Dafür können aber Titel der Weltliteratur aufgenommen werden, die tatsächlich gelesen werden.

In Bezug auf die Genderdarstellungen reichen die Texte von der Erwähnung der nicht-stereotypen Genderidentität einer Figur (wie in Murakamis *Kafka am Strand*) bis hin zu einer Fülle nicht-stereotyper Genderidentitäten, die zum Hauptthema des Erzähltextes werden. Letzteres ist z.B. bei Evaristos *Mädchen, Frau, etc.* der Fall, sodass wir diesen Roman für unser Machine-Learning-Training als Testtext gewählt haben. Eine linguistische Besonderheit zeigt Leckies *Maschinen-Trilogie*, die durchgehend im generischen Femininum geschrieben wurde. Neben Murakami und Leckie wurden bisher acht weitere Texte ins Trainingskorpus übernommen. Von jedem der zehn Texte wurden zwei 2.000 Tokens umfassende Passagen ins Trainingskorpus integriert – eine vom Beginn und eine vom Ende des Textes (um eventuelle Transitionen berücksichtigen zu können). Erste Tests mit einem auf diesem 40.000 Tokens umfassenden Trainingskorpus trainierten Modell zeigen noch keine zufriedenstellenden Ergebnisse. Der F1-Score liegt insgesamt bei 0,35, die Erkennung der Kategorie „Divers“ bei 0. Ein Blick auf die annotierten Beispiele im Trainingskorpus zeigt, dass für diese Kategorie nur 60 Vorkommnisse annotiert wurden, während die Kategorien „Frau“ und „Neutral“ jeweils rund 500 Vorkommnisse aufweisen, „Mann“ sogar 902. Um hier zu einem ausgewogeneren Verhältnis zu kommen, könnten statt Anfangs- und Endpassagen, Ausschnitte aus den Texten ausfindig gemacht werden, in denen sich Genderzuschreibungen der Kategorie „Divers“ häufen. Eine andere Möglichkeit wäre die Annotation kompletter Erzähltexte, in denen non-binäre Genderdarstellungen zum Hauptthema gemacht werden. Nach Abschluss der Pilot-Trainingsphase werden wir darum ein erneutes Training mit relevanteren Samples oder Volltexten anschließen.

Fazit

Die Korpuskonstituierung ist für Projekte, die Verfahren des maschinellen Lernens einsetzen, ein Dreh- und Angelpunkt. Alle weiteren Verfahrensschritte und Ergebnisse wie z.B. die Performanz eines Classifiers oder Analyseergebnisse, die durch dessen Einsatz erzielt werden, werden vom genutzten Korpus massiv beeinflusst. Unser Konzept der Korpuskonstituierung greift sowohl arbeitspraktische Ansätze der Digital Humanities wie das *Random Sampling* oder die opportunistische Auswahl als auch kritische Betrachtungen von Aspekten wie Representation-Bias und Empowerment auf. Mithilfe einer Citizen-Humanities-Komponente begegnen wir zentralen Herausforderungen beim Aufbau des Diversitäts-Korpus DisKo. Die strategische Korpuskonstituierung, die wir in unserem Vortrag präsentieren und zur Diskussion stellen wollen, ist dabei nicht nur ein optionaler Bestandteil, sondern wird zur Gelingensbedingung für eine möglichst diverse und ausgewogene Datenbasis.

Bibliographie

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. 1. Aufl. Bd. 4. Digital Humanities Research. Bielefeld, Germany: transcript Verlag / Bielefeld University Press. <https://doi.org/10.14361/9783839459256>.

Chenier, Elise. 2014. „Oral History and Open Access: Fulfilling the Promise of Democratizing Knowledge“ <https://nanocrit.com/issues/issue5/notes-women-who-rock-making-scenes-building-communities-participatory-research-community-engagement-and-archival-practice> [zugegriffen: 6. Juli 2021].

Christie, Alex, Jana Millar Usiskin, Jentery Sayers und Kathryn Tanigawa. 2020. „Introduction: Digital Humanities, Public Humanities - Nanocrit.Com.“ <https://nanocrit.com/issues/issue5/introduction-digital-humanities-public-humanities> [letzter Zugriff: 6. Juli 2021].

Dunn, Stuart E. und Mark Hedges. 2012. *Crowd-Sourcing Scoping Study: Engaging the Crowd with Humanities Research*. <https://www.semanticscholar.org/paper/Crowd-Sourcing-Scoping-Study%3A-Engaging-the-Crowd-Hedges-Dunn/9940b0520332a6b0605559fd7c8c46672b3f-b655> [zugegriffen: 6. Juli 2021].

Flüh, Marie und Mareike Schumacher. 2021. „Digitale Diachrone Korpusanalyse Am Beispiel Des Projekts „m*W – Gender Stereotype in Der Literatur“. *Digital Humanities and Gender History*. <https://doi.org/10.22032/dbt.49173>.

Flüh, Marie, und Mareike Schumacher. 2022. „Jung, wild, emotional? Rollen und Emotionen Jugendlicher in zeitgenössischer Fantasy-Literatur“. Gehalten auf der DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2022), Potsdam, März 7. <https://doi.org/10.5281/zenodo.6327983>.

Flüh, Marie, Jan Horstmann und Mareike Schumacher. 2022. „Distant Gender Reading

Genderaspekte in Fantasy-Jugendromanen von 2008 bis 2020". In: Weertje Willms (Hg.): *Gender in der deutschsprachigen Kinder- und Jugendliteratur: Vom Mittelalter bis zur Gegenwart*. *Gender in der deutschsprachigen Kinder- und Jugendliteratur*. Berlin: De Gruyter. <https://www.degruyter.com/document/isbn/9783110726404/html?lang=de>.

Flüh, Marie, Mark Lemke und Mareike Schumacher. 2022: The model of choice. Using pure CRF- and BERT-based classifiers for gender annotation in German fantasy fiction. In: *Digital Humanities 2022 – Responding to Asian Diversity (DHTokyo)*.

Flüh, Marie und Mareike Schumacher (forthcoming): "Macht versus Emotion. Handlungstreibende Muster in Günderrodes Dramen digital, distant und scalable gelesen". In *Noch Zukunft haben. Das Werk Karoline von Günderrode neu gelesen. Neue Romantikforschung*, hg. von Roland Borgards, Martina Wernli und Frederike Middelhoff Berlin: Springer.

Geier, Andrea und Markus Gottschling. 2019. „Wissenschaftskommunikation auf Twitter? Eine Chance für die Geisteswissenschaften!" *Mitteilungen des Deutschen Germanistenverbandes* 66 (3): 282–91. <https://doi.org/10.14220/mdge.2019.66.3.282>.

Gius, Evelyn, Mareike Schumacher, Dominik Gerstorfer, Malte Meister, Sandra Bläß, Marie Flüh, Jan Horstmann, Janina Jacke, Christian Bruck und Marco Petris (2021): forTEXT. Literatur digital erforschen. URL: <https://fortext.net> [zugegriffen: 6. Juli 2021].

Gius, Evelyn, Krüger Katharina und Carla Sökefeld. 2019. „Korpuserstellung als literaturwissenschaftliche Aufgabe". Gehalten 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, März 16. <https://doi.org/10.5281/zenodo.4622112>.

Habell-Pallán, Michelle, Sonnet Retman und Angelica Macklin. 2014. „Notes on Women Who Rock: Making Scenes, Building Communities: Participatory Research, Community Engagement, and Archival Practice - nanocrit.com", 2014. URL: <https://nanocrit.com/issues/issue5/notes-women-who-rock-making-scenes-building-communities-participatory-research-community-engagement-and-archival-practice> [zugegriffen: 3. August 2021].

Heinisch, Barbara. 2020. "Citizen Humanities as a Fusion of Digital and Public Humanities?" *Magazén*, no. 2 (December): JournalArticle_3442. <https://doi.org/10.30687/mag/2724-3923/2020/02/001>.

Henny-Kramer, Ulrike und Frederike Neuber. 2017. „Criteria for Reviewing Digital Text Collections, version 1.0 |". *IDE – Institut für Dokumentologie und Editörrik* (blog). URL: <https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/>. [zugegriffen: 3. August 2021].

Lassner, David. 2020. *Bericht aus dem Workshop zu Bias in Datensätzen und ML-Modellen. Erkennung und Umgang in den DH*. URL: <https://digitalintellectuals.hypotheses.org/3262> [zugegriffen: 27. Juli 2022].

Schöch, Christof. 2017. „Aufbau von Datensammlungen". In *Digital Humanities: Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Reh-

bein, 223–33. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05446-3_16.

Schumacher, Mareike und Evelyn Gius. 2022. "forTEXT – Literatur digital erforschen". *Mitteilungen des Deutschen Germanistenverbandes* Jg. 69, Heft 2. 2022. Vandenhoeck & Ruprecht Verlage.

Schumacher, Mareike, und Flüh, Marie. 2020. „m*w Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen. Genderstereotype und -bewertungen in der Literatur des 19. Jahrhunderts". In *DHd2020: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, hg. von Christof Schöch 162–167. <https://doi.org/10.5281/ZENODO.4621892>.

Suresh, Harini und John V. Guttag. 2019. "A framework for understanding unintended consequences of machine learning." arXiv preprint arXiv:1901.10002.

Schumacher, Mareike und Marie Flüh. 2022. "Jung, wild, emotional? Rollen und Emotionen Jugendlicher in zeitgenössischer Fantasy-Literatur". In *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum" (DHd 2022)*, Potsdam. <https://doi.org/10.5281/zenodo.5555952>.

Schumacher, Mareike, und Marie Flüh. 2022. „DisKo – das Diversitäts-Korpus". Projektwebseite. m*w (blog). 2022. URL: <https://msternchenw.de/diversitaets-korpus/>. [zugegriffen: 3. August 2021].

Schumacher, Mareike, und Jan Horstmann. 2019. „Social Media, YouTube und Co: Multimediale, multimodale und multicodierte Dissemination von Forschungsmethoden in forTEXT". Frankfurt am Main und Mainz, März 16. <https://doi.org/10.5281/zenodo.4622253>.

Woodward, Kathleen. 2009. "The Future of the Humanities- in the Present & in Public." *Daedalus* 138 (1): 110–23. <https://doi.org/10.1162/daed.2009.138.1.110>.

EGRAPHSEN. Von einem Nebenprodukt des Supervised Machine Learnings zu einer evidenzbasierten Malerzuweisung auf attischen Vasen

Kipke, Marta

marta.kipke@uni-goettingen.de

Georg-August-Universität Göttingen, Deutschland

Einleitung

Woran erkennt man die Handschrift eines Malers? Was macht den Stil eines Künstlers aus? Diese Fragen be-

schäftigen die Klassische Archäologie bezüglich antiker griechischer Vasenbilder seit über einem Jahrhundert. Im Projekt EGRAPHSEN¹ wird die Methode der Malerzuweisung als Klassifikationsproblem mit Supervised Machine Learning Verfahren untersucht. Das Ziel ist es, einerseits neue Erkenntnisse über Maler und ihre Stile zu gewinnen, andererseits über die Methode zu reflektieren und das Vorgehen eines neuronalen Netzes dem traditionellen menschlichen Zugang gegenüberzustellen.

In der Klassischen Archäologie hat sich mit dem Vasenforscher John D. Beazley (* 13. September 1885; † 6. Mai 1970) eine Expertise und Kennerschaft ausgebildet, die ihm eine kaum anfechtbare Autorität verliehen hat. Er hat hunderttausenden Vasenbildern Maler zugewiesen, indem er sie insbesondere in ihren zeichnerischen Details verglichen hat (zu Beazley und seiner Methode s. Neer 1997, 7-16; Driscoll 2019, 106-110). Wenn auch die Methodenkritik und -reflexion in den letzten Jahren zugenommen hat (Graepler 2016, 18-21), so werden auch noch in aktuellen Publikationen Maler identifiziert und ihre Œuvre erweitert (z.B. Padgett 2017, 392-399). In den letzten Jahren fand mit dem Aufkommen des maschinellen Lernens in den Geisteswissenschaften jedoch auch ein Perspektivwechsel statt, und sowohl die Archäologie und Kunstgeschichte als auch die Informatik sind sehr am Erkenntnisgewinn und an der Methodenreflexion durch die Verwendung künstlicher neuronaler Netze in diesem Gebiet interessiert (Ma et al 2017, 1174-1176; Elgammal und Kang und DenLeeuw 2018, 42-49; Bell und Offert 2021, 4-9; Langmead 2021, 2-19). Das führt auch zu Veränderungen im Anspruch an die der Publikation von Forschungsergebnissen und -daten.

Beim Training eines Convolutional Neural Network (CNN) entstehen große Datenmengen: Annotationen, vorverarbeitete Bilder, Merkmalsvektoren, Meta- und Paradata. In unserem Projekt streben wir an, diese Daten auf eine Art und Weise zu veröffentlichen, die einen forschungsorientierten und methodenkritischen Zugang erlauben. Das Konzept dieser Veröffentlichung in Form einer Datenbank soll im Zentrum dieses Beitrags stehen. Um das Problem zu verdeutlichen, soll zu Beginn der methodische Zugang kurz umrissen werden. Dann soll zunächst die Publikation der Bild- und Metadaten vor dem Hintergrund bestehender Tools der digitalen Kunstgeschichte erläutert werden. Schließlich wird auf die trainierten Modelle eingegangen und besprochen werden, inwiefern diese Verwendung in der Datenbank finden können.

Versuchsaufbau und Vorgehen

Zunächst soll die Datengrundlage benannt werden: Welche Informationen werden dem CNN zugeführt, um es zu trainieren? Bei der traditionellen Methode der Malerzuweisung stehen sehr spezifische Details im Vordergrund. Wir möchten mit computergestützten Methoden untersuchen, welche Bildelemente und -eigenschaften tatsächlich die Handschrift eines Malers erkennen lassen. Allerdings sind nur wenige Vasenbilder tatsächlich mit einer Malersignatur versehen, und auf diese Weise ergibt sich keine kritische Menge für das Training eines CNN. Deswegen haben wir uns in EGRAPHSEN entschie-

den, zusätzlich zu signierten Vasen auch die Malerzuweisungen von John D. Beazley als Ground Truth für die Klassifikation zu verwenden.

Da es um die Details im Bild gehen soll, nutzen wir außerdem nicht die gesamte Darstellung für das Training. Stattdessen trainieren wir mit unterschiedlichen Kombinationen von vordefinierten Bildausschnitten. Für diesen Zweck wurde eine kleinteilige Ontologie entwickelt, die die Körperteile der Figuren und die Bildbestandteile in ihren räumlichen Ausmaßen, Bezeichnungen und Bezügen zueinander klar festgelegt. So sind die Bildbestandteile nicht nur benannt, sondern meistens auch mit weiteren Informationen zu ihrer Darstellung angereichert. Sie sind außerdem in einem hierarchischen System strukturiert, sodass eine Figur auf unterschiedlichen Detailebenen betrachtet werden kann: Es könnte eine gesamte Figur für das Training verwendet werden, auf der nächsten hierarchischen Detailstufe lediglich der Arm, oder auf der nächsten hierarchischen Detailstufe auch nur die einzelnen Bestandteile des Armes (für eine ausführliche Beschreibung der Ontologie s. Kipke und Brinkmeyer, 2022, 3-5). Auf diese Weise kann man die Bilder ihrer Komplexität angemessen analysieren und mit unterschiedlichen Merkmalen auf verschiedenen Detailstufen experimentieren, ohne den Bildkontext vollständig zu verlieren.

Nach diesem System wurden 4.188 einzelne Figuren und damit insgesamt über 200.000 kleinteiligen Einzelannotationen von 38 Malern vorgenommen.² Damit liegt eine hohe Dichte an Informationen pro Vasenbild vor. Diese Einzelannotationen werden extrahiert, vorverarbeitet und dem CNN zugeführt. Der Nutzen der ausgeschnittenen Annotationen soll an dieser Stelle jedoch nicht enden. Im Gegenteil sollen diese der Klassischen Archäologie als weitere Hilfestellung bei der Malerzuweisung dienen und der unkontrollierbaren Abstraktion des CNN sowie der autoritätsgesteuerten Zuweisung einzelner Forscher:innen gegenüberstehen. Somit soll der Malerzuweisung eine visuelle Evidenz verschafft werden, die in einer Abwägung unterschiedlicher Methoden zu neuen Erkenntnissen führen soll.

Denn während das Modell erfreuliche Ergebnisse bei Malern mit zahlreichen bekannten Werken und häufig verwendeten Labels wie Augen und Händen liefert, werden auch die Grenzen und Gefahren schnell deutlich: Erstens bleibt der Mensch dem neuronalen Netz dort überlegen, wo nur wenige Werke bekannt oder wo diese sehr heterogen sind. Schließlich reichten John D. Beazley zuweilen nur zwei Vasenbilder, um einen Maler zu identifizieren (z.B. Beazley 1963, 21). Das ist eine Datenlage, auf der Supervised Machine Learning Verfahren aktuell noch keine befriedigenden Lösungen liefern können. Zweitens besteht noch Unklarheit darüber, ob nicht stärkere Eigenschaften des Bildes, wie etwa Zeit-, Gattungs-, oder Gefäßstil trotz Bildausschnitten zu einem unerwünschten Bias im Training führen und so den Erkenntnissen über den persönlichen Stil der Maler im Weg stehen könnten. Dies soll in unseren Daten mithilfe strukturierter Analysemöglichkeiten problematisiert werden können.

Konzeption: Der digitale Bildvergleich als Grundlage visueller Evidenz

Wie kann eine Publikation der Annotationen nun bestmöglich diesen Zweck erfüllen? Der detaillierte Bildvergleich steht nicht nur im Zentrum der Meisterforschung und Malerzuweisung, sondern bildet den Kern aller Bildwissenschaften. So steht die digitale Kunstgeschichte bereits in einer Tradition von Bilddatenbanken, die ein assoziatives Vorgehen und Sortieren ermöglichen und sich dabei auf Aby Warburg und den Entstehungsprozess seines Bilderatlas (Hristova 2016, 117-120; Du Preez 2020) berufen. Ein Beispiel hierfür ist die Anwendung Meta-Image, die im Rahmen eines gleichnamigen, DFG geförderten Projekts in Köln und Lüneburg entwickelt wurde. Die Anwendung erlaubt das stetige Neuankordnen von Bildern in Netzwerke, was eine Nachvollziehbarkeit des Erkenntnisgewinns ermöglicht, und damit eine visuelle Evidenz für die Beantwortung ikonographischer oder gestaltungstechnischer Fragestellungen schafft (Dieckmann und Warnke 2018, 79-90). Die Anwendung simuliert den Leuchttisch von Kunsthistoriker:innen, jedoch ohne von seinen physischen Grenzen eingeschränkt zu sein. Dieser Ansatz scheint auch für den Detailvergleich einzelner Bildausschnitte sehr lohnend.

Mehr Funktionalitäten im Analyseprozess und Unabhängigkeit zu Bilddatenbanken bietet das webbasierte Tool *ARIES*. Es wurde Team von Forschern aus Amerika (New York) und Brasilien (Rio de Janeiro) entwickelt (Projektwebsite: <https://artimageexplorationspace.com/>). Dort können eigene Meta- und Bilddaten importiert und mithilfe unterschiedlicher Tools analysiert werden. So kann man beispielsweise unterschiedliche Formen des Überlagerns der Bilder (Crissaff 2017 1-8; Deuch 2021, 7-12) simulieren. Jedoch ist die Möglichkeit zur Verwendung der Metadaten in einem Umfang und einer Komplexität, wie sie in diesem Projekt vorliegen, nicht möglich. Um den bestmöglichen Nutzen in einer kontrollierten Umgebung zu gewährleisten, wird stattdessen die Entwicklung einer eigenen Benutzeroberfläche und Exportmöglichkeiten für die Weiterverwendung in anderen Anwendungen angestrebt.

Eine solche Datenbank soll dabei nicht nur die annotierten Bildausschnitte zur Verfügung stellen, sondern auch Metadaten zu den Vasenbildern im Umfang des Beazley Archives (Smith 2005, 23-24; Kurtz 2009, 39-46) enthalten, die im Projekt um weitere Informationen wie beispielsweise eine kleinteilige Datierung der untersuchten Vasenbilder und Maße der Vasen ergänzt wurden. Zusätzlich zu den Metadaten soll das hierarchische Annotationssystem mit all seinen weiteren Informationsebenen als Grundlage für die Suchmaske dienen. Dabei soll die Suche nach drei Kategorien aufgefächert sein:

1. Annotationslabel: Es ist möglich, ein oder mehrere Labels (z. B. Hände, s. Fig. 1) auszuwählen, die für den Vergleich verwendet werden sollen. Dabei können auch bestimmte Zustände des Labels gewählt werden, die ebenfalls in der Annotation berücksichtigt wurden (z. B. nur Hände, die etwas halten oder Münder, die Flöte spielen, etc.).

2. Malerauswahl: Man kann einen oder mehrere Maler auswählen, die mit den gewählten Labels untersucht werden sollen. Dabei werden die Zuweisungen in vier Stufen von Zuordnungssicherheit geteilt: 1. Signierte Werke, 2. von Beazley zugeordnet, 3. von *Ceuvre*-Forschern zugeordnet (z.B. J. H. Oakley beim Achilleus Maler (Oakley 1997)) und 4. von weiteren Vasenforscher:innen zugeordnet. Dies soll Transparenz und Nachvollziehbarkeit über die Sicherheit der Zuordnung gewährleisten.

3. Externe Kriterien: Die Zuweisung selbst ist bereits ein subjektives Kriterium. Deswegen soll es auch möglich sein, die Labels nach übergeordneten Kriterien in unterschiedlichen Kombinationen zu suchen und sie so im Kontext ihrer Datierung, Gefäßform, Figurengröße und ihres Motivspektrums zu betrachten, um damit das Verhältnis von Zeit- und Gattungsstil zum persönlichen Stil des Malers beurteilen zu können.

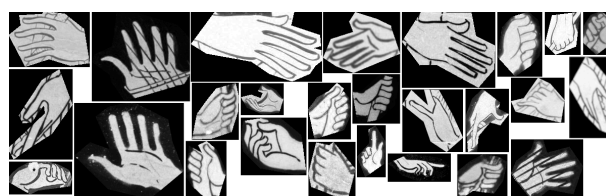


Abb. 1: Ein Vergleich der Hände lässt die 'Handschrift' des Berliner Malers erkennen.

Schließlich können sowohl die Zuweisung als auch die externen Kriterien vernachlässigt und die Bilder unabhängig davon angezeigt werden. Diese facettierte Suche soll dazu beitragen, sich von der Malerzuweisung durch bestimmte Forscher:innen zu lösen und einen Erkenntnisgewinn aus den Bildern heraus zu ermöglichen. Das Ergebnis dieser Suche soll dann ebenfalls eine Leinwand sein, auf der die Bilder nach den bereits genannten Kriterien sortiert werden können. Weiterhin hat jede Einzelannotation einen Datenbankeintrag, in dem die Metadaten eingesehen werden können. Ein Export des Suchergebnisses soll es schließlich ermöglichen, die Bilder auch in anderen Anwendungen zu importieren um weitere Untersuchungen durchzuführen.

Konzeption: Einsatz künstlicher neuronaler Netze für die Bildsuche

Diese Suchmaske basiert auf den Metadaten zur Vase und der Ontologie, die im Projekt entwickelt wurde. Die Merkmalsvektoren, Zuweisungen und Funktionalität des CNN sind darin noch nicht inbegriffen. Im Folgenden soll erörtert werden, inwiefern diese Daten nutzbar gemacht werden sollen.

Die keyword basierte Suche in *EGRAPHSEN* steht einem Trend entgegen, der eine gewisse Loslösung von Schlagworten in Bilddatenbanken anstrebt. Im Bereich des Content Based Image Retrieval suchen Forscher:innen nach bildimmanenten Eigenschaften, mit denen Deskriptoren für jedes einzelne Bild definiert werden können. Zwischen diesen Deskriptoren können Ähnlich-

keitsbeziehungen berechnet und so Suchergebnisse generiert werden, die sowohl den Suchprozess erleichtern, als auch das Bild mit seinen Eigenschaften in den Mittelpunkt stellen (Tyagi 2017, 1-22). Dabei können neben einfach auslesbaren low-level-features wie Farben und Formen auch künstliche neuronale Netze verwendet werden, um Features zu extrahieren und für die Beschreibung der Bilder zu verwenden (Aasia und Sharma 2017, 1049; Hameed 2021, 21-32). Insbesondere Methoden der Computer Vision können bei komplexen und heterogenen Bildern, wie sie von den digitalen Geisteswissenschaften erforscht werden, den Umgang mit großen Bildkorpora erleichtern (Bell und Ommer 2016, 71-72; Bell und Ommer 2018, 67-72; Resig 2014).

Da es in EGRAPHSEN explizit um Stilanalyse geht, ist die Verwendung von low-level-features zu banal. Die Experimente im Projekt haben unterschiedliche trainierte Modelle ergeben, die genutzt werden können, um Features zu extrahieren und zu visualisieren. Diese Features könnten zwar auch für ein Content Based Image Retrieval verwendet werden, jedoch ist für EGRAPHSEN eine derartige Funktionalität aus verschiedenen Gründen nicht vorgesehen. Im Projekt ist das Ausmaß der experimentellen Abstraktion durch das CNN sehr deutlich geworden: Noch mehr als bei Zuweisungen durch Archäolog:innen ist die Nachvollziehbarkeit der Ergebnisse stark mit einer Interpretation dieser verbunden. Das führt zu spannenden Erkenntnissen über die Methode der Malerzuweisung und über die Funktion neuronaler Netze als solche, würde eine Datenbank in ihrer Funktionalität jedoch zu stark mit einer Subjektivität färben, die nicht mehr nachvollziehbar sein kann. Statt also die Ergebnisse der Experimente in der Datenbank funktional zu nutzen, soll sie ihnen gegenüberstehen und zur weiteren Forschung, Verifizierung und Vertiefung der methodischen Reflexionen dienen – insbesondere dort, wo die Verfahren des maschinellen Lernens derzeit an ihre Grenzen kommen. Auf der einen Seite steht also die Analyse der Maler und ihrer Beziehungen zueinander mithilfe eines CNNs, und auf der anderen Seite eine Anwendung zur Nachvollziehbarkeit dieser Ergebnisse und Vertiefung der Forschung durch menschliche Expert:innen.

Für den Kern der Datenbank – ihre Strukturierung und Funktionalität – ist also kein Einsatz von neuronalen Netzen vorgesehen. An zwei weiteren Stellen sollen aber die Ergebnisse der Experimente und Nebenprodukte des Vorgehens genutzt werden.

So werden die vom CNN extrahierten Features, die Merkmalsvektoren und die Zuweisungen als reine Werte in der Datenbank enthalten sein, um die Nachvollziehbarkeit der Experimente nutzerfreundlich zu halten und in einer Domäne zu dokumentieren.

Zudem sollen für das Wachstum und die Pflege der Datenbank Teile unseres semi-automatischen Annotations-Workflows nutzbar gemacht werden. Um auf eine kritische Datenmenge für das Training der Modelle zu kommen, wurde ein Annotationstool entwickelt, das auf einer Open Source Software Version des Tools LabelMe (Wada 2022) basiert und in EGRAPHSEN um eine Object Detection Komponente erweitert wurde. Der Workflow sieht vor, dass die Object Detection Vorschläge zur Annotation macht, die dann individuell angepasst werden können (Kipke und Brinkmeyer, 2022, 5-6). Da im Hinter-

grund von EGRAPHSEN unsere Projektdatenbank steht, liegen bereits Metadaten zu den Vasenbildern vor. Eine Datenpipeline mit Nutzung dieses Tools und der Pre-Processing Algorithmen soll es ermöglichen, weitere annotierte Ausschnitte aus Vasenbildern der Datenbank hinzuzufügen (vgl. Fig. 2).

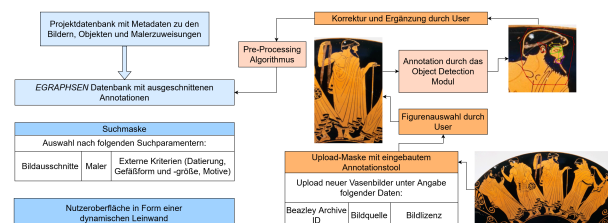


Abb. 2: Pipeline zur weiteren Anreicherung und Nutzung der Datenbank. Blau betrifft die Datenbank, ihren Aufbau und Nutzung, orange den Anreicherungsprozess.

Fazit

In EGRAPHSEN wurde die Malerzuweisung als traditionelle Methode untersucht und diese der Funktionsweise von CNNs gegenübergestellt. Obwohl beide Methoden spezifische Bilddetails in den Mittelpunkt stellen, lassen sich auf beiden Seiten Vorteile benennen: Das menschliche Auge hat die Fähigkeit, auch bei geringen Bildmengen komplexe Transferleistungen zu erbringen und die Bilder in ihrer Heterogenität sowie im Aufbau zu verstehen, während das CNN einen Blick auf die Bilder ermöglicht, der nicht durch spezifisch menschliche Expertenkenntnis, motivische Zusammenhänge und unterbewusste Annahmen gefärbt sein muss. Jedoch besteht auch stets die Gefahr, dass menschliche Expert:innen bereits in der Bildauswahl und im Training die Ergebnisse beeinflussen und das CNN durch fehlendes Bildverständnis andere Fehlannahmen, z. B. über die Bildqualität, aufweisen kann.

Deswegen wurde in EGRAPHSEN großer Wert darauf gelegt, dass die Kategorisierung der Bildausschnitte auf einer Ontologie basiert, die bewusst auf interpretative Aspekte verzichtet und die Bilder primär in ihrer Form beschreibt. Dadurch, dass die Bilder aus ihrem Kontext extrahiert und nach frei wählbaren, formalen Kriterien sortiert werden können, bekommt die Malerzuweisung ihrerseits eine Evidenz, wie sie häufig sonst nicht vorhanden ist. Denn sei es Mensch oder Maschine – viele, stärkere Bildmerkmale beeinflussen die Zuweisung häufig erheblich.

Es soll eine Anwendung geschaffen werden, in der der Einfluss solche Merkmale, wie etwa des Motivs oder der Vasenform, möglichst reduziert werden. Mithilfe unterschiedlicher dynamischer Kriterien können sich Expert:innen zwischen den reinen Bilddaten auf der einen Seite und experimenteller Abstraktion durch das CNN auf der anderen Seite positionieren. So entsteht eine Forschungsumgebung, in der das menschliche Auge und die hochkomplexen Transferleistungen ausgebildeter Bildwissenschaftler:innen im Wechselspiel mit der KI ihr Potential weiter entfalten können.

Fußnoten

1. Gefördert durch das Niedersächsische Kultusministerium und SPRUNG (ehemals „Niedersächsisches Vorab“). In Kooperation mit dem Machine Learning Lab der Universität Hildesheim. Unter der Mitarbeit von Prof. Dr. Martin Langner, Prof. Dr. Lars Schmidt-Thieme und Lukas Brinkmeyer M. Sc.
2. Besonderer Dank gilt an dieser Stelle den studentischen Hilfskräften, die die Annotationen durchgeführt haben: Firmin Forster (B. A.), Manuel Janda (B. A.), Maja Leone (B. A.) und Max Maletzki (B. A.).

Bibliographie

- Ali, Aasia und Sanjay Sharma. 2017. "Content based image retrieval using feature extraction with machine learning". In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1048-1053. <https://doi.org/10.1109/ICCONS.2017.8250625>.
- Beazley, John D. 1963. *Attic Red-figure Vase-painters*. 2. Aufl. Oxford: Clarendon Press.
- Bell, Peter und Björn Ommer. 2016. "Visuelle Erschließung (Computer Vision als Arbeits- und Vermittlungstool)". *Elektronische Medien & Kunst, Kultur und Historie* 23: 67-73.
- Bell, Peter und Björn Ommer. 2018. "Computer Vision und Kunstgeschichte. Dialog zweier Bildwissenschaften". In *Computing Art Reader: Einführung in die digitale Kunstgeschichte*, hg. von Piotr Kuroczyński, Peter Bell und Lisa Dieckmann, 61-75. <https://doi.org/10.11588/art-historicum.413.c5769>.
- Bell, Peter und Fabian Offert. 2021. "Reflections on connoisseurship and computer vision." *Journal of Art Historiography* 24. <https://doi.org/10.48352/uobx-jah.00003418>.
- Crissaff, Lhaylla, Louisa Wood Ruby, Samantha Deutch, R. Luke DuBois, Jean-Daniel Fekete, Juliana Freire und Claudio Silva. 2017. "ARIES: enabling visual exploration and organization of art image collections". *IEEE computer graphics and applications* 38, Nr. 1: 91-108. <https://doi.org/10.1109/MCG.2017.377152546>.
- Deutch, Samantha. 2021. "Art Image Exploration Space (ARIES): A response to the image needs of art library patrons". *Art Libraries Journal* 46, Nr. 1: 7-12. <https://doi.org/10.1017/alj.2020.31>.
- Dieckmann, Lisa, und Martin Warnke. 2018. "Meta-Image und die Prinzipien des Digitalen im Mnemosyne-Atlas Aby Warburgs". In *Computing Art Reader: Einführung in die digitale Kunstgeschichte*, hg. von Piotr Kuroczyński, Peter Bell und Lisa Dieckmann, 79-96. <https://doi.org/10.11588/art-historicum.413.c5770>.
- Driscoll, Eric. 2019. "Beazley's Connoisseurship: Aesthetics". Natural History, and Artistic Development. In *Dossier. Corps antiques: morceaux choisis*, hg. von Florence Gherchanoc und Stéphanie Wyler, 101-120.
- Du Preez, Amanda. 2020. "Approaching Aby Warburg and Digital Art History: Thinking Through Images". In *The Routledge Companion to Digital Humanities and Art History*, hg. von Kathryn Brown, 374-385. London: Routledge.
- Elgammal, Ahmed, Yan Kang und Milko Den Leeuw. 2018. "Picasso, Matisse, or a Fake? Automated Analysis of Drawings at the Stroke Level for Attribution and Authentication". *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 42-50. <https://doi.org/10.1609/aaai.v32i1.11313>.
- Graepler, Daniel. 2016. "Künstlerhand und Kennerauge. Die Zuschreibung als archäologisches Methodenproblem". In *Töpfer, Maler, Werkstatt. Zuschreibungen in der griechischen Vasenmalerei und die Organisation antiker Keramikproduktion*, hg. von Norbert Eschbach und Stefan Schmidt, 14-24. München: C. H. Beck.
- Hameed, Ibtihal M, Sadiq H. Abdulhussain und Basheera M. Mahmmod. 2021. "Content-based image retrieval: A review of recent trends". *Cogent Engineering* 8, Nr. 1. <https://doi.org/10.1080/23311916.2021.1927469>.
- Hristova, Stefka. 2016. "Images as Data: Cultural Analytics and Aby Warburg's Mnemosyne". *International Journal for Digital Art History*, Nr. 2, 116-133. <https://doi.org/10.11588/dah.2016.2.23489>.
- Kipke, Marta und Lukas Brinkmeyer. 2022. "Deep Level Annotation for Painter Attribution on Greek Vases utilizing Object Detection". In *SUMAC'22: Proceedings of the 4th workshop on Structuring and Understanding of Multimedia heritAge Contents*. <https://doi.org/10.1145/3552464.3555684>.
- Kurtz, D. 2009. "www.beazley.ox.ac.uk. From apparatus of scholarship to web resource. The Beazley Archive 1970-2008". *Archeologia e Calcolatori*, Nr. 20: 37-46.
- Langmead, Alison, Christopher J. Nygren, Paul Rodriguez und Alan Craig. 2021. "Leonardo, Morelli, and the Computational Mirror." *DHQ: Digital Humanities Quarterly* 15, Nr. 1. <http://www.digitalhumanities.org/dhq/vol/15/1/000540/000540.html>.
- Ma, Daiqian, Feng Gao, Yan Bai, Yihang Lou, Shiqi Wang, Tiejun Huang und Ling-Yu Duan. 2017. "From part to whole: who is behind the painting?" In *Proceedings of the 25th ACM international conference on Multimedia*, 1174-1182. <https://doi.org/10.1145/3123266.3123325>.
- Neer, Richard. 1997. "Beazley and the Language of Connoisseurship". *Hephaistos* 15: 7-30.
- Oakley, John Howard. 1997. *The Achilles Painter*. Mainz: Phillip von Zabern.
- Padgett, J. Michael, hg. 2017. *The Berlin Painter and his World. Athenian Vase-Painting in the Early Fifth Century BC*. New Haven: Yale University Press.
- Resig, John. 2014. "Using computer vision to increase the research potential of photo archives". *Journal of Digital Humanities* 3, Nr. 2. <http://journalofdigitalhumanities.org/3-2/using-computer-vision-to-increase-the-research-potential-of-photo-archives-by-john-resig/>.
- Smith, Tyler Jo. 2005. "The Beazley archive: inside and out". *Art Documentation: Journal of the Art Libraries Society of North America* 24, Nr. 1: 22-25.
- Tyagi, Vipin. 2017. *Content-based image retrieval*. Singapur: Springer Nature. <https://doi.org/10.1007/978-981-10-6759-4>.
- Wada, Kentaro. Labelme: Image Polygonal Annotation with Python, 2022. <https://doi.org/10.5281/zenodo.5711226>. <https://github.com/wkentaro/labelme> (zugegriffen 02. August 2022).

Einfluss des häufigen Lesens auf Textwahrnehmung: Ergebnisse eines Leseexperiments

Glawion, Anastasia

anastasia.glawion@tu-darmstadt.de
TU Darmstadt, Deutschland

Weitin, Thomas

thomas.weitin@tu-darmstadt.de
TU Darmstadt, Deutschland

In dem Vortrag werden Ergebnisse eines Leseexperiments vorgestellt, welches unter anderem darauf abzielte, die Lücke zwischen psychologisch orientierten Lesereaktionsstudien und literaturwissenschaftlich fundierten Rezeptionsstudien (Kavanagh, 2021) zu schließen. Die Stimuli umfassten Passagen aus der beliebten "Harry Potter"-Buchreihe in deutscher Sprache sowie Auszügen aus „Harry Potter“-Fanfictions. In dem Experiment sollten folgende Forschungsfragen beantwortet werden:

1. Wie beeinflusst der emotionale Gehalt der Texte die Reaktion der Lesenden?
2. Welche Wirkung hat der Hintergrund der Lesenden: gibt es Unterschiede in der Wahrnehmung, die durch Leseerfahrung und Fandom-Affinität bedingt sind?

In der Studie wurde eine Reihe von Messmethoden verwendet, darunter die Messung der Augenbewegungen (inklusive der Pupillengröße) und des Hautleitwerts (GSR) der Teilnehmer:innen. Diese beiden Messmethoden werden am häufigsten als Marker von emotionaler Reaktion in Betracht gezogen.

Die Originaltexte unter den Stimuli umfassten 40 "neutrale" Texte, 40 Texte, die als "furchteinflößend", und 40 Texte, die als "fröhlich" gekennzeichnet waren (s. Tabelle 1). Diese Textstellen sowie ihre Sentimentmarkierungen wurden aus einer früheren Lesestudie von Hsu (2015) übernommen. Die Liste der Stimuli wurde um Fanfiction-Texte erweitert, die zuvor von 82 Fanfiction-Leser:innen in einer Umfrage ausgewählt wurden, weil sie besonders starke Emotionen bei ihnen ausgelöst hatten.

Tabelle 1.

Stimulus	Sentiment
Als Hagrid Harrys Gesicht sah, strahlte er, ohne die verdutzten Blicke der vorübergehenden Muggel zu bemerken. "Harry!" dröhnte er, und kaum war Harry aus dem Wagen gestiegen, schloss Hagrid ihn auch schon in eine knochenbrechende Umarmung.	Fröhlich
Das Wetter draußen vor den Zugfenstern war so durchwachsen, wie es den ganzen Sommer über gewesen war; sie fuhren streckenweise durch kalten Nebel, dann wieder in schwaches klares Sonnenlicht.	Neutral
Mindestens hundert Dementoren, die verummten Gesichter ihm zugewandt, standen dort unter ihm. Es war, als würde eiskaltes Wasser in seiner Brust aufsteigen und ihm die Eingeweide abtöten. Und dann hörte er es wieder... Jemand schrie, schrie im Innern seines Kopfes... eine Frau.	Furchteinflößend

Die vielseitige Auswahl der Stimuli in der Studie von Hsu deckte unterschiedliche Aspekte auf, die mit Wirkung von Literatur verbunden sind. Einer davon war für unsere Analysen besonders anregend: es wurde ein Zusammenhang zwischen Immersion und emotionalen Inhalten, "especially negative, arousing and suspenseful ones" (Hsu, Conrad, Jacobs 2014; 1359) festgestellt. Daher interessieren wir uns zunächst dafür, ob die gemessenen Indikatoren für Erregung, Pupillengröße und Hautleitwert (Bradley et al. 2008) vergleichbare Ergebnisse wie andere Studien zur Fiction-Feeling-Hypothese aufzeigen, z. B. dass als "furchteinflößend" markierte Passagen stärkere Reaktionen hervorrufen als diejenigen, die das Label "fröhlich" oder "neutral" tragen (Hsu, Conrad, Jacobs 2014, Eekhof et al. 2021). Diese Reaktion würde in der Klassifikation der Leseemotionen von Miall und Kuiken (2002) in den Bereich der "narrative feelings" fallen, also Gefühlen, die gegenüber literarischen Figuren entwickelt werden bzw. auf eine Resonanz mit der Stimmung und dem Schauplatz eines literarischen Textes hindeuten. Von dieser Art der Emotionen erwartet man, dass sie den Emotionsgehalt des Textes "spiegeln" (Miall, Kuiken, 2002; 224). Dies ist das erste Experiment in einer Reihe von geplanten Studien am LitLab der TU Darmstadt, die als Ziel die Erforschung des Zusammenhangs zwischen Textsentiment und empirischen Untersuchungen von Leseprozessen haben.

Insgesamt haben im Rahmen des aktuellen Experiments 40 deutsche Muttersprachler:innen 150 Textpassagen gelesen (120 Originale, 15 Fanfictions und 15 Badfictions). Anschließend füllten die Teilnehmer:innen einen Fragebogen aus. Entgegen der Vorläuferstudie, wurden keine Fragen zur Immersion gestellt: Die Stimuli waren recht kurz (40-50 Wörter) und wurden in einer zufälligen Reihenfolge präsentiert, was die Immersion behindern würde. Wir erwarteten, dass andere Faktoren das Leseverhalten beeinflussen würden und befragten die 40 Teilnehmer auf drei verschiedene Arten zu ihren Lesegewohnheiten. Da bereits in Experimenten zur Verbindung zwischen Lesen und Theory of Mind die Unterteilung in Gruppen nach Leseerfahrung signifikante Unterschiede aufgedeckt hat (Kidd und Castano 2013, Panero et al., 2016), wollten wir den Einfluss auch bei der emotionalen Reaktion überprüfen.

Zunächst wurden Lesegewohnheiten der Teilnehmer:innen mit Hilfe des Reading Habits Questionnaire (Kuijpers et al., 2020) erfasst. Der Fragebogen nimmt die selbst angegebene Lesemenge im Laufe des letzten Jahres auf. Obwohl der Fragebogen vielseitig in seinen Auswertungsmöglichkeiten ist, kann er als Selbstauskunft

nur bedingt als zuverlässig eingestuft werden. Intensive Lesephasen und Mehrfachnennungen sind hierbei nur schwer erfassbar. Für die Auswertung wurden Angaben über unterschiedliche Genres summiert und drei fast gleich große Gruppen gebildet: die Teilnehmenden wurden in Vielleser:innen (13 Teilnehmer, Summe der Punktzahlen: 24-80), Durchschnittsleser:innen (13 Teilnehmer, Summe der Punktzahlen: 13-22) und Selten-Leser:innen (14 Teilnehmer, Summe der Punktzahlen: 4-11) eingeteilt.

Außerdem haben Proband:innen die deutsche Version des Tests zur Autorenerkennung ausgefüllt (Grolig et al. 2020), ein bewährte Methode um die Kenntnis des Literatursystems oder die langfristige Auseinandersetzung mit Literatur zu erfassen (Panero et al. 2016; Stanovich et al. 1989). Auch hier teilten wir die Proband:innen in drei Gruppen auf: Literaturkenner:innen (14, erkannten 13-38 Autoren richtig), Literatureinsteiger:innen (12, erkannten 1-6 Autoren richtig) und Mittelfeld (14, erkannten 7-12 Autoren richtig). Wie erwartet, gab es einige Überschneidungen zwischen den Gruppierungen, doch die Rangkorrelation zwischen den beiden Angaben war schwach ($\tau = 0,21$).

Der dritte Bereich, den wir in Hsus Originalexperiment als nicht ausreichend untersucht betrachteten, war die Einbeziehung des Fandom-Wissens: es wird lediglich erwähnt, dass alle Proband:innen mindestens ein Buch aus der „Harry Potter“-Reihe gelesen haben. Der Fragebogen der aktuellen Studie enthielt zwei Fragen zum Wissen über das Harry-Potter-Fandom, differenziert nach Filmen und Büchern. Wir erwarteten, dass Fans stärker auf die präsentierten Stimuli reagieren würden. Ein Wert von 0 stand für Teilnehmer, die keines der Bücher gelesen und keinen der Filme gesehen haben, während 5 bedeuten würde, dass alle Filme, alle Bücher und zusätzliches Material gelesen wurden. Insgesamt wurden Teilnehmer mit einer Punktzahl von 4 und 5 der Gruppe "Fans" (17) zugeordnet, Teilnehmer mit einer Punktzahl von 0 und 1 galten als "Nicht-Fans" (10) und Teilnehmer mit einer Punktzahl von 2 und 3 wurden dem "Fandom-Mittelfeld" (13) zugerechnet. Die Rangkorrelation zwischen dem Context Score und den beiden anderen Gruppeneinteilungen ist ebenfalls schwach ($\tau = 0,26$ mit den Autorenerkennungsergebnissen; $\tau = 0,35$ mit der selbstberichteten Lesehäufigkeit).

Datenanalyse

Die Hautleitwertdaten wurden mit Brainvision Recorder aufgenommen und mit Hilfe von Ledalab (Benedek 2010) analysiert und exportiert. Ledalab ist eine Software, die die Segmentierung von Hautleitwertdaten sowie eine automatische Ermittlung von Hautleitwertreaktionen durchführt. Dafür werden zwei unterschiedlichen Methoden verwendet: die TTP-Analyse (trough-to-peak), die auf vorgegebenen zeitlichen Kriterien basiert (Bousséin et al. 2012), und die CDA (continuous decomposition analysis), die das Signal zunächst in seine kontinuierlichen (tonischen) und stimulusbezogenen (phasischen) Komponenten unterteilt und dann mit Hilfe eines Algorithmus die Reaktionen identifiziert (für eine genaue Beschreibung und Auswertung der Methoden s. Kuhn et al. 2022). Anschließend wird eine Reihe von Statistiken ausgegeben: die Summe der Amplituden der signifikanten

Hautleitwertreaktionen, der durchschnittliche stimulusbezogene Hautleitwert und der durchschnittliche Hautleitwert inklusive der tonischen Komponente.

Die Pupillometriedaten wurden aus der Eye-tracking-Software SR Research Data Viewer exportiert und normalisiert: für jede Testperson wurde eine Baseline der Pupillengröße ermittelt, die auf der durchschnittlichen Pupillengröße basiert, die zwischen den Trials aufgenommen wurde. Um eine mittlere Veränderung der Pupillengröße zu bestimmen, wurde von der mittleren Pupillengröße pro Trial die Baseline subtrahiert.

Bei der Datenverarbeitung der Hautleitwertreaktion und der Pupillometrie-Daten wurde besonders darauf geachtet, ob die Daten die Voraussetzungen für einen ANOVA-Test erfüllten: Unabhängigkeit (die durch das Experimentdesign gegeben war), Normalverteilung und Homogenität der Varianz. Es zeigte sich, dass die Pupillengröße und die Daten zur Anzahl der Hautleitwertpeaks normalverteilt waren, die anderen Hautleitwertdaten jedoch nicht.

Die normalverteilten Daten wurden mit einem ANOVA-Test untersucht, während für die nicht normal verteilten Werte der Kruskal-Wallis-Test durchgeführt wurde. Nach den Berechnungen mit ANOVA und dem Kruskal-Wallis-Test wurde eine Auswertung der Effekte mit Epsilon-Quadrat durchgeführt, die zeigte, dass die signifikanten Ergebnisse starke Effekte aufweisen und die meisten Ergebnisse, die sich der Signifikanz näherten, mittlere Effekte zeigten.

Ergebnisse

Unsere Ergebnissen zufolge gab es keinen signifikanten Einfluss von Textsentiment auf die Anzahl oder Stärke der Reaktionen, weder beim Hautleitwert noch bei der Pupillengröße. Doch wir konnten einen weiteren signifikanten Faktor ausfindig machen, der bei der Erforschung der Leser:innenreaktionen eine Rolle spielt. Sobald Proband:innen in Gruppen nach den Ergebnissen des Author Recognition Tests eingeteilt wurden, konnte man sehen, dass Literaturkenner:innen signifikant stärkere Reaktionen im Bereich des Hautleitwertes gezeigt haben im Vergleich zu Literatureinsteiger:innen und den Mittelfeld-Proband:innen.

Tabelle 2: p-Werte des Kruskal-Wallis-Tests bei der Unterteilung in Gruppen nach den Lesegewohnheiten.

Kruskal-Wallis test	Author Recognition Test	Fandom	Reading frequency
Summe der Reaktionsamplituden (CDA)	0.08622	0.7147	0.6135
Summe der Reaktionsamplituden bei fröhlichen Stimuli (CDA)	0.07356	0.8639	0.7932
Summe der Reaktionsamplituden bei furchteinflößenden Stimuli (CDA)	0.1113	0.6392	0.4509
Summe der Reaktionsamplituden bei neutralen Stimuli (CDA)	0.1542	0.7298	0.5716
Durchschnittlicher stimulusbezogener Hautleitwert (CDA)	0.1682	0.7588	0.6101
Durchschnittlicher stimulusbezogener Hautleitwert bei fröhlichen Stimuli (CDA)	0.1396	0.8018	0.8356
Durchschnittlicher stimulusbezogener Hautleitwert bei furchteinflößenden Stimuli (CDA)	0.1682	0.7643	0.4392
Durchschnittlicher stimulusbezogener Hautleitwert bei neutralen Stimuli (CDA)	0.3272	0.7298	0.5087
Summe der Reaktionsamplituden (TTP)	0.06908	0.4902	0.7619
Summe der Reaktionsamplituden bei fröhlichen Stimuli (TTP)	0.03789	0.749	0.8883
Summe der Reaktionsamplituden bei furchteinflößenden Stimuli (TTP)	0.0675	0.3206	0.6087
Summe der Reaktionsamplituden bei neutralen Stimuli (TTP)	0.1279	0.4754	0.7932
Durchschnittlicher Hautleitwert	0.7751	0.1183	0.9248
Durchschnittlicher Hautleitwert bei fröhlichen Stimuli	0.7451	0.01945	0.7118
Durchschnittlicher Hautleitwert bei furchteinflößenden Stimuli	0.8082	0.2041	0.4508
Durchschnittlicher Hautleitwert bei neutralen Stimuli	0.9802	0.3553	0.6587

Tabelle 2 zeigt die p-Werte des Kruskal-Wallis-Tests. Bei den fettgedruckten Werten wurde der Effekt der Gruppeneinteilung als "mittel" eingestuft, während bei unterstrichenen Werten der Effekt als "stark" bewertet wurde. Wir sehen, dass die durch den Autorenerkennungstest gebildeten Gruppen bei den meisten Werten signifikante Unterschiede in ihren Mittelwerten aufweisen.

Tabelle 3: p-Werte der ANOVA bei der Unterteilung in Gruppen nach den Lesegewohnheiten.

ANOVA test	Author Recognition Test	Fandom	Reading Frequency
Anzahl der signifikanten Hautleitwertreaktionen (CDA)	0.1973	0.7381	0.5607
Anzahl der signifikanten Hautleitwertreaktionen in fröhlichen Stimuli (CDA)	0.2014	0.7028	0.6636
Anzahl der signifikanten Hautleitwertreaktionen in furchteinflößenden Stimuli (CDA)	0.1529	0.6464	0.4287
Anzahl der signifikanten Hautleitwertreaktionen in neutralen Stimuli (TTP)	0.2689	0.8548	0.6125
Anzahl der signifikanten Hautleitwertreaktionen (TTP)	0.05916	0.9627	0.4471
Anzahl der signifikanten Hautleitwertreaktionen in fröhlichen Stimuli (TTP)	0.04836	0.8797	0.457
Anzahl der signifikanten Hautleitwertreaktionen in furchteinflößenden Stimuli (TTP)	0.04812	0.9499	0.3572
Anzahl der signifikanten Hautleitwertreaktionen in neutralen Stimuli (TTP)	0.09878	0.945	0.5045

Tabelle 3 zeigt, dass eine ähnliche Tendenz in den normalverteilten Variablen auffindbar ist: während die Gruppen, die auf der Basis des Fandomscores und des Reading Habit Questionnaire gebildet wurden, keine signifikanten Unterschiede aufzeigen, zeigen die Gruppen der Autorenerkennungstests mittlere bis starke Effekte.

Meistens manifestieren sich die signifikanten Effekte in Unterschieden zwischen dem Verhalten der Literatur-

kenner:innen auf der einen und Literatureinsteiger:innen und dem Mittelfeld auf der anderen Seite, wie in Abbildung 1. Es scheint, als würde die Kenntnis des literarischen Feldes Voraussetzung für häufigere Hautleitwertreaktionen sein.

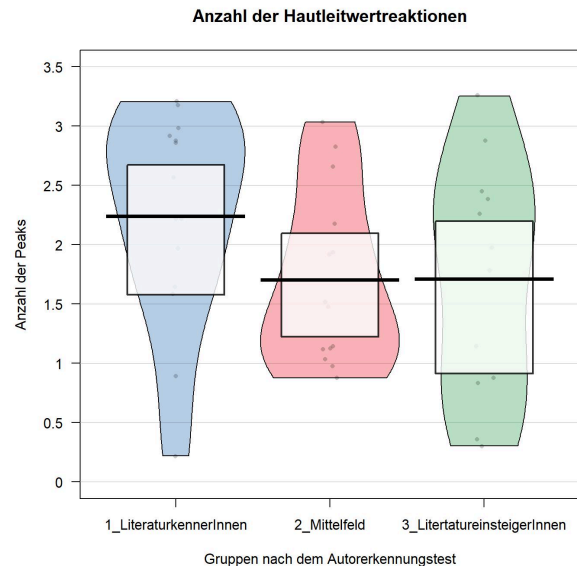


Abbildung 1. Durchschnittliche Anzahl der Hautleitwertreaktionen bei Literaturkenner:innen, Mittelfeld-Proband:innen und Literaturkenner:innen.

Wird eine Unterteilung nach der Leistung im Autorenerkennungstest vorgenommen, so zeigt sich, dass Literaturkenner:innen bei den meisten Werten signifikant größere Reaktionen zeigten: Es gibt eine höhere Anzahl von Peaks, die Summe der Amplituden ist höher und die durchschnittliche Pupillengröße ist größer. Bei den Literaturkenner:innen sind die Reaktionen auf fröhliche Stimuli am höchsten und auf neutrale Stimuli fast immer am niedrigsten. Die Literatureinsteiger:innen hingegen zeigen meist minimale Werte bei fröhlichen Stimuli und am häufigsten höchste Werte bei neutralen Stimuli. Keiner dieser Werte wich signifikant vom Mittelwert ab.

Diskussion

Unsere Ergebnisse zeigen, dass die nach unterschiedlichem Sentiment gelabelten Texte keine signifikanten Unterschiede in der Hautleitwertreaktion und in der Pupillengröße aufzeigen. Die Analysen der Lesegewohnheiten der Proband:innen lassen hingegen darauf schließen, dass diejenigen, die mehr lesen, auch stärkere Reaktionen auf Texte insgesamt aufweisen.

Vor allem die Kenntnis des Literatursystems – wie der Autorenerkennungstest oft interpretiert wird – beeinflusst die körperlichen Reaktionen auf das Lesen in erheblichem Maße. Leser:innen, die mehr Erfahrung mit Literatur haben, reagieren stärker auf literarische Werke und spiegeln dabei den Textsentiment wieder, die ein Text enthält (stärkere Reaktionen auf emotionale Inhalte, schwächere auf neutrale Passagen).

Leser mit geringerem Wissen über Literatur scheinen auch auf neutrale Stimuli stark zu reagieren, vielleicht weil sie einen emotionalen Stimulus erwarten und diesen nicht erhalten. Diesen Ergebnissen zufolge ist das Wissen über Literatur für eine andere Art der Reaktion auf Texte verantwortlich.

Entgegen unseren Erwartungen zeigte sich kein signifikanter Einfluss von höherer Kenntnis des Werks, obwohl der Gesamtmittelwert des Hautleitwerts bei fröhlichen Stimuli bei Fans höher war. Möglicherweise ist dies auf eine Kombination aus Nostalgie und narrativen Gefühlen zurückzuführen.

Schließlich hat die jüngste Leseaktivität, die mit dem RHQ ermittelt wurde, fast keinen Einfluss auf die physiologischen Reaktionen - nur als zusätzlicher Faktor bei der Berücksichtigung der Pupillengröße. Diese korrelativen Zusammenhänge bieten allerdings noch keine Antwort auf die Frage nach der Kausalität - die Frage, ob Lektüre die emotionale Reaktion trainiert oder ob empfindsame Menschen sich mehr zu Literatur hingezogen fühlen, bleibt offen.

Die Aussagekraft der Ergebnisse ist durch einige Schwachstellen eingeschränkt: beispielsweise sind die Proband:innen überwiegend Studierende und können daher nur schwer als absolute Wenigleser:innen bezeichnet werden. Vielleicht ist das der Grund, warum die Daten so selten Unterschiede zwischen Literatureinsteiger:innen und Mittelfeld-Proband:innen aufzeigen.

Darüber hinaus hat die Anzahl der Proband:innen eine eher geringe statistische Aussagekraft (40 Teilnehmer), wovon allerdings nur die Analyse der Proband:innengruppen betroffen ist: Für die Analyse des Sentimenteinflusses auf die Leser:innenreaktion wird die statistische Signifikanz durch die große Anzahl an Trials derselben Sentimentklasse wieder angehoben. Die Ergebnisse dienen zum Anlass, über mehrere Studien hinweg den Einfluss der Lesekompetenz zu berücksichtigen.

Zuletzt wären Vergleichsstudien mit anderen literarischen Gegenständen interessant, um die Zusammenhänge der hier vorgestellten Variablen über „Harry Potter“ hinaus zu beobachten und weitere Aspekte von literarischen Texten wie Stil und Epoche ebenfalls in ihrer Wirkung zu untersuchen.

Bibliographie

Benedek, Mathias, und Christian Kaernbach. 2010. "A Continuous Measure of Phasic Electrodermal Activity." *Journal of Neuroscience Methods* 190, Nr. 1: 80–91. <https://doi.org/10.1016/j.jneumeth.2010.04.028>.

Boucsein Wolfram, Don C. Fowles, Sverre Grimnes, Gershon Ben-Shakhar, Walton T. Roth, Michael E. Dawson, Diane L. Fillion. 2012. "Publication recommendations for electrodermal measurements." *Psychophysiology* 49, Nr. 8: 1017–34. <https://doi.org/10.1111/j.1469-8986.2012.01384>.

Bradley, Margaret M., Laura Miccoli, Miguel A. Escrig, und Peter J. Lang. 2008. "The Pupil as a Measure of Emotional Arousal and Autonomic Activation." *Psychophysiology* 45, Nr. 4: 602–7. <https://doi.org/10.1111/j.1469-8986.2008.00654.x>.

Eekhof, Lynn S., Kobie van Krieken, José Sanders, und Roel M. Willems. 2021. "Reading Minds, Reading Stories:

Social-Cognitive Abilities Affect the Linguistic Processing of Narrative Viewpoint." *Frontiers in Psychology* 12: 698986. <https://doi.org/10.3389/fpsyg.2021.698986>.

Grolig, Lorenz, Simon P. Tiffin-Richards, und Sascha Schroeder. 2020. "Print Exposure across the Reading Life Span." *Reading and Writing* 33, Nr. 6: 1423–41. <https://doi.org/10.1007/s11145-019-10014-3>.

Hsu, Chun-Ting. 2015. "Textual Emotion Potential, Fiction Feelings, and Immersion: An fMRI Study Testing the Neurocognitive Poetics Model of Literary Reading." FU Berlin, <http://dx.doi.org/10.17169/refubium-14086>.

Hsu, Chun-Ting, Markus Conrad, und Arthur M. Jacobs. 2014. "Fiction Feelings in Harry Potter: Hemodynamic Response in the Mid-Cingulate Cortex Correlates with Immersive Reading Experience." *NeuroReport* 25, Nr. 17: 1356–61. <https://doi.org/10.1097/WNR.0000000000000272>.

Hsu, Chun-Ting, Arthur M. Jacobs, Ulrike Altmann, und Markus Conrad. 2015. "The Magical Activation of Left Amygdala When Reading Harry Potter: An fMRI Study on How Descriptions of Supra-Natural Events Entertain and Enchant." *PLOS ONE* 10, Nr. 2: e0118179. <https://doi.org/10.1371/journal.pone.0118179>.

Hsu, Chun-Ting, Arthur M. Jacobs, und Markus Conrad. 2015. "Can Harry Potter Still Put a Spell on Us in a Second Language? An fMRI Study on Reading Emotion-Laden Literature in Late Bilinguals." *Cortex* 63: 282–95. <https://doi.org/10.1016/j.cortex.2014.09.002>.

Hsu, Chun-Ting, Arthur M. Jacobs, Francesca M.M. Citron, und Markus Conrad. 2015. "The Emotion Potential of Words and Passages in Reading Harry Potter – An fMRI Study." *Brain and Language* 142: 96–114. <https://doi.org/10.1016/j.bandl.2015.01.011>.

Hsu, Chun-Ting, Markus Conrad, und Arthur M. Jacobs. 2014. "Fiction Feelings in Harry Potter: Hemodynamic Response in the Mid-Cingulate Cortex Correlates with Immersive Reading Experience." *NeuroReport* 25, Nr. 17: 1356–61. <https://doi.org/10.1097/WNR.0000000000000272>.

Kavanagh, Ciarán. 2021. "Refiguring Reader-Response: Experience and Interpretation in J.G. Ballard's Crash." In *Powerful Prose*, hg. von R. L. Victoria Pöhls und Mariane Utudji. Bielefeld, Germany: transcript Verlag, 77–96. <https://doi.org/10.14361/9783839458808-006>.

Kidd, David Comer, und Emanuele Castano. 2013. "Reading Literary Fiction Improves Theory of Mind." *Science* 342, Nr. 6156: 377–80. <https://doi.org/10.1126/science.1239918>.

Kuijpers, Moniek, Shawn Douglas, und Don Kuiken. 2020. "Capturing the Ways We Read." *Anglistik* 31, Nr. 1: 53–69. <https://doi.org/10.33675/ANGL/2020/1/6>.

Kuhn, Manuel, Anna M. V. Gerlicher, und Tina B. Lonsdorf. 2022. "Navigating the Manyverse of Skin Conductance Response Quantification Approaches – A Direct Comparison of Trough-to-Peak, Baseline Correction, and Model-based Approaches in Ledalab and PSPM." *Psychophysiology* 59, Nr. 9. <https://doi.org/10.1111/psyp.14058>.

Miall, David S., und Don Kuiken. 2002. "A Feeling for Fiction: Becoming What We Behold." *Poetics* 30, Nr. 4: 221–41. [https://doi.org/10.1016/S0304-422X\(02\)00011-6](https://doi.org/10.1016/S0304-422X(02)00011-6).

Panero, Maria Eugenia, Deena Skolnick Weisberg, Jessica Black, Thalia R. Goldstein, Jennifer L. Barnes, Hiram Brownell, und Ellen Winner. 2016. "Does Reading a

Single Passage of Literary Fiction Really Improve Theory of Mind? An Attempt at Replication. " *Journal of Personality and Social Psychology* 111, Nr. 5: e46–54. <https://doi.org/10.1037/pspa0000064>.

Raffaelli, Quentin, Caitlin Mills, und Kalina Christoff. 2018. "The Knowns and Unknowns of Boredom: A Review of the Literature." *Experimental Brain Research* 236, Nr. 9: 2451–62. <https://doi.org/10.1007/s00221-017-4922-7>.

Stanovich, Keith E., und Richard F. West. 1989. "Exposure to Print and Orthographic Processing." *Reading Research Quarterly* 24, Nr. 4: 402. <https://doi.org/10.2307/747605>.

FakeNarratives – First Forays in Understanding Narratives of Disinformation in Public and Alternative News Videos

Tseng, Chiao-I

tseng@uni-bremen.de
Universität Bremen

Liebl, Bernhard

liebl@informatik.uni-leipzig.de
Universität Leipzig

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de
Universität Leipzig

Bateman, John

bateman@uni-bremen.de
Universität Bremen

Introduction: Narratives of Disinformation

Audiovisual media, such as film, TV, webseries, YouTube-videos and so on, have long become one, if not the, dominant cultural communication form of our times (Mirzoeff, 1999; Mitchell, 2001). Their role in shaping sociocultural configurations of all kinds appears incontestable. However, description and analysis of such media presents immense challenges which have so far resisted scalable solutions. Although it is becoming increasin-

gly possible to conduct larger-scale stylistic and formal feature analysis (shot lengths, brightness, color profiles, dynamicity, etc.: cf., e.g., Heftberger, 2018), productive engagement with such media as powerful sociocultural artifacts demands in addition hermeneutic and functional interpretations as pursued in many branches of the humanities. Analyses of these kinds have not so far been possible within the digital humanities.

In an ongoing BMBF project on FakeNarratives¹, we combine expertise, experience and tools crucial for making general advances towards these goals, practically directed at a soundly delimited and, at the same time, socially significant class of audiovisual artifacts: TV news. We address the research hypothesis that news of this kind increasingly employs strategies of *audiovisual narrative* that may undercut, undermine or construct ideological positions and evaluations of news content *independently* of what may be simply stated, for example, in accompanying spoken text. The increasing sophistication and technical possibilities available to news channels makes it possible to draw on techniques for storytelling well established in film, but this can by no means always be guaranteed to work as intended. Indeed, filmic storytelling techniques may also serve to more effectively dis-inform.

Although consideration of news reporting as narrative already has a considerably history and is nowadays hardly controversial (Sperry, 1981; Bell, 1999; Liebes, 1994; Langer, 1998; Hickethier, 2000; Dunn, 2005; Machill et al., 2006), methods for engaging specifically with their *audiovisual* non-verbal properties remain limited. Analyses addressing the audiovisual are overwhelmingly based on small datasets and, as is usual within the humanities, are performed interpretatively. Consequently, just how widespread 'narrativization' strategies are, and to what effect and aims they are being applied, urgently demands in depth and larger-scale research from within the digital humanities.

Case study: Tagesschau vs. Bild TV

The current paper presents some early results of the FakeNarratives project illustrating how narrative audiovisual strategies might begin to be addressed at variable scales. Several distinct kinds of narrative strategies have been discussed in the journalism and newsreporting literature, but for the purposes of our first case study we focus on the production of 'individual-centered narrative' by means of editing devices such as showing layperson's talk, individualized events and actions, close-ups of individual faces, and emotion shown either visually or exhibited in accompanying language. These features appear to be widely used across alternative and public news outlets (Vettehen et al., 2008). Our research hypothesis is that both kinds of news outlets use narrative strategies to some degree, but that there may still be significant differences in why and how often they are actually used. We consider it highly implausible that the use of narrative strategies can be reduced to any binary "good vs. bad" or "informing vs. dis-informing" characterisation and so more finely discriminating accounts are in any case ne-

cessary. The basic challenge we address here is then to set out how hypotheses concerning narrative construction may be made amenable to larger-scale data analysis.

Here we select some of the most commonly discussed narrative features in news that contribute to highlighting individual stories, a technique that can effectively increase the viewers' memories of news contents and their perception of news severity (Aust & Zillmann, 1996; Zillmann & Brosius, 2012). We examine whether and how these features are presented in both public and online alternative news channels. In particular, we consider how these features are employed to individualise, personalise or emotionalise people shown in news videos. In other words, we seek to operationalise just what constitutes the narrative strategy of *individualisation* so that we can explore how the strategy is used across our selected news channels.

Data

Our data include a preliminary selection of 166 news videos, 70 videos from Tagesschau, which we use as an example for more traditional, serious media, and 96 comparable videos from Bild TV, which we use as an example for a more recent, alternative news channel. These videos were all produced between 1 January 2022 and 15 March 2022. The length of entire news programs in both channels is around 15-25 minutes. For current purposes, we analyze the first news report of each news video, namely, the top news story of the entire program, which usually lasts from 3 to 5 minutes. The most reported top news themes from January to the middle of March in 2022 were Covid-19 and the outbreak of the Ukraine war.

Method

For systematically analyzing the news videos, we employed the multimodal discourse approach of film cohesion (Tseng, 2013), which operates by picking out how the deployment of visual image, sound, verbal language, written language, camera movement, framing, color, and many more pattern together so as to introduce and coherently track people, places, objects and events within an unfolding event sequence. In other work, we have found that the cohesion structures resulting from such analyses appear to play significant roles for scaffolding an audience's perception and understanding of narratives (cf., e.g., Tseng et al., 2021).

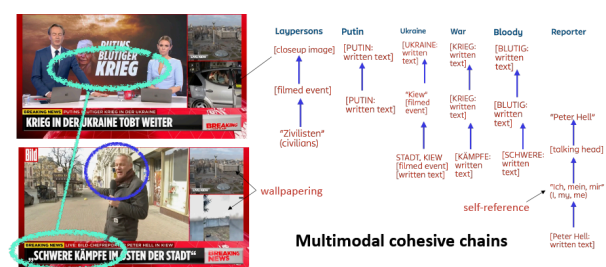
Five distinct types of people commonly depicted in news reported were singled out for attention: (1) anchor persons, (2) reporters, (3) news commentators, (4) protagonists in the reported news events and (5) laypersons. The first three traditionally represent the authoritative voice in news and so we track and analyze whether these voices are personalized by including their own attitudes and emotions. The fourth type, protagonists, includes the main involved characters of news events, whereas the fifth type, laypersons, refers to people represented as generic exemplars effected by the news event.

1. *Layperson in [talking head] mode*: whether a news video employs layperson's interviews.
2. *Specific protagonist*: whether a particular individual in the news report is cohesively identified as a specific protagonist.
3. *Putin*: whether the Russian president Putin is depicted as protagonist co-patterned with particular personalized quality and emotional features.
4. *Close-up of laypersons and protagonists*: whether laypersons or protagonists are additionally shown with close shots.
5. *Self references of anchors, reporters or commentators*: whether the 'authoritative' voices make recurring uses of self-references, such as the uses of "I", "my", "me" suggesting subjective opinions or personal attitudes of the journalists.

In addition to these five multimodal categories derived directly from the cohesive chains concerning individuals, we also analyze how these individuals are combined with the following four narratively relevant features:

1. *Conflict or violent events*: whether any conflict-related event is being mentioned, seen or written in news reports.
2. *Negative evaluation*: whether particular, recurring quality features with negative evaluation can be found in news reports, for instance, the negative descriptions about the war or politicians.
3. *Emotions*: whether there are multimodal realization of emotions, for instance, emotional terms mentioned in spoken or written texts and emotional reactions shown in visual images or sounds.
4. *Wallpapering*: whether, in addition to the individuals and their evaluative features derived from the analysis of cohesive chains, background images that are not strictly related to the contents of the news report running in the foreground are shown.

The process of deriving the applicability of these categories for a video segment on the basis of the analysis of cohesive chains is depicted graphically in Figure 1.



Illustrative example of cohesive chains (right) developed on the basis of the unfolding video segment (Bild TV news, 25.02.2022, left). The chains show how the identities and elements tracked may occur in any modality and that their combination provides direct support for the narratively relevant features being targeted.

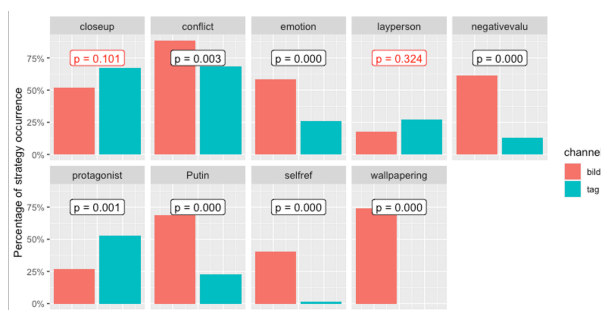
Initially, we coded and analyzed these nine categories in 166 video segments manually. Each segment was coded as '1' when the category applies and '0' when it does not apply. For simple coding of single news stories, the analyses could be entered into spreadsheets; for more complex coding, either of further narrative strategies or inclu-

ding more of the videos beyond the first news story, more structured annotations are essential, typically performed currently using tools such as ELAN (ELAN, 2021).

Results and discussion

The occurrences of the nine categories in our two selected TV channels are displayed in Figure 2. This shows the complex similarities and differences of the two channels when employing the distinct fine-grained strategies contributing to individualization. From the differing distributions among the fine-grained strategies, it becomes clear that, although both channels might be said to be employing individualization as a technique for raising the interest-value of their reports, they do this in differing ways. For example, the occurrences of closeup faces and the talk of laypersons show no significant difference across the two channel (Fisher's exact test, $p=0.101$ and $p=0.324$ respectively); thus both channels can be seen to use the individualizing feature of closeups and laypeople reports. However, the comparison also shows that Tagesschau uses significantly more *protagonists* to report news events than does Bild TV (Fisher's exact test, $p=0.001$). This means that Tagesschau actually individualizes news stories more than Bild TV does, but in a very specific way. Nevertheless, the result in the category of *Putin* indicates that Bild TV does specifically individualize Putin significantly more than Tagesschau. Indeed, in our data, Bild TV largely labels the Ukraine war or the violent conflict as Putin's war, while Tagesschau mostly refers the war to the 'invasion of Russia' or Russian soldiers. Figure 2 also shows that the categories of conflict events, emotional reaction, negative evaluation and wallpapering all pattern similarly with significantly more use in Bild TV than in Tagesschau. Moreover, the substantial use of self references of reporters in Bild TV indicates the common insertion of personalized attitudes and opinions into their news reports.

In summary, Tagesschau minimizes Putin's personification, negative evaluation, decorative editing and reporters' self-references, presumably in order to balance its heavy use of protagonist-centered narrative model with objective representation and visualization, while the Bild TV news videos employ evaluative and decorative features to dramatize the news events.



The percentage occurrences of the nine categories in Tagesschau and Bild TV and corresponding p-values indicating the significance of difference (Fisher's exact test).

Apart from the narrative strategy of *individualisation*, we are currently analysing and formalising other crucial news narrative strategies such as *dramatisation* and *fragmentation* (Bennett, 1988). Nevertheless, for enlarging the corpus of news videos and the scale of annotation categories, we require more effective annotation tools beyond the time-intensive manual annotation methods such as ELAN. In this pursuit, automatic annotation is one central goal of the FakeNarratives project.

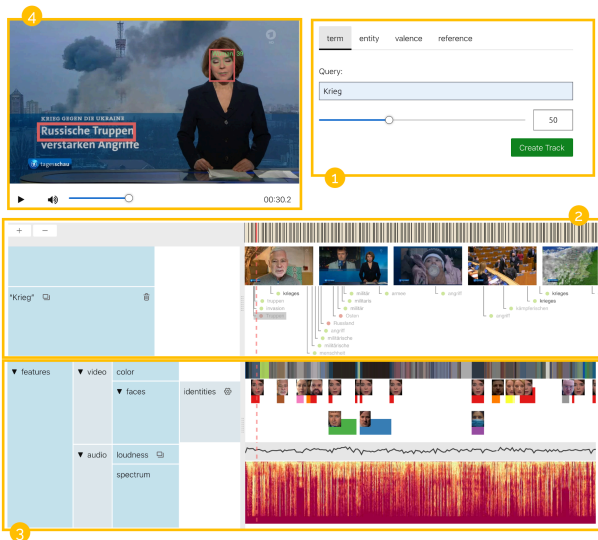
What's ahead: Automatic annotation and exploratory visualization tool

The results of the previous section are encouraging concerning the use of more abstract audiovisual analyses as a means for revealing differences and similarities between audiovisual materials in ways that naturally relate to issues of narrativization. The characterizations of our selected news media clearly indicate where differences in broader strategies are occurring. It is also nevertheless clear that for more reliable and extensive results, expanding both the kinds of audiovisual materials analyses and the kinds of strategies at issue, it will be essential to move beyond time-intensive manual annotation. Augmenting this work further by the development and application of automatic and semi-automatic annotation is consequently a further central goal of the FakeNarratives project, for which we adopt a two-pronged approach combining foci on visualization and automated analysis.

We now combine a range of computational techniques for which there are already good solutions available, such as shot detection, face detection, etc., into a flexible tool named *Zoetrope*, supporting the progressive annotation of larger data sets by means of a more interactive interface (see Figure 3). *Zoetrope* enhances the existing landscape of tools for audiovisual annotation and analysis (for an overview see Pustu-Iren et al., 2020), as it is not only capable of visualizing automatic annotations, but also provides some basic functionalities for querying the video, for instance for specific key words or named entities. By these means, annotation becomes an activity that can leverage both automatic results of audiovisual processing and more abstract characterizations of data in terms such as cohesive structures as used above.

The current *Zoetrope 1.0* prototype already allows researchers to search for keywords (*segment 1*). Any keyword can be used as a query, which will then be searched in the spoken (Mozilla DeepSpeech² with a German model by Agarwal and Zesch³) and written (scene detection framework *easyOCR*⁴) language of the video. We also embed the query using word embeddings⁵ so that we can find words semantically related to the query as well. Via a slider, the similarity threshold can be set looser or more exact (=100%). The results of such a query are visualized in *segment 2*, which basically uses a timeline metaphor. Query results found in spoken language are rendered in green, results in written language are rendered in red. Note that for the query "Krieg" we also find semantically related concepts such as "Truppen", "Invasion" or "Mili-

tär", due to the semantic similarity threshold, which was set here rather low, at 50%. In segment 3, we see some exemplary automatically determined features, including face detection and audio analysis by means of a spectrogram. Finally, segment 4 shows an interactive video player, which can be navigated by means of the timeline or by clicking on specific results, such as a keyword or a face. Visible features, such as written keywords or faces, can also be rendered with their bounding box withing the video.



Zoetrope 1.0 prototype for visualizing and exploring news videos.

The tools developed within the project will continue to pursue advanced visualization and more extensive automatic analysis. The latter goal is supported by our project partners in Hanover and their continuing development of the web-based audiovisual analytics platform TIBAVA (TIB AV- Analytics).

Conclusions and Outlook

We have shown in this paper some of the benefits that may accrue when the analysis of audiovisual media can build on more abstract discourse patterns carried out at scale. We have argued that to take this further, however, a progressive increase in computational support is necessary. To this end, we have set out some first steps taken towards bridging some of the gaps between current technical feature processing possibilities and discourse-related characterizations of data. This is being enabled first and foremost by an interactive tool which already seamlessly incorporates a variety of state of the art processing techniques relevant for more abstract annotation practices. We believe such a tool will greatly speed up the manual annotation that our current corpus analysis relies on and will allow us to discover and annotate many more narrative strategies in the future. Once the tool is beyond its prototype stage, we also plan to release it for the DH community, which might use it to search other video material than news videos, for instance YouTube Let's Play videos for game studies, or any mo-

vie or TV series for film studies scenarios. A live demo of the prototype will be included in the presentation at DHd 2023.

Fußnoten

1. <https://fakenarratives.github.io/index>
2. <https://github.com/mozilla/DeepSpeech>
3. <https://github.com/AASHISHAG/deepspeech-german>
4. <https://github.com/JaidedAI/EasyOCR>
5. Using spaCy's `de_core_news_md` model

Bibliographie

- Aust, Charles & Dolf Zillmann. 1996. "Effects of Victim Exemplification in Television News on Viewer Perception of Social Issues". *Journal of Mass Media and Communication Quarterly* 4(73), 787-803.
- Benett, Lance W. 1988. "News: The Politics of Illusions". New York: Longman.
- Bell, Allan. 1999. "News stories as narratives." In A. Jaworski and N. Coupland (Eds.), *The Discourse Reader*, Chapter 13, 236-251. London: Routledge.
- Dunn, Anne. 2005. "Television news as narrative." In H. Fulton, R. Huisman, J. Murphet, and A. Dunn (Eds.), *Narrative and Media*, 140-152. Cambridge, UK: Cambridge University Press.
- ELAN. 2021. "ELAN – Linguistic Annotator (Version 6.2)." computer software. Nijmegen: Max Planck Institute for Psycholinguistics. The Language Archive.
- Heftberger, Adelheit. 2018. "Digital Humanities and Film Studies: Visualising Dziga Vertov's Work." Switzerland: Springer Nature.
- Hickethier, Knut. 2000. "Fernsehnachrichten. Geschichten aus 1001 Nachricht." *Message* 2(2), 70-74.
- Langer, John. 1998. "Tabloid Television: Popular Journalism and the 'Other News'." London and New York: Routledge.
- Liebes, Tamar. 1994. "Narrativization of the News: An Introduction." *Journal of Narrative and Life History* 4(1-2), 1-8.
- Machill, Marcel, Sebastian Köhler, & Markus Waldhauser. 2006. "Narrative Fernsehnachrichten: Ein Experiment zur Innovation journalistischer Darstellungsformen." *Publizistik* 51(4), 479-497.
- Mirzoeff, Nicholas. 1999. "An introduction to visual culture." London & New York: Routledge.
- Mitchell, William. 2001. "What is visual culture?" In *Meaning in the visual arts: views from the outside*, 202-217. Princeton: Institute for Advanced Study.
- Pustu-Iren, Kader, Julian Sittel, Roman Mauer, Oksana Bulgakowa & Ralph Ewerth. 2020. "Automated Visual Content Analysis for Film Studies: Current Status and Challenges". *Digital Humanities Quarterly*, 14(4).
- Sperry, Sharon L. 1981. "Television News as Narrative." In R. P. Adler (Ed.), *Understanding Television: Essays on Television as a Social and Cultural Force*, 295-312. New York: Praeger.
- Tseng, Chiao-I. 2013. "Cohesion in Film: Tracking Film Elements." Basingstoke: Palgrave Macmillan.

Tseng, Chiao-I., Jochen Laubrock, & John A. Bateman. 2021. "The impact of multimodal cohesion on attention and interpretation in film." *Discourse, Context & Media* 44, aop.

Vettehen, Paul, Koos Nuijten, & Allerd Peeters. 2008. "Explaining effects of *sensationalism* on liking of television news stories: The role of emotional arousal." *Communication Research* 35, 319–338.

Zillmann, Dolf & Hans-Bernd Brosius. 2012. "Exemplification in Communication the influence of Case Reports on the Perception of Issues". Routledge.

Forschung, Informationswissenschaft und Archiv = drei Perspektiven auf eine Aufgabe

Grundig de Vazquez, Katja

katja.gundig.de.vazquez@uni-jena.de
Friedrich-Schiller-Universität Jena

Krefft, Annett

krefft@dipf.de
DIPF | Leibniz-Institut für Bildungsforschung und
Bildungsinformation

Thoden, Klaus

thoden@ktho.de
DIPF | Leibniz-Institut für Bildungsforschung und
Bildungsinformation

Kurzbeschreibung des Projekts

Das DFG-geförderte Projekt „Erziehung über Grenzen denken - Wilhelm Reins pädagogischer Korrespondenznachlass“¹ zielt auf eine transnationale und vergleichende Analyse vielfältiger historischer pädagogischer Kontakte und länderübergreifender pädagogisch-reformerischer Diskurse und leistet damit Grundlagenforschung mit bildungstheoretischem Schwerpunkt. Quellengrundlage ist der umfangreiche, langjährige und vielfältige internationale pädagogische Korrespondenznachlass Wilhelm Reins. Der in seiner Wirkungszeit international einflussreiche Erziehungswissenschaftler, Lehrerbildner und pädagogische Netzwerker Rein (1847–1929) hatte an der Universität Jena das erste Ordinariat für Pädagogik in Deutschland inne. Das Forschungsvorhaben wird verbunden mit der archivarisches Bearbeitung der Originalquellen, der Aufbereitung des detailliert vorliegenden Metadatenkorpus sowie der Bereitstellung des gesamten pädagogischen Briefnachlasses in edierter

ter Form als ein frei verfügbares nachnutzbares Instrument für weiterführende Forschungen.

Ausgangslage

Das Projekt² baut auf Arbeitsergebnisse aus einer Vorbereitungsphase (2016–2017) und einem Anschubprojekt (1/2018–6/2019)³ auf, in deren Verlauf das Quellenkorpus durch die Forschenden im Projekt erfasst, kategorisiert, strukturiert, systematisiert und digitalisiert wurde. Während der Vorbereitungsphase wurde der pädagogische Korrespondenznachlass aus dem Gesamtbestand des schriftlichen Nachlasses Reins herausgelöst und in (A) Briefe an Rein, (B) Briefe von Rein und (C) Kondolenzschreiben zum Ableben Reins unterteilt. Diese Dokumentmenge bildet das Quellenkorpus, das im Rahmen des vorgestellten Projekts bearbeitet und ausgewertet wird.

Im Anschubprojekt wurde die pädagogische Korrespondenz alphabetisch nach Korrespondenzpartner*innen sowie chronologisch strukturiert, erfasst und gemäß den *DFG-Praxisregeln „Digitalisierung“* (Deutsche Forschungsgemeinschaft 2016) digitalisiert. Den Korrespondent*innen wurden projekteigene Grundsignaturen, den Briefen Dokumentsignaturen zugewiesen. Die grundlegenden Metadaten der Dokumente (1) Signatur, (2) Verfasser*in/Empfänger*in, (3) Ausstellungsdatum, (4) Ausstellungsort, (5) Umfang, (6) Dokumenttyp, (7) Sprache, (8) Prüfvermerke und (9) Dateinamen der zugehörigen Images wurden im Bibliographieprogramm Zotero⁴ erfasst. Als Grundlage zur Auszeichnung und Auswertung der Quellen wurden eine Indexsystematik und ein fachsystematischer Index zur kontrollierten Indexierung entworfen. Im Juli 2019 wurde mit der Recherche bio-bibliographischer Angaben zu den Verfasser*innen begonnen, welche in der aktuellen Projektphase fortgeführt wird.⁵

In der Anschubphase wurden, basierend auf der Sichtung der Briefe und den erzeugten Metadaten, erste inhaltliche Erkenntnisse gewonnen, welche auf ein vielfältiges Erkenntnispotenzial des Bestandes verweisen und die zur Herleitung der u.g. Forschungsschwerpunkte geführt haben. So kann man z. B. aus Erkenntnissen zu Umfang, Dauer und Reichweite der Korrespondenz und zur Diversität der beteiligten Korrespondent*innen auf eine bemerkenswert heterogene Struktur des Korrespondenzgefüges schließen. Diese heterogene Struktur wird durch den Korrespondenznachlass wesentlich dokumentiert, so dass dieses Quellenkorpus vergleichende wie transzendierende Analysen (s. u.) von Diskursen und Dynamiken zwischen vielfältigen Akteur*innen über vielfältige Begrenzungen hinweg ermöglicht.

Quellenkorpus als Datengrundlage

Der pädagogische Korrespondenznachlass Reins umfasst insgesamt 6.301 Korrespondenzdokumente (Briefe, Postkarten, Telegramme) von mehr als 3.500 Korrespondent*innen aus 42 Ländern. Diese Originalquellen – überwiegend in deutscher Kurrentschrift

- bilden in einem Zeitraum von sechs Jahrzehnten (1869-1929) interne Perspektiven vielfältiger Akteur*innen ab, die in kontroverse und einflussreiche bildungspolitische Diskurse mit oftmals internationaler Reichweite eingebunden waren. Zu den Korrespondent*innen zählten Personen unterschiedlichster Hintergründe und Weltanschauungen, darunter Vertreter*innen einer pädagogisch-akademischen Elite (z. B. Nicolas Murray Butler, Adolf Damaschke, Friedrich Wilhelm Förster, Helene Lange, Paul Geheeb, Paul Natorp, Friedrich Paulsen, Eduard Spranger) und einflussreiche Personen aus Politik und Gesellschaft (z. B. Gertrud Bäumer, Houston Stewart Chamberlain, Else Fisch, Marie Fischer-Lette, Friedrich Naumann), aber auch zu einem wesentlichen Teil Personen, die aufgrund vielfältiger Faktoren (z. B. Geschlecht, sozialer, fachlicher oder akademischer Status, Religion oder geographische Herkunft) in pädagogischen Diskursen und bildungsgeschichtlich unterrepräsentiert sind. Rein zählte in mehreren Bereichen (insbesondere Lehrerbildung, Volksschulbildung, höhere Frauenbildung) als progressiver pädagogischer Reformator. Sein Expert*innennetzwerk war weitverzweigt und im zeitgenössischen Vergleich bemerkenswert heterogen und inklusiv. Der Korrespondenznachlass dokumentiert dieses in wesentlichen Auszügen und konserviert sowohl Stimmen von Akteur*innen, die ihr Wirken und ihre Expertise in der (fach)öffentlichen Wahrnehmung breiter sichtbar machen konnten, als auch von solchen Personen, die maßgeblich in die Verbreitung und Rezeption pädagogischer Theorie und Praxis eingebunden waren, aber keine nennenswerte Sichtbarkeit erlangen konnten. Das macht dieses bisher kaum erforschte Quellenkorpus aus bildungsgeschichtlicher Sicht bemerkenswert. Das Wirken unterschiedlicher Akteur*innengruppen, ihre Kontakte und Synergien, pädagogische Entwicklungen und Phänomene können vergleichend und vielfältige Abgrenzungen transzendierend (z. B. Kultur, Status, geographische und geopolitische Grenzen, Ideologien, Geschlecht) exemplarisch erforscht werden. Bei der Auswertung muss die besondere Art der Quelldokumente berücksichtigt werden. Briefe⁶ stellen besondere Anforderungen an die Auswertung, bieten aber auch Forschungspotenzial, das andere historische Quellentypen nicht aufweisen (vgl. Baillot 2020, S. 390, Budde 2020, Henzel 2020, S. 226 ff.). Sie transportieren als Lebenszeugnisse (Nutt-Koth 2016) und Ereignisse (Stadler, Illetschko und Seifert 2016), die in größere kommunikative Zusammenhänge eingebunden sind, mehr Informationsschichten als z. B. Fachpublikationen. Als solche sind Briefe sensible Dokumente, die ursprünglich nicht für eine (breitere) Öffentlichkeit bestimmt waren. Sie repräsentieren einen im zeitgenössischen Kontext auch durch das Postgeheimnis gesetzlich geschützten (vgl. Standhartinger 2020, S. 272) Kommunikationsraum, in dem Verfasser*innen persönliche Meinungen, fachliche Positionen oder auch private Belange tendenziell unverstellt darlegen können.

Zielperspektiven des Projektes

In der Grundlagenforschung bzw. Quellenauswertung folgt das Projekt der Annahme, dass sich pädagogische Reform als ein verbindendes wie mehrdeutiges Motiv auf die Verbreitung, Rezeption, Gestaltung und Entwick-

lung pädagogischer Theorie und Praxis ausgewirkt hat. Schwerpunktmäßig soll untersucht werden, (1) ob pädagogische Reform nachweislich ein wesentliches Motiv des zeitgenössischen pädagogischen Austausches war, (2) in welchen thematischen Kontexten pädagogische Reform ggf. diskutiert worden ist, (3) ob sich verschiedene bzw. welche unterschiedlichen Konnotationen von pädagogischer Reform sich identifizieren lassen, (4) ob bzw. in welchem Maße das Sprechen über pädagogische Reform einen fachlichen Austausch über Grenzen hinweg begünstigt hat und (5) ob bzw. inwiefern unterschiedliche Konnotationen oder Verwendungen des Motivs pädagogische Reform zu Störungen in der fachlichen Kommunikation geführt haben. Ergänzend dazu sollen bio-bibliographische Daten und Erkenntnisse zu Kontakt- und Netzwerkstrukturen zwischen den beteiligten Korrespondent*innen gewonnen werden. Ein besonderes Interesse gilt bildungshistorisch unterrepräsentierten Akteur*innen, ihrem Einfluss und Wirken und den Themen und Positionen, die durch sie bearbeitet und vertreten worden sind.

Als strukturelles Ziel wird im Laufe des Projekts ein übersichtliches recherchier- und zitierfähiges, digitales Korpus als nachnutzbare Grundlage für die inhaltliche Auswertung und weiterführende Forschungen erstellt. Dazu werden die Quellentexte in digitale Volltexte in TEI/XML transkribiert, indexiert und inhaltlich ausgewertet. Aufgrund des Entstehungszeitraumes der Briefe sind zwar keine Persönlichkeitsrechte der Korrespondenzpartner*innen sowie der in den Briefen Erwähnung findenden dritten Personen zu beachten, hingegen gelten noch Urheberrechte für einen Teil der Briefe. Daher werden die Volltexte gemeinsam mit den digitalen Faksimiles der Korrespondenzdokumente über ein abgestimmtes Rechtemanagement sukzessive im Projektverlauf über die Editions-umgebung der BBF (EditionenBildungsgeschichte⁷) soweit möglich unter freier Lizenz bzw. gemäß § 60f UrhG verfügbar gemacht. Die gewonnenen bio-bibliographischen, geographischen sowie Kontakt- und Netzwerkdaten werden nach Abschluss der Auswertung als wertvolle Ergänzung der edierten Texte veröffentlicht. Zudem ist die Anreicherung der GND um bisher unbekannte Personen bzw. um hinzugewonnene biographische Informationen zu bereits erfassten Personen ein weiteres Projektziel. Von zentraler Bedeutung im Projekt ist darüber hinaus die dauerhafte Sicherung sowohl der unikatlen Originaldokumente durch entsprechende konservatorische Maßnahmen als auch aller digitalen Projektdaten.

Daten, Methoden und Werkzeuge

Im Projekt werden vier Datenarten unterschieden, deren Erzeugung aufeinander aufbaut: Er-schließungsdaten (A), digitale Faksimiles (B), digitale Volltexte (TEI/XML) (C) und semantische Daten (D). Am Anfang der Forschungsdatenerzeugung steht die archivische Nacherschließung (A) und Image-Publikation (B). Die Volltexterzeugung (C) erfolgt im TEI/XML DTA-Basisformat, das schließlich die Grundlage eines strukturierten, zitierfähigen Korpus bildet (Ausgangsdaten). Durch die darauf aufbauende semantische Anreicherung, Auswer-

tung und automatisierte Analyse werden Forschungsdaten (Arbeitsdaten) erzeugt (D).

Neben der dauerhaften Sicherung dieser bildungs- und kulturhistorisch wertvollen Originalquellen verfolgt das Infrastrukturteilprojekt die Nachnutzung der im Anschubprojekt erzeugten Forschungsdaten. Der für das Projekt entwickelte Workflow vereinbart archivarchivische Erschließungskonventionen mit den Bedarfen eines bildungshistorischen Forschungsprojektes. Die im Anschubprojekt erzeugten Metadaten werden zunächst mithilfe des Datenbereinigungstools Open Refine⁸ normalisiert, mit den digitalen Repräsentanten verknüpft sowie Orts-, Personen- und Körperschaftsangaben mit GND-Nummern angereichert. Nach dem Import in die Archivdatenbank erfolgen eine Qualitätsprüfung und die Vergabe von eindeutigen, dauerhaften Archivsignaturen. Die XML-Exporte aus der Archivdatenbank bilden die Basis für die Erzeugung von TEI/XML-Vorlagen für jeden Brief, die so bereits alle relevanten Metadaten mitführen und damit als Grundlage für die Transkription der Briefe dienen. Die Transkription erfolgt gemäß einer leicht angepassten Version des DTA-Basisformats (Haaf, Geyken und Wiegand 2014), sodass die Ergebnisse direkt in die Editionsinfrastruktur der BBF eingebunden werden können. Parallel dazu erfolgt die Veröffentlichung der Faksimiles über ScriptaPaedagogica⁹, dem digitalen Textarchiv der BBF. Das System sichert die dauerhafte Verfügbarmachung der Bilder durch die Vergabe von URN, und der integrierte IIIF-Server erlaubt den Export der Digitalisate und der zugehörigen Metadaten in einschlägigen Formaten (z. B. .jpg, .tiff, .mets). Zudem werden für die Digitalisate standardmäßig DFG-Viewerkompatible METS/MODS-Pakete erstellt, die in Zusammenhang mit den hochauflösenden Masterscans im TIFF-Format eine Voraussetzung für den digitalen Langzeiterhalt bieten. Für die weitere wissenschaftliche Auswertung bilden die Transkription, die Metadaten sowie die Auszeichnung und Kodierung der Volltexte die Grundlage. Diese werden im freien, portablen und offenen Format TEI/XML gespeichert und sind somit ebenfalls langzeitarchivierungsfähig. Zudem ist so eine plattformunabhängige Nachnutzbarkeit der Daten möglich. Im Projekt werden die angereicherten TEI/XML-Dateien in strukturierte Textdateien überführt, welche wiederum als Grundlage für die weitere Auswertung in der Analysesoftware MAXQDA Analytics Pro¹⁰ dienen.

Bei der Erzeugung und Auswertung derjenigen Forschungsdaten, die der inhaltlichen Analyse dienen, werden (teil)automatisierte Analyseverfahren in ein klassisches hermeneutisches Vorgehen integriert. Die Volltexte werden auf Grundlage von deduktiv wie induktiv generierten Indizes ausgezeichnet. Angewendet werden ein fachsystematischer Index, ein Kontext- und ein Beziehungsindex. Die induktive Weiterentwicklung der Indizes stellt dabei bereits eine eigene Forschungsleistung dar, wobei die Indizes die Grundlage für die Auszeichnung expliziter wie impliziter Nennungen bilden. Als zentrales Werkzeug dient MAXQDA Analytics Pro, das durch die Anwendung und Generierung von sehr umfangreichen „lebenden“ Indizes eher unkonventionell zur Unterstützung einer hermeneutischen Textanalyse durch qualitative und quantitative Verfahren genutzt wird. MAXQDA bietet in dieser Version Funktionen (z. B. Keyword-in-

Context- und Mixed-Method-Anwendungen), die eine Anreicherung eines qualitativ-quantitativ-hermeneutischen Vorgehens durch teilautomatisierte Verfahren wie Kollokations- bzw. Kookkurrenzanalysen ermöglichen. Über die initialen Forschungsfragen hinaus soll das Textkorpus perspektivisch über Topic-Modeling-Verfahren genauer beschrieben werden.

Die inhaltliche Analyse folgt einem vergleichend-transzendierenden Ansatz. Das besondere Potenzial des Quellenkorpus wird genutzt, um Dynamiken der Verbreitung und Rezeption pädagogischer Theorie und Praxis zwischen vielfältigen Akteur*innen über vielfältige Grenzen hinweg nachzuzeichnen. So wird ein Beitrag geleistet, Bildungsgeschichte(n) kritisch zu hinterfragen und aktuelle Bedingungen im Bildungswesen besser zu verstehen. Als besonders interessant erscheinen Unterschiede zwischen typischen und untypischen, wie zwischen repräsentierten und unterrepräsentierten Akteur*innen; eine systematische Unterscheidung, die im Projekt als zentrales Forschungswerkzeug entwickelt wurde. Im Ansatz werden vergleichende Perspektiven mit Perspektiven zusammengeführt, bei denen es um die Überwindung oder Transzendierung vielfältiger Abgrenzungen geht. Für letztere Perspektiven wird in dem hier beschriebenen Forschungsprojekt der Terminus „transzendierende Ansätze“ neu gesetzt, um Zugänge zu benennen, die Phänomene Grenzen überwindend und abgegrenzte Bereiche durchdringend beschreiben. Transzendierende Ansätze (die über transnationale Ansätze hinausgehen¹¹) werden genutzt, um Transfer- und Zirkulationsprozesse und -zusammenhänge zu erforschen und darzustellen. Vergleichende Perspektiven werden hinzugezogen, um spezifische Ausprägungen zu untersuchen. Transzendierend wie vergleichend werden pädagogische Akteur*innen(gruppen), ihr Wirken, ihr Einfluss sowie Kontakte und Netzwerke zwischen ihnen analysiert. Mit Blick auf inhaltliche, theoretische und paradigmatische Grenzen wird z. B. mit Fokus auf unterschiedliche Milieus oder Berufsgruppen u. a. untersucht, wie Motive, Positionen und Diskurse voneinander abgegrenzt und wie um Deutungshoheiten gerungen wurde, aber auch wie Motive, Positionen und Diskurse, die zunächst als widersprüchlich erscheinen (z. B. traditionelle bzw. konservative versus reformorientierte Positionen) in Theorien und Praktiken vermischt oder miteinander in Einklang gebracht wurden (vgl. Geiss/Reh 2020).

Herausforderungen

Wie oben beschrieben, werden die Daten je nach Verwendungszweck in verschiedenen Formaten benötigt. Also muss sichergestellt werden, dass erstens alle Daten ohne Verlust oder Mehraufwand sowohl für die Ansprüche eines umfänglich durchsuchbaren Volltextkorpus, als auch für die Analyse mit MAXQDA aufgearbeitet werden und zweitens die erzeugten Datenarten idealerweise aufeinander aufbauen und sich gegenseitig ergänzen. Es mussten u. a. Lösungen gefunden werden, Auszeichnungsdaten z. B. zu Personen, Orten und Werken, die während der Transkription in TEI/XML vorgenommen werden, so in der Textdatei auszugeben, dass diese mit

den Volltexten nach MAXQDA importiert und dort als bereits ausgezeichnet übernommen werden.

Bei der Bearbeitung und Auswertung des Textkorpus wird mit unterschiedlichen Repräsentationsformen der Dokumente gearbeitet. Dabei müssen die Charakteristika der unterschiedlichen Formate – sowohl der physischen Originaldokumente wie aller erzeugten Repräsentanten – stets mitgedacht und Hinzufügungen wie Annotationen mitgeführt werden. Gleichzeitig bedeutet die forschungsgestützte Auszeichnung und Migration von Texten in andere Datenformate nicht nur eine Anreicherung der Originaltexte, sondern auch das Hervorbringen von neuen Daten bzw. Quellen. Das geht mit der Herausforderung einher, durch die Auswahl von geeigneten Tools die gewonnenen Informationen nachnutzbar und dauerhaft verfügbar zu machen.

Fazit

Das Projekt baut auf Ergebnisse aus zwei vorhergehenden Projektphasen auf. Es operiert mit unterschiedlichen Datenformaten und verbindet zentrale Aufgaben und Verpflichtungen von Forschenden mit den Aufgaben einer Forschungsbibliothek mit angegliedertem Archiv (vgl. Müller 2019; Reh et al. 2020). Daraus ergeben sich komplexe Herausforderungen: (1) in der Kommunikation zwischen den unterschiedlichen Beteiligten, (2) in den Verfahrensweisen und (3) in der technischen Umsetzung. Diese gehen neben der erwähnten Verwendung unterschiedlicher Datenformate auf verschiedene (Konnotationen von) Begrifflichkeiten und Zielperspektiven zurück. Als ein Projekt, das wesentlich mit digitalen Methoden arbeitet, verorten wir es als ein einschlägiges DH-Projekt für die Bildungsgeschichte. Die Projektergebnisse können in gleichem Maße als End- und Ausgangspunkt von Forschung verstanden werden, da aus Forschendenperspektive tatsächlich neue Quellen erzeugt werden, die durch neue Daten angereichert sind. Aufgrund des erzeugten Mehrwertes werden weiterführende Forschungen eher die erzeugten Abbilder und Datenvarianten der Originalquellen als die ursprünglichen physischen Dokumente auswerten. Alle Daten werden dabei gemäß den FAIR-Prinzipien dauerhaft auffindbar, zugänglich, interoperabel und nachnutzbar bereitgestellt (vgl. Wilkinson et al. 2016).

Bei der Überlegung, welche Datenformate im Projekt genutzt werden und auf welche Weise sie in andere Formate überführt werden, muss zudem deren mehrdimensionale Wirksamkeit – Dokumentation, Erforschung und Erzeugung bildungsgeschichtlicher Daten bzw. Quellen – stets mitgedacht werden. Diese Anforderung geht über den Anspruch hinaus, dass Daten so erzeugt werden müssen, dass sie sowohl Forschung wie auch editorische Bearbeitungen ermöglichen. Tatsächlich zeigt das hier vorgestellte Projekt exemplarisch, welche Bedeutung auch in den Sozial- und Geisteswissenschaften einer engen Kooperation zwischen Akteur*innen in Forschung und Infrastruktur zunehmend zukommt (vgl. Cremer et al. 2021). Um anschlussfähige und langfristig nachnutzbare Forschungsergebnisse und Forschungsinfrastrukturen zu gewinnen, müssen ab der Planung über die Durchführung bis zur Nachbereitung wissenschaftlicher Projekte Synergien zwischen Forschenden

und Expert*innen aus Bereichen der Infrastrukturen genutzt werden. So können nicht nur punktuell die Infrastrukturen möglichst genau auf die Bedarfe von Forschenden angepasst werden, sondern auch erprobte Verfahrensweisen und Strukturen bei der dauerhaften Sicherung von analogen und digitalen Quellen genutzt werden. Durch die Verständigung über Vorgänge, Methoden, Werkzeuge, Begrifflichkeiten und Konzepte und deren gemeinsame (Weiter-)Entwicklung leisten solche Kooperationen auch wertvolle Übersetzungsarbeit zwischen unterschiedlichen Akteur*innengruppen. Erfahrungen und Ergebnisse aus einer solchen gegenseitigen Verständigung kann zukünftigen Vorhaben zugutekommen, da Akteur*innen sowohl aus der Forschung wie in Infrastruktureinrichtungen für die Bedarfe, Denkungsarten und Vorgehensweisen der jeweils anderen Seite sensibilisiert werden. So werden auch Forschende geübt und befähigt, bei der Wahl von Forschungsmethoden und der Arbeit mit Forschungsdaten Potenziale, Bedarfe und Grenzen von Forschungsinfrastrukturen mitzudenken und für aktuelle wie weiterführende Projekte nutzbar zu machen.

Fußnoten

1. Das Projekt wird unter der Leitung von Dr. Katja Grundig de Vazquez, Institut für Bildung und Kultur, Lehrstuhl Historische Pädagogik und Globale Bildung der Friedrich-Schiller-Universität Jena (<https://www.uni-jena.de/>, zuletzt zugegriffen am 03.08.2022, so auch alle im folgenden angegebenen Links) in Kooperation mit Prof. Dr. Sabine Reh, BBF | Bibliothek für Bildungsgeschichtliche Forschung des DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation (<https://bbf.dipf.de/de>) realisiert. Die Leitung des infrastrukturellen Teilprojekts an der BBF verantwortet Dr. Bettina Irina Reimers.
2. Projektwebseite unter <https://www.erziehungsforschung.uni-jena.de/dfgprojektgdv>.
3. Projektdokumentation zum Anschubprojekt unter <https://www.uni-due.de/allgemeine-didaktik/projekte.php>.
4. <https://www.zotero.org>.
5. Im Juli 2019 ging der Nachlass von Wilhelm Rein, der bis dahin als Dauerleihgabe von der Projektleiterin verwahrt und mit Zustimmung des Eigentümers wissenschaftlich bearbeitet worden war, als Schenkung des Urenkels Andreas Rein in das Eigentum des DIPF über und wird im Archiv der BBF als sammelndem Spezialarchiv für die Bildungspraxis und Bildungsgeschichte dauerhaft verwahrt.
6. Zur Merkmalsfestlegung von Briefen vgl. Thiedeke 2020, S. 188 ff.
7. Auf Basis des TEI-Publisher werden gegenwärtig fünf Editionen und Textkorpora verwaltet und – wo rechtlich möglich – veröffentlicht (<https://editionen.bbf.dipf.de>).
8. <https://openrefine.org/>.
9. <https://scripta.bbf.dipf.de>.
10. <https://www.maxqda.de/produkte/maxqda-analytics-pro>.
11. Zum Begriff des Transnationalen vgl. u.a. Roldán Vera/Fuchs 2019, Mayer 2019, Popkewitz 2019.

Bibliographie

Baillot, Anne. 2020. „Digitalisierung und ihre Einflüsse auf den Umgang mit alten wie neuen ‚Briefen‘ in deutscher wie internationaler Perspektive.“ In *Handbuch Brief*, hg. von Matthews-Schlinzig, Marie Isabel, Jörg Schuster, Gesa Steinbrink und Jochen Strobel, 387–398. Berlin: De Gruyter. 10.1515/9783110376531-025.

Budde, Gunilla. 2020. „Geschichtswissenschaft.“ In *Handbuch Brief*, hg. von Matthews-Schlinzig, Marie Isabel, Jörg Schuster, Gesa Steinbrink und Jochen Strobel, 61–80. Berlin: De Gruyter. 10.1515/9783110376531-004.

Cremer, Fabian, Silvia Daniel, Marina Lemaire, Katrin Moeller, Matthias Razum und Arnost Stanzel. 2021. „Data meets history: A research data management strategy for the historically oriented humanities.“ In *Cultural Sovereignty beyond the Modern State*, hg. von Feindt, Gregor, Bernhard Gissibl und Johannes Paulmann, 155–178. Berlin: De Gruyter. 10.1515/9783110679151-009.

Deutsche Forschungsgemeinschaft. 2016. *DFG Praxisregeln „Digitalisierung.“* https://www.dfg.de/formulare/12_151/index.jsp (zugegriffen: 3. August 2022).

Geiss, Michael, und Sabine Reh. 2020. „Konservatismus und Pädagogik im Europa des 20. Jahrhunderts: Einleitung in den Themenschwerpunkt.“ In *Jahrbuch für Historische Bildungsforschung* 26: 9–27.

Haaf, Susanne, Alexander Geyken, Frank Wiegand. 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* (Issue 8): 10.4000/jtei.1114.

Henzel, Katrin. 2020. „Materialität des Briefs.“ In *Handbuch Brief*, hg. von Matthews-Schlinzig, Marie Isabel, Jörg Schuster, Gesa Steinbrink und Jochen Strobel, 222–231. Berlin: De Gruyter. 10.1515/9783110376531-013.

Mayer, Christine. 2019. „The Transnational and Transcultural: Approaches to Studying the Circulation and Transfer of Educational Knowledge.“ In *The Transnational in the History of Education*, hg. von Fuchs, Eckhardt und Eugenia Roldán Vera, 49–68. Cham: Springer International Publishing. 10.1007/978-3-030-17168-1_2.

Müller, Lars. 2019. „Kooperatives Management geisteswissenschaftlicher Forschungsdaten.“ In *ABI Technik* 39 (3): 194–201. 10.1515/abitech-2019-3003.

Nutt-Kofoth, Rüdiger. 2016. „Briefe herausgeben: Digitale Plattformen für Editionswissenschaftler und die Grundfragen der Briefedition.“ In „*Ei, dem alten Herrn zoll ich Achtung gern*“. *Festschrift für Joachim Veit zum 60. Geburtstag*, hg. von Richts, Kristina und Peter Stadler, 575–586. München: Allitera Verlag. 10.25366/2018.2.

Popkewitz, Thomas S. 2019. „Transnational as Comparative History: (Un)Thinking Difference in the Self and Others.“ In *The Transnational in the History of Education*, hg. von Fuchs, Eckhardt und Eugenia Roldán Vera, 261–291. Cham: Springer International Publishing. 10.1007/978-3-030-17168-1_10.

Reh, Sabine, Lars Müller, Stefan Cramme, Bettina Reimers und Marcelo Caruso. 2021. „Warum sich Forschende um Archive, Zugänge und die Nutzung bildungswissenschaftlicher Forschungsdaten kümmern sollten – historische und informationswissenschaftliche Perspektiven.“ In *Erziehungswissenschaft* 31 (61 (2-2020)), 9–20. 10.3224/ezw.v31i2.02.

Roldán Vera, Eugenia und Eckhardt Fuchs. 2019. „Introduction: The Transnational in the History of Education.“ In *The Transnational in the History of Education*, hg. von Fuchs, Eckhardt und Eugenia Roldán Vera, 1–47. Cham: Springer International Publishing. 10.1007/978-3-030-17168-1_1.

Stadler, Peter, Marcel Illetschko und Sabine Seifert. 2016. „Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing.“ In *Journal of the Text Encoding Initiative* (Issue 9): 10.4000/jtei.1433.

Standhartinger, Angela. 2020. „Briefzensur und Briefgeheimnis in der Neuzeit.“ In *Handbuch Brief*, hg. von Matthews-Schlinzig, Marie Isabel, Jörg Schuster, Gesa Steinbrink und Jochen Strobel, 269–275. Berlin: De Gruyter. 10.1515/9783110376531-016.

Thiedeke, Udo. 2020. „Der Brief als individualmediale Kommunikationsform: Eine mediensoziologische Beobachtung.“ In *Handbuch Brief*, hg. von Matthews-Schlinzig, Marie Isabel, Jörg Schuster, Gesa Steinbrink und Jochen Strobel, 187–202. Berlin: De Gruyter. 10.1515/9783110376531-011.

Wilkinson, Mark D. et al. 2016. „The FAIR Guiding Principles for scientific data management and stewardship.“ In *Scientific Data* 3 (1): 160018. 10.1038/sdata.2016.18.

Forschungsperspektiven zur Interaktion mit Musiknotation

Nowakowski, Matthias

matthias.nowakowski@hfm-detmold.de
Center of Music and Film Informatics

Berndt, Axel

axel.berndt@th-owl.de
Center of Music and Film Informatics

Hadjakos, Aristotelis

aristotelis.hadjakos@hfm-detmold.de
Center of Music and Film Informatics

Startpunkt Notensatzprogramme

Musikalische Notation ist das vordringliche Medium zur Musiküberlieferung, aber auch der alltäglichen musikalischen Praxis und ist durch ihre Textualität auch Gegenstand der Digital Humanities. Die wissenschaftliche Beschäftigung mit Musik wie auch die Kommunikation über Musik und musikalische Ideen ist in weiten Teilen der heutigen Musikkultur ohne eine Notenschrift nicht denkbar. Selbst die Interaktion zwischen Musizierenden und mit dem musikalischen Material findet vorrangig

vermittels dieses Mediums statt. Dabei kann die Verschriftlichung von Musik unterschiedliche Formen annehmen, sei es die sogenannte Common Western Music Notation (CWMN), Notenschriften aus anderen Kulturen oder zeitlichen Epochen, die vielfältigen eher technischen Darstellungsformen in Software-Werkzeugen zur Musikproduktion oder die analytischen Visualisierungen aus dem Bereich des Music Information Retrieval (Khuslusi 2020).

Trotz dieser Vielfalt gilt die Interaktion mit Musiknotation aber noch immer primär als Domäne klassischer Notensatzprogramme¹ und wird jenseits dessen kaum tiefergehend thematisiert. Die Funktionen von Notensatzprogrammen umfassen in erster Linie die Eingabe und das Editieren von Notenmaterial nach der CWMN sowie das Platzieren im Layout für den Druck. Sie sind damit Textbearbeitungsprogrammen nicht unähnlich. Ihre Benutzeroberflächen folgen ähnlichen Gestaltungskonzepten, visualisieren das interaktive Notenblatt im Bildzentrum und umgeben es mit vielfältigen Werkzeugpaletten und Modusschaltern, z.B. um Noten zu schreiben, editieren und abzuspielen. Es sind klassische WYSIWYG²-Desktopanwendungen, die für eine filigrane Maus- und Tastatursteuerung konzipiert sind. Diese etablierten Konventionen der Interaktion in einer Desktopumgebung stoßen bei den vielfältigen Anwendungsszenarien von Musiknotation jedoch an ihre Grenzen. Dies geht über das bloße Erstellen, Anzeigen und Ausdrucken der Noten hinaus. Der Abschnitt „Vielfalt der Anwendungsszenarien“ wird dieses Spektrum Überblickhaft umreißen. Abschnitt 3 „Interaktion jenseits von Notensatz“ wird den Fokus entsprechend weiten und Interaktionskonzepte und Technologien aus angrenzenden Forschungs- und Anwendungsgebieten im Bereich Musik in den Blick nehmen. Leere Flecken auf der „Forschungslandkarte“ werden im Abschnitt „Bestimmung des Design Space“ herausgearbeitet und anhand von Beispielen verdeutlicht.

Vielfalt der Anwendungsszenarien

Die heute etablierten graphischen Darstellungs- und Interaktionsformen für Musik möchten wir zunächst nach dem Anwendungsbezug wie folgt systematisieren und ihre jeweiligen Besonderheiten herausarbeiten: *Erstellung*, *Musizieren*, *Präsentation* und *Gaming*.

Erstellung: Neben der manuellen Noteneingabe spielt der Notenscan eine immer wichtigere Rolle. Solche Scans werden mittels Optical Music Recognition-Verfahren in editierbaren Notentext übersetzt (Calvo-Zaragoza 2020). In der historisch-kritischen Musikedition liegen der Erstellung meist mehrere Quellen zu Grunde, die nicht nur (digital) erfasst, sondern auch verglichen werden. Werkzeugen für vergleichende Ansichten kommt dabei eine zentrale Rolle zu (Kepper 2007, Waloschek 2019). Grundsätzlich fanden diese Arbeitsprozesse seitens der Interaktionsforschung bislang kaum Beachtung.

Es ist vor allem aber die Kreativarbeit, welche von spontaneren und direkteren Eingabemöglichkeiten zur Er-

stellung profitieren würde. Beim Komponieren entstehen Skizzen nicht notwendigerweise nur am Schreibtisch oder Klavier, sondern in Alltagssituationen, in denen allenfalls ein Smartphone gerade zur Hand ist. Bei größeren Werken geht oft eine Planung der Formteile und ihrer Proportionen voraus, die dann in beliebiger Reihenfolge oder auch parallel ausgearbeitet werden. Notensatzprogramme können solchen nichtlinearen, kreativen Prozessen aber kaum gerecht werden.

Musizieren: Im Ensemble spielen die Musizierenden oft nicht aus der Partitur, sondern aus Einzelstimmen, und sehen damit nicht, was ihre Mitmusizierenden spielen. In der Bandmusik und in improvisatorischen Kontexten reduziert sich auch das eigene Notenmaterial weiter auf Lead Sheets oder nur mündliche Absprachen zu Akkordfolgen und Tempo. Die Kommunikation beim Ensemblespiel geschieht über den vermittelnden Dirigenten, Sichtkontakt, Bewegungsgesten und das eigene Gehör. Das Notenmaterial ist in dieser Konstellation allerdings kein starres Objekt, das nur noch gelesen wird. Das wird vor allem bei den zahlreichen Eintragungen der Musizierenden deutlich: Interpretationsanweisungen, Hinweise zur Kommunikation, Ergänzungen und Veränderungen von Noten. Bei vernetzten, digitalen Noten könnten diese auch mit den anderen geteilt werden. Maus und Tastatur sind in diesem Szenario nicht praktikabel. Für das Weiterblättern kommen daher Pedale und Taster zum Einsatz. Im Idealfall hört das Gerät sogar mit, führt ein Audio-to-Score Alignment aus und blättert vollautomatisch weiter.

Auf der Seite der Musikproduktion werden Eintragungen in Partituren gezeichnet und dienen dazu, Takes im musikalischen Kontext zu verorten, gelungene oder weniger gelungene Stellen zu annotieren und den Schnittplan zu erstellen. Dabei interagieren die Noten mit den Aufnahmedaten in einer Digital Audio Workstation (DAW) (Waloschek 2017).

Gaming: Visualisierungen sind oft stilisierte und abstrahierte Ableitungen klassischer Musiknotation. Diese können, ebenso wie die erklingende Musik, fest vorgegeben und unveränderlich sein, wobei sie dann Tempo und Rhythmus von Geschicklichkeitsübungen diktieren. Sie können aber auch interaktiv sein, sodass die Spielenden durch ihre Interaktion Einfluss auf die Musik nehmen, sie spielend erzeugen oder arrangieren (Berndt 2011).

Präsentation: Dies zielt vor allem darauf ab, neue Zugänge zum Verständnis der Musik zu schaffen. In den Anwendungsgebieten Musikwissenschaft, Music Information Retrieval und Musikvermittlung gilt die Interaktion daher vor allem der Annotation von Analyseergebnissen. In Videos werden Notentext und klingende Musik synchron dargestellt, um ein mitlesendes Hören zu ermöglichen. Einige YouTube-Kanäle inszenieren ihre Musikanalysen in Form aufwendig angefertigter, annotierter Partiturreduktionen³. Im schulischen und akademischen Musikunterricht findet sich das Notenbild großformatig auf der (digitalen) Tafel wieder, wo kurze Notentexte verfasst, editiert, angehört und nachgespielt werden. Da meist klassische Notensatzprogramme an die Wand projiziert werden, fällt auf, dass hierfür immer wieder zu Maus und Tastatur gegriffen werden muss, also nicht mit dem Tafelbild direkt interagiert wird. In rein virtuellen und hybriden Unterrichtsformen betreiben die Beteiligten jeweils lokal ein Notensatzprogramm, können zwar

Bild und Ton mit den anderen teilen, nicht aber gemeinsam am selben Notentext arbeiten. Hierfür lohnt ein Blick in Museen, wo musikbezogene Medien zumeist von mehreren Besuchern gleichzeitig erlebt und bedient werden können, sei es an großen Multitouch Displays, Tabletops, mittels Freihandgesteninteraktion (Berndt 2016) oder in raumgreifenden Klanginstallationen (Berndt 2019).

Die meisten der hier aufgeführten Anwendungsszenarien erfordern aber mehr als nur neue Funktionen innerhalb der etablierten und durch ihre Konventionen geprägten Notensatzprogramme. Sie erfordern ein unverstelltes Neudenken von interaktiven Zugängen zum Medium Notentext.

Interaktion jenseits von Notensatz

Interaktionen mit Musiknotation lassen sich nur schwer von ihren musikalischen Realisationen trennen. Daher überrascht es nicht, dass die Exploration von neuen Eingabemodalitäten vor allem in der künstlerischen Forschung geschieht. Hier sind meist die schnelle und unmittelbare Eingabe von Noten Hauptaspekt der Betrachtung. Der Fokus auf die künstlerische Forschung konnte vor allem durch eine Institutionalisierung der „New Interfaces for Musical Expression“ (NIME) geschehen. In einer Analyse der Texte der gleichnamigen Konferenzen der letzten 20 Jahre zeigt sich, dass diese vor allem neuartigen Musikinstrumenten, Aufführungen und elektronischen Kontrollmöglichkeiten gewidmet sind (Fasciani 2021). Viele dieser Controller, Software und Interaktionsformen sind jedoch meist schwer generalisierbar und werden nur durch ein Werk oder eine Menge an Werken desselben Künstlers exemplifiziert. Solche Formen sind etwa die Synthese von Noten durch Sensoren am Körper oder Objekte, wie etwa Taktstöcke. Dadurch können aus einer vorher festgelegten Geste Notationen entstehen, die dann von Maschine oder Musizierenden umgesetzt werden. Eine Geste wird übersetzt in musikalische Parameter, wie z.B. Tonhöhe, Dynamik etc., und ermöglichen abstraktere Interaktionen als Notensatzprogramme es erfordern. Es entstehen live notierte Abschnitte, welche als eine Reihe von musikalischen Gesten dargestellt werden. Dabei macht es keinen Unterschied, ob diese Notationen in Form von CWMN oder anderer graphischer Elemente dargestellt werden (Frame 2022). Darüber hinaus sind aktionsbasierte Notationen gestische Anweisungen, die so dargestellt werden, dass sie direkt und ohne Kenntnis von anderen Notationssystemen ausgeführt werden können. Sie werden im Voraus erstellt und ähneln in ihrem Mapping Tabulaturen, erweitern sie aber durch stilisierte Animationen von Hand- oder Körperbewegungen, welche den entsprechenden Klang erzeugen sollen (Clay 2010). Diese Bewegungen können elektronisch analysiert werden, um z.B. automatisch „weiterzublättern“ (Dori 2020).

Als Erweiterung der sensorbedingten Interaktion kann man die Nutzung von Virtual Reality (VR) verstehen. Die Immersion der Nutzenden geschieht durch die Übertragung der körperlichen Bewegung in einen simulierten, dreidimensionalen Raum. Bewegungen können in

den realen Raum abgebildet werden und können, z.B. als Nachzeichnungen dieser Bewegungen, Ausgang einer Notation für Musiker*innen sein (Santini 2022). Dafür benötigt es u.a. Controller, welche das Greifen und Zeigen ersetzen. Ebenso muss die Körperposition verfolgt werden können, um eine Beziehung zu den simulierten Objekten herzustellen. Solches Greifen wird auch in Anwendungen zu Lernszenarien mit haptischen und räumlichen Komponenten genutzt, in welchen Nutzer Noten oder Harmonieverläufe direkt vertikal und horizontal anordnen und so allein durch die Kopfbewegung einen direkten Überblick über das Geschaffene gewinnen können (Shvets 2022).

Im Unterschied dazu wird in der Augmented Reality (AR) kein Raum simuliert, sondern Informationen im realen Raum ergänzt. Neben spezialisierten Geräten, wie der HoloLens,⁴ sind auch viele Smartphones in der Lage, das Livekamerabild zu analysieren und mit Augmentierungen anzureichern. Beispiele für die AR-Nutzung gibt es im Bereich der Instrumentallehre. Darstellung der Noten können als Pianorolle direkt über der Klaviatur dargestellt werden und ermöglicht so das gleichzeitige Beobachten von Notation, Fingern und Instrument (Kim-Boyle 2022).

Jenseits vom Notensatz erhält die Notation also immer mehr räumliche Bedeutung. Durch Web-Technologien wird auch die Entfernung zwischen den Zusammenspielenden beliebig. So können etwa durch Interaktionen der Dirigierenden Notationen direkt auf Displays im Orchester verteilt werden (Andersen 2022), oder es ermöglicht den Musizierenden sich über verschiedene Geräte sowohl in Proben- als auch in Aufführungssituationen zu synchronisieren (Bell 2017, Bell 2021).

Die realweltliche Interaktion mit dem Notenblatt auf dem Notenpult findet über Handgesten (z.B. Umblättern des Notenblattes, Zeigen) und Stift (z.B., Schreiben von Noten, Ergänzen von Vortragsanweisungen) statt. Vor diesem Hintergrund kommt auch der mittlerweile sehr umfangreichen Forschung zu Touch- und Stiftinteraktion - und im genannten Nutzungsszenario insbesondere auf Tablets - eine große Relevanz zu (Baró 2019).

In den Besprechungen der NIMEs werden allerdings nichtkünstlerische Bereiche oft vernachlässigt. So werden z.B. in der Pädagogik musikbezogene Interfaces auch als Mittel genutzt, um das Lernen in anderen Domänen zu erleichtern. Das „Computational Music Thinking“ (Repenning 2019) ist ein Ansatz, um informatische und musikalische Konzepte zu verknüpfen und Wissen aus dem jeweils anderen Bereich zu übertragen. Grundlegend dafür ist die Überzeugung, dass musikalische und programmatische Muster (*Patterns*) übersetzbar seien. Daraus entstehen Programmierumgebungen, die durch die Abstraktion der musikalischen Elemente eher visuellen Programmiersprachen, wie Max/MSP⁵ ähneln. Die Interaktion ist hier aber klar auf Maus und Tastatur ausgerichtet.

Als Ziel der Interaktionsforschung sticht also deutlich die künstlerische Aufführung hervor. Für deren Verwendung in Forschungsbereichen der Digital Humanities möchten wir sie in einem möglichen Design Space einbetten.

Bestimmung des Design Space

Als Design Space verstehen wir den Raum, in dem die Elemente vorhanden sind, die für die Konstitution eines Forschungsbereiches herangezogen werden können. Der Raum sei hier aber nicht nur als Metapher verstanden. Als Instrument der Positionsbestimmung im Design Space kann die Kombination der Elemente als Vektor dargestellt werden. Abbildung 1 zeigt nur eine flache Darstellung der verschiedenen Dimensionen, welche in verschiedenen Farben kodiert sind.

Die Auswahl der Elemente basiert auf der Analyse von Interaktionsformen mit musikalischer Notation. Die Dimensionen werden definiert durch die Objekte der Interaktion, der Granularität ihrer Interaktionsgegenstände, des musikalischen und sozialen Kontexts der Nutzung, sowie Fragen nach verschiedenen Darstellungsformen und Interaktionsgeräten. Diese Auflistung ist beileibe nicht vollständig, geben aber eine Orientierung und Klassifikation aktueller und zukünftiger Forschungsfelder.

In klassischen Musiknotationsprogrammen interagieren die Nutzenden mit Notationszeichen auf der logischen Ebene nach der Klassifikation von Maxwell (Maxwell 1981): Es können nur syntaktisch wohlgeformte Notationen erstellt werden. Die kleinste Granularität der Interaktion ist hier also die Manipulation von Musikzeichen auf der logischen Ebene. In handgeschriebenen Notationen finden sich allerdings manchmal Verletzungen dieser „Logik“, zum Beispiel durch Auslassungen, abgekürzte Notationen oder absichtliche Überschreitungen. Daher erlaubt das bei digitalen historisch-kritischen Ausgaben häufig benutzte MEI-Format (Hankinson 2011) die Modellierung von Inhalten u.a. auf der grafischen Ebene, um damit auch Merkmale von Handschriften nachbilden zu können. Eine weitere Ebene tiefer kann es auch von Interesse sein, mit den grafischen Grundelementen des Notentextes als Vektorgrafik zu interagieren, um besondere gestalterische Ziele umzusetzen. Andererseits ist es auch sinnvoll mit Notentexten auf höheren Granularitätsstufen zu interagieren: Beispielsweise werden in digitalen Musikeditionen häufig verschiedenen Quellen taktweise verlinkt, um Ähnlichkeiten und Abweichungen zu erkennen. Interfaces sollten auf unterschiedlichen Stufen bedienbar sein, unter anderem um graphische Primitive, Musikzeichen auf der grafischen oder logischen Ebene, Takte, Stimmen, von Anwender*innen ausgewählte Teile bis hin zu Musiksammlungen zu manipulieren.

Solche Interfaces werden auch in vielfältigen sozialen Kontexten verwendet, ob gemeinsam oder allein, beim Einstudieren eines Stückes, in einer Probe, bei der gemeinsamen Arbeit an einem Notentext oder im Musikunterricht. Auch der Musikstil hat einen Einfluss auf das Interface Design, insbesondere welche Darstellungsformen zum Einsatz kommen: Wenn improvisiert wird, könnten Lead Sheets zum Einsatz kommen. Bei Neuer Musik oder in NIME-Performances mit Controllern und Live-Elektronik, kommen meist grafische Partituren zum Einsatz. Piano Roll und auf Spektrogramm basierende Darstellungsformen können unabhängig von musika-

lichen Vorkenntnissen in verschiedenen Stilen genutzt werden.

Zuletzt spielen auch die eingesetzten Interaktionsgeräte und -technologien eine Rolle. Noch vor WIMP⁶-basierten Interfaces für PCs und Laptops, sind heute Touch-Interfaces für Smartphones und Tablets das wichtigste Medium für den Umgang mit Noten, während Interfaces für AR/VR sich noch in einem experimentellen Stadium befinden.

Ein Beispiel: Die Elemente sind aus den jeweiligen Kategorien frei kombinierbar. So ließe sich z.B. ein Vektor erstellen, der alle fünf Kategorien umfasst: [AR, Sammlungen, Skizzenhaft, NIME, Lehre]. In diesem Fall könnte man sich eine AR-Anwendung (vielleicht auf einem Smartphone) vorstellen, welche auf vordefinierte Sammlungen angewandt werden könnte. Die Skizzen die daraus automatisiert erstellt werden, wären Grundlagen für Performances mit einem NIME, welche in einem Lehrkontext (im Musikunterricht oder an einer Musikhochschule) besprochen werden. Beiträge dazu würden beispielsweise die kreative Aneignung von klassischer Musik durch AR erforschen. Verkleinert man den Vektor und lässt den sozialen Kontext „Lehre“ und das Interaktionsmedium „AR“ heraus, so hätte man eine Fokussierung auf Sammlung beispielsweise als Grundlage generativer Musik oder für die Nachmodellierung aktueller kompositorischer Prozesse. Bei der Aufführung bietet sich darüber hinaus eine Aufzeichnung als Forschungsgegenstand an, die selbst weiteren digitalen Analysen unterzogen werden kann.

Mit diesem Paper haben wir die vielfältigen Perspektiven für Interaktionen mit Musiknotationen analysiert. Dabei haben wir auch gezeigt welches Innovationspotential hier noch offen liegt – ein lohnendes Feld für zukünftige Forschungen und Entwicklungen.

Fußnoten

1. Zu den bekanntesten Vertretern zählen etwa MuseScore (Werner Schweer & The MuseScore developer community 2002), Dorico (Steinberg 2022), Sibelius (Avid Technology, Inc. 1993), Finale (MakeMusic, Inc. 1988) und Capella (capella-software AG 1992).
2. What You See Is What You Get.
3. Z.B. <https://www.youtube.com/c/BigDaddyDave/videos> (zugegriffen am 29. Juli 2022)
4. <https://www.microsoft.com/de-de/hololens> (zugegriffen am 29. Juli 2022)
5. <https://cycling74.com/products/max> (zugegriffen am 29. Juli 2022)
6. Windows, Icons, Menus, Pointer

Bibliographie

Andersen, Drake. 2021. „Indra: A Virtual Score Platform for Networked Musical Performance“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’20/21*, hg. von Rama Gottfried, Georg Hajdu, Jacob Sello, Alessandro Anatrini und John MacCallum, 227–234. Hamburg, Ger-

many: Hamburg University for Music / Theater. <https://doi.org/10.5281/zenodo.4764757>

Baró, Arnau, Pau Riba, Jorge Calvo-Zaragoza und Alicia Fornés. 2019. „From Optical Music Recognition to Handwritten Music Recognition: A Baseline“. *Pattern Recognition Letters* 123: 1–8. <https://doi.org/10.1016/j.patrec.2019.02.029>

Bell, Jonathan. 2021. „Distributed Notation in the Browser, an Overview“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’20/21*, hg. von Rama Gottfried, Georg Hajdu, Jacob Sello, Alessandro Anatrini und John MacCallum, 251–259. Hamburg, Germany: Hamburg University for Music / Theater. <https://doi.org/10.5281/zenodo.4764764>

Bell, Jonathan, und Benjamin Matuszewski. 2017. „SMARTVOX. A Web-Based Distributed Media Player as Notation Tool for Choral Practices“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’17*, hg. von Helena Lopez Palma, Mike Solomon, Emiliana Tucci und Carmen Lage, 99–104. A Coruna, Spain: Universidade da Coruna. <https://doi.org/10.5281/zenodo.924143>

Berndt, Axel. 2011. „Diegetic Music: New Interactive Experiences“. In *Game Sound Technology and Player Interaction: Concepts and Developments*, hg. von M. Grimshaw, 60–76. Hershey, PA: IGI Global. <http://dx.doi.org/10.4018/978-1-61692-828-5.ch004>

Berndt, Axel, Simon Waloschek und Aristotelis Hadjakos. 2016. „Hand Gestures in Music Production“. In *Proc. of the Int. Computer Music Conf. (ICMC)*, hg. von H. Timmermans. Utrecht, The Netherlands: International Computer Music Association, HKU University of the Arts Utrecht, Gaudeamus Muziekweek.

Berndt, Axel, und Joachim Veit, Hrsg. 2019. *Inside Beethoven! Das begehbare Ensemble - Begleitpublikation zur Klanginstallation der Hochschule für Musik Detmold zum Septett op. 20 und Trio op. 38 (mit CD)*. Bonn, Germany: Beethoven-Haus Bonn.

Calvo-Zaragoza, Jorge, Jan Haji# Jr und Alexander Pacha. 2020. „Understanding optical music recognition“. *ACM Computing Surveys (CSUR)* 53, Nr. 4: 1–35. <https://doi.org/10.48550/arXiv.1908.03608>

Clay, Arthur, und Jason Freeman. 2010. „Preface: Virtual Scores and Real-Time Playing“. *Contemporary Music Review* 29, Nr. 1: 1–1. <https://doi.org/10.1080/07494467.2010.509587>

Dori, Gil. 2020. „Using Gesture Data to Generate Real-Time Graphic Notation: a Case Study“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’20/21*, hg. von Rama Gottfried, Georg Hajdu, Jacob Sello, Alessandro Anatrini und John MacCallum, 68–74. Hamburg, Germany: Hamburg University for Music / Theater.

Fasciani, Stefano, und Jackson Goode. 2021. „20 NIMes: Twenty Years of New Interfaces for Musical Expression“. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Shanghai, China. <https://doi.org/10.21428/92fbeb44.b368bcd5>

Frame, Ciaran, Alon Ilisar und Sam Trolland. 2022. „Mutable Gestures: A New Animated Notation System for Conductor and Chamber Ensemble“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’2022*, hg. von Vin-

cent Tiffon, Jonathan Bell und Charles de Paiva Santana, 1–7. Marseille, France: PRISM Laboratory.

Hankinson, Andrew, Perry Roland und Ichiro Fujinaga. 2011. „The Music Encoding

Initiative as a Document-Encoding Framework.“ In *ISMIR*, 293–298. <https://doi.org/10.5281/zenodo.1417609>

Kepper, Johannes und Daniel Rößenstrunk. 2007. „Das Edirrom-Projekt. Werkzeuge für digitale Formen wissenschaftlich-kritischer Musikeditionen“. *Forum Musikbibliothek* 28: 36–49.

Khulusi, Richard, Jakob Kusnick, Christofer Meinecke, Christina Gillmann, Josef Focht und Stefan Jänicke. 2020. „A Survey on Visualizations for Musical Data“. In *Computer Graphics Forum*, 39:82–110. <https://doi.org/10.1111/cgf.13905>

Kim-Boyle, David. 2022. „The Twittering Machine“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’2022*, hg. von Vincent Tiffon, Jonathan Bell und Charles de Paiva Santana, 15–21. Marseille, France: PRISM Laboratory.

Maxwell, John Turner. 1981. „Mockingbird: An interactive composer’s aid“. Diss., Massachusetts Institute of Technology.

Repenning, Alexander, Jürg Zurmühle, Anna Lamprou und Daniel Hug. 2020. „Computational Music Thinking Patterns: Connecting Music Education with Computer Science Education through the Design of Interactive Notations“. In *Proceedings of the 12th International Conference on Computer Supported Education - Volume 1: CSME*, 641–652. INSTICC, SciTePress. <http://dx.doi.org/10.5220/0009817506410652>

Santini, Giovanni. 2022. „Linear: A Multi-Device Augmented Reality Environment for Interactive Notation and Music Improvisation“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’2022*, hg. von Vincent Tiffon, Jonathan Bell und Charles de Paiva Santana, 37–43. Marseille, France: PRISM Laboratory, 2022. isbn: 979-10-97498-03-0.

Shvets, Anna, und Samer Darkazanli. 2022. „Conditional Semantic Music Generation in a Context of VR Project “Graphs in Harmony Learning”“. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’2022*, hg. von Vincent Tiffon, Jonathan Bell und Charles de Paiva Santana, 62–69. Marseille, France: PRISM Laboratory.

Waloschek, Simon. 2017. „Audioschnitt in digitalen Notaten“. In *INFORMATIK 2017, 47. Jahrestagung der Gesellschaft für Informatik*, hg. von M. Eibl und M. Gaedke. LNI. Chemnitz, Germany: Chemnitz University of Technology, Gesellschaft für Informatik, GI. https://dx.doi.org/10.18420/in2017_13

Waloschek, Simon, Aristotelis Hadjakos und Alexander Pacha. 2019. „Identification and Cross-Document Alignment of Measures in Music Score Images“. In *Proc. of the 20th Int. Society for Music Information Retrieval Conf. (ISMIR)*. Delft, The Netherlands: Int. Society for Music Information Retrieval. <https://doi.org/10.5281/zenodo.3527760>

From the Secret Archive to open and fair access. Ways of modelling legal ecclesiastical data from the XVI and XVII centuries

Albani, Benedetta

albani@lhl.mpg.de

Max Planck Institut for Legal History and Legal Theory, Deutschland

Anokhina, Alexandra

anokhina@lhl.mpg.de

Max Planck Institut for Legal History and Legal Theory, Deutschland

Park, Yohan

park@lhl.mpg.de

Max Planck Institut for Legal History and Legal Theory, Deutschland

A complex starting point: from an inaccessible archive to FAIR data approach

The Vatican Archive is the private archive of the pope. Until 2019 it was called the Vatican Secret Archives and this particular name has given rise to legends, novels, films, rumours... Despite the fact that the archive has been open to the public since 1881 without any restrictions, and although the connotation of 'secret', which has always tickled the fancy of writers and journalists, derives from the ancient meaning of the adjective *secretum*, indicating that the holdings were the pope's private and personal property,¹ this archive and its very rich heritage remain inaccessible to many in several respects. This concerns both the archive holdings as a whole (archival fonds and series) and the historical documents preserved there.

In contrast to other historical archives whose inventories are based on International Standard Archival Description (International Council on Archives, 2000) that allow for interoperability of data, the Vatican Archive is so immense and complex² that it has not yet been equipped with a modern archival description system. Bear in mind that there is still no uniform archive guide, no comprehen-

sive inventory, and that researchers often have to consult indexes and handwritten inventories, even dating back to the 14th century, in order to know the contents of the archive series and select the documents they are interested in. Moreover, for reasons of document ownership and authenticity, the archive's policy is clearly resistant to an open access approach to data: documents cannot be photographed by researchers. Digital reproductions are expensive and can only be used for personal research purposes. Handwriting recognition software,³ which usually foresee the sharing of digitized images with other users, cannot be used to 'read' these documents.⁴ A final aspect that makes the documents preserved in the Vatican Archives hardly accessible concerns the specific knowledge and skills needed to read, understand and interpret them. Although common to historical research in general, this aspect takes on an important significance in the case of papal documents for the exegesis of which it is necessary to master certain specific disciplines and techniques.⁵

Among the various scientific objectives of our research project is also to improve the accessibility of data obtained from historical sources held in the Vatican Archive and to offer them to the scientific community according to FAIR principles (Findability, Accessibility, Interoperability, Reuse) (Wilkinson *et al.*, 2016.), thus overcoming the inaccessibility of papal sources through modern technologies. In this paper we describe the methods and tools we are developing to address these challenges.

Our research project focuses on one of the most active bodies of the Roman Curia – the complex of organs and authorities that constitute the administrative apparatus of the Holy See – between the modern and contemporary ages: the Congregation of the Council, which we affectionately call "SCC"⁶. This dicastery was appointed for more than 350 years to oversee the correct interpretation and implementation of the Council of Trent⁷ and the administration of justice around all disciplinary matters contained in the council and later on other papal laws. The Congregation of the Council was composed of cardinals and other personnel and met periodically to discuss and decide legal cases that came to it from all over the Catholic world. It had consultative, judicial and gracious functions and its jurisdiction ideally extended to the whole world. Based on the 1.5 km of the SCC archive, preserved in the Vatican Archive, the group has compiled a dataset describing approximately 35,000 *positiones*, that is judicial cases that took place in front of the SCC between 1564 and 1680 and involved thousands of people and institutions from all over the Catholic world and beyond.

In this paper we describe how we developed 1) modern standards for processing historical data, 2) the semantic model and 3) visualization strategies for contextualizing data in global history.

knowledge graph. Using the Semantic Web technologies, we constructed an event-based ontology with key elements (Event, Time, Place, Institution, Source) that trace legal administrative events and changes of each diocese quoted in the *positiones*.

Our model proposes an extension of CIDOC-CRM (Doerr, 2003) through suggesting additional subclasses and object properties. Although CIDOC-CRM provides strong semantic expressiveness, this instrument was designed for modelling within the cultural heritage domain that set semantic limitations for modelling the structures of the administrative acts in historical perspective. Therefore, we proposed additional subclasses and object properties in order to fill a methodological gap.¹³

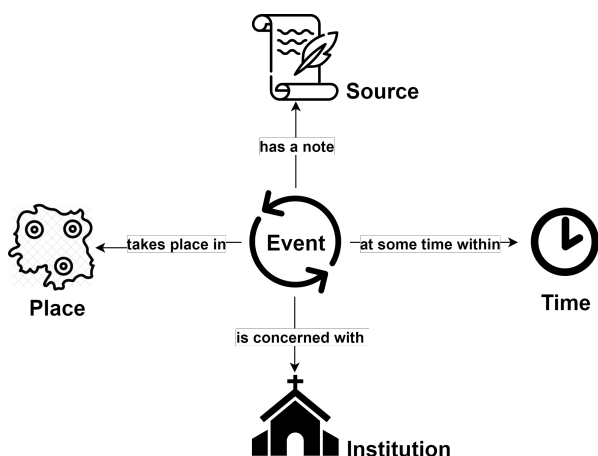


Figure 4: Key elements in our ontology

Publishing open and FAIR data

Based on our ontology data model and in accordance with FAIR principles, we prepare the publication of our data as linked open data by using the combination of Resource Description Framework (RDF) triples and Internationalized Resource Identifiers (IRIs) which ensures for maintaining the semantic interoperability of data (Berners-Lee, 2009). Since for many of our actors, places, institutions there are no entries in any authority databases yet, therefore we will assign an URI to each resource and offer our data IDs as authority records for other projects.

In accordance with FAIR principles, this achieves and guarantees, on the one hand, the findability and accessibility of research data, and on the other hand, it enables researchers to use our data as a reference as well as for their own purposes. This provides also a fundamental framework for reusability of the data. We believe that providing our data in RDF triple format, we fulfil the requirements for interoperability and standardization of data according to the World Wide Web Consortium (W3C) recommendations. To achieve these goals, we currently identify and match entries from our dataset, such as dioceses and other actors, by using the reconciliation web service API providing OpenRefine that enables to align datasets to entries from the already existing dataset¹⁴ in OpenRefine (Thalbach et al., 2021). Therefore, we

enrich semantically our dataset via authority data and achieve useful degree of interoperability and connectivity between our dataset and external data.

Developing open source visualization instruments for contextualizing the SCC data and metadata in history

We consider that adequate historical contextualisation of data is necessary in order to avoid anachronistic or teleological biases and allow to interpret the results in scientific ways that correspond academic standards of historical research. To contextualize the SCC data and its metadata we developed visualization approaches that allow us to show our data not only as tables but also in interactive graphic format, to improve accessibility from the user perspective. There are various methods for visualizing historical data, focused on modern ideas of time¹⁵ and space¹⁶ that however don't count uncertainty, unclearness and incompleteness of historical data. We stress on methods of visualization of unclear, uncertain and incomplete data for reducing the impact of modern gaze on time and history. We also aim to remain the complexity of the sources from the perspective of an historian, i.e. preserve the original data as presented in the sources even if unclear, uncertain or incomplete. This part of our project aims to extend the functionality of this kind of instrument with setting up the custom controllers, impossible to maintain in already existed tools and therefore make uncertainty, unclearness and incompleteness visible and accessible as an important specifics of the data. The principal languages we use for this part of the project are R and JavaScript.

We developed the *SCC Timeline Explorer* web app in which the history of the SCC is placed in global historical context. Through the visualization of parallel timelines, this application allows to explore different aspects of the history of the SCC (evolution of competences, turnover of the personnel, frequency of the meetings) within global history (Global Legal History, History of the Roman Curia, Pontificates). We enriched the original data with descriptive metadata, which provide information in case of uncertainty, unclearness and incompleteness of data. Since many inputs have incomplete information, we decided to use vis.js¹⁷ and R, for working with both – the data in an advanced way and the graphic visualization in dynamics, which also allows to visualize the uncertainty, unclearness and incompleteness in a functional (R) and aesthetic way (JS + CSS). We also combined the data and controllers, as we want to let users choose the settings.

Using R we created a reactive dataset, which connects the original dataset and controllers. Since the graphic part of vis.js works only with dd-mm-yyyy format, we added an external CSS file to stylize uncertain entities, which set the uncertainty as semi-transparent elements. For setting the controllers, firstly, in R we wrapped them as a reactive function. In a basic way, this solution uses shiny for calling reactive and dynamic functions in-

side a server part. Secondly, these functions are visualized in vis.js.

```
subset <- reactive({
  category <- paste0(c(input$sub_group_a,
    input$sub_group_b), collapse = "|")
  category <- gsub("","|",category)
  type_search <- paste0(c(input$type), collapse = "|")
  type_search <- gsub("","|",type_search)
  dataset[grepl(category, sub_group) & grepl(type_search, type_of_act)]})

#Setting the type controller
checkboxGroupInput("check_act_reform", "Type of the event",
  choiceNames = mapply(type, icon, FUN = function(type, iconUrl),
    SIMPLIFY = FALSE, USE.NAMES = FALSE),
  choiceValues = c("act", "reform", "foundation"),
  selected = c("act", "reform", "foundation"))
```

Figure 5: A basic example of setting a subset with controllers.

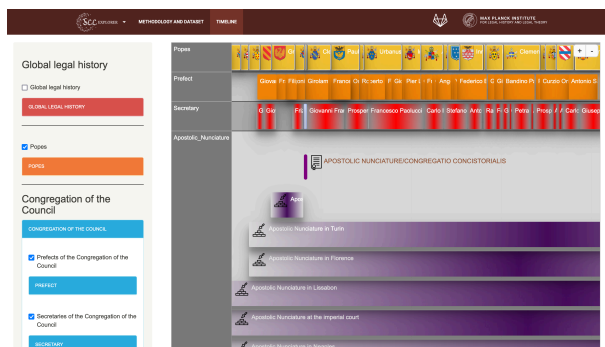


Figure 6: A timeline prototype using vis.js and R.

Each label has a trigger on click, which shows the information about the event or range, which is based on reactive values. What is unique in this approach is the possibility to set up reactive controllers for groups, custom reactive values, advanced aesthetics, and a way to bypass limits of the vis.js in the case of unclear, uncertain or incomplete information that is essential for historical datasets.

Products of our development will be published in the format of web applications in the SCC Explorer Platform. The commented code will be available on our GitHub. The results can be validated and repeated with other datasets. R functions are designed universal and scalable for other Digital Humanities requests.

Conclusion

Our research project offers various methods of modeling and visualization of a large scaled database and metadata, considered in a FAIR way with open access, open data and open code. Firstly, we developed a knowledge graph using a semantic data model, which offers a data-centred approach for the Congregation of the Council in global context in the early modern period. Secondly, we created open access research tools for contextualizing historical data, including unclear, uncertain and incomplete information, into a big picture of global legal history, providing a graphic and accessible visualization of the data. For the time being, our data covers the period from 1564 to 1680 and are concentrated on the SCC, but

data collection will continue for the later periods and our methods can be profitably applied also to other bodies of the Roman Curia. Therefore, our research and the tools we are developing can be of great importance for the understanding of the administration of justice in the Western World from the Late Middle Ages to Contemporary period.

Fußnoten

1. On the history and heritage of the Vatican Apostolic Archive there is an extensive, though very fragmentary, bibliography. For an initial overview, see *Religiosa archivorum custodia. IV Centenario della Fondazione dell'Archivio Segreto Vaticano (1612-2012)*. 2015. Città del Vaticano: Archivio Segreto Vaticano and Gualdo, Germano. 1989. *Sussidi per la consultazione dell'Archivio Vaticano*. Città del Vaticano: Archivio Vaticano.
2. Today, the archive consists of more than 600 archival fonds from different types of institutions and covers approximately 85 linear kilometres of shelving.
3. For example the software *Transkribus*: <https://readcoop.eu/transkribus/>.
4. The University of Roma Tre in collaboration with the Vatican Apostolic Archive in the frame of the project *In codice ratio* (<http://www.inf.uniroma3.it/db/icr/>) is developing a software for text recognition of the volumes of the *Registra Vaticana*. This is an ambitious project, yielding excellent results, but unfortunately, due to the profound differences between medieval, early modern and contemporary writing systems, it will not be applicable to manuscript documents that are not written in the same style as the Vatican Registers. Firmani, Donatella, Paolo Merialdo, Elena Nieddu and Simone Scardapane. 2017. "In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks." In *11th Italian Workshop on Artificial Intelligence for Cultural Heritage*; Lastilla, Lorenzo, Serena Ammirati, Donatella Firmani, Nikos Komodakis, Paolo Merialdo, Simone Scardapane. 2022. "Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library." In *Information Processing and Management* 59(3).
5. For example papal diplomatics and specific branches of palaeography, sphragistics, chronology and chronography as well as the history of the Papacy and the Roman Curia.
6. In this paper, we will refer to the Congregation of the Council as 'SCC', an acronym used in specialist literature and derived from the Latin name of the institution: *Sacra Congregatio Concilii*.
7. The Council of Trent (1545-1563) was the 19th ecumenical council of the Catholic Church and remained in force until the 19th century. It was of central importance to the history of the Western world and beyond. Politically, the council attempted, unsuccessfully, to settle the rift between Catholics and Protestants that had arisen from Lutheran ideas by addressing important theological and ecclesiological issues (doctrine of justification, role of grace, existence of saints, doctrine of the sacraments, etc.). Within the Catholic world, the council constituted an important point of reference on a pastoral and juridical level. It is the council that remained in

force the longest of all the councils recognised by the Catholic Church (307 years) and thus left an important imprint on Catholic societies, an imprint that is still visible today.

8. This is due both to the very high costs of such an operation and other factors more related to source criticism such as the fragmentary nature of some of the cases decided by the SCC (some are divided into several phases also preserved in different volumes), the precarious state of preservation of some volumes, the extreme complexity of the structure of the processes (identification of the parties involved, the role of the cardinals who were members of the congregation, the institutions mentioned often related to canon law issues, the locations etc.) and the handwritings with which the documents were written (abbreviations, symbols etc.).

9. This phase of the work (2013-2019) was carried out by Dr. Benedetta Albani and Dr. Francesco Russo between 2013 and 2019 and was coordinated and financed by the Max Planck Research Group "Governance of the Universal Church after the Council of Trent" directed by Dr. Albani.

10. Vatican Apostolic Archive (AAV), *Congr. Concilio, Positiones*, 1-271.

11. Actors and their semantic roles (petitioners, members of the dicastery, lawyers and procurators, senders, addressees, sponsors etc.), institutions (dioceses, parishes, bodies of the Roman Curia, secular authorities etc.), places (spatial data, including coordinates, toponyms, etc.), temporal entities (events, dates, time periods), legal procedures, legal subject matters. For the moment, we have evidence of at least 8,000 petitioners, 1,500 places, 900 dioceses, 700 abbeys, 80 religious orders, 130 cardinals, 17 pontificates. These are preliminary results. Definitive data will be provided after the ongoing data cleaning process will be completed.

12. Biographical data of persons, metadata on the history of the mentioned institutions (dioceses, religious orders, churches, monasteries, abbeys), geographical coordinates of places and historical evolution of place names, bibliographical references, etc.

13. The limitation of the CIDOC model already shows in mapping on our data model, for example, the class *crm:E8 Acquisition* did not exactly fit our modelling notion, because the class was primarily intended to design the legal process in the museum landscape, such as lending artwork to a gallery.

14. For example, some data were provided by the project *Monasteries, Collegiate churches, and Convents of the Holy Roman Empire and neighbouring countries* (<https://adw-goe.de/germania-sacra/klosterdaten-bank/datenservice/>), by the Göttingen Academy of Sciences.

15. For example the *Timeline JS* by Northwestern University Knight Lab (<http://timeline.knightlab.com>) or *MIT HyperStudio's Chronos Timeline* (<http://hyperstudio.mit.edu/software/chronos-timeline/>) cannot visualize unclear and incomplete dates.

16. Almost all projects in geo-spatial visualization are based on GIS, which proposes a modern gaze on geography that does not allow to operate with unclear, uncertain and incomplete data in historical sense. For example, *Esri Story Maps* and the *Digital Humanities* projects (<https://collections.storymaps.esri.com/hu->

manities/). On modern geography based visualization in R see Weinberg Eric. 2018. "Using Geospatial Data to Inform Historical Research in R." In *Programming Historian* 7.

17. The official website of vis.js library: <https://visjs.org/>.

Bibliographie

Berners-Lee, Tim. 2009. "Linked Data - Design Issues." <https://www.w3.org/DesignIssues/LinkedData.html>

Doerr, Martin. 2003. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." In *AI Magazine* 24(3): 75. <https://doi.org/10.1609/aimag.v24i3.1720>.

Firmani, Donatella, Paolo Merialdo, Elena Nieddu and Simone Scardapane. 2017. "In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks." In *11 th Italian Workshop on Artificial Intelligence for Cultural Heritage*.

Gualdo, Germano. 1989. *Sussidi per la consultazione dell'Archivio Vaticano*. Città del Vaticano: Archivio Vaticano.

International Council on Archives. 2000. *ISAD(G): General International Standard Archival Description*. Ottawa: International Council on Archives.

Lastilla, Lorenzo, Serena Ammirati, Donatella Firmani, Nikos Komodakis, Paolo Merialdo, Simone Scardapane. 2022. "Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library." In *Information Processing and Management* 59(3). <https://doi.org/10.1016/j.ipm.2022.102875>.

Religiosa archivorum custodia. IV Centenario della Fondazione dell'Archivio Segreto Vaticano (1612-2012). 2015. Città del Vaticano: Archivio Segreto Vaticano.

Thalhath, Nishad, Nagamori, Mitsuharu, Sakaguchi, Tetsuo, and Sugimoto Shigeo. 2021. "Wikidata Centric Vocabularies and URIs for Linking Data in Semantic Web Driven Digital Curation." In *Metadata and Semantic Research*, ed. Emmanouel Garoufallou and Maria-Antonia Ovalle-Perandones, 336-344. Metadata and Semantic Research. MTSR 2020. Communications in Computer and Information Science, vol 1355. Springer, Cham. https://doi.org/10.1007/978-3-030-71903-6_31.

Weinberg Eric. 2018. "Using Geospatial Data to Inform Historical Research in R." In *Programming Historian* 7. <https://doi.org/10.46430/phen0075>.

Wilkinson Mark D., Dumontier, Michel, Aalbersberg, I. Jsbrand Jan et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." In *Sci Data* 3. <https://doi.org/10.1038/sdata.2016.18>.

Zhang, Zuopeng Justin. 2017. "Graph Databases for Knowledge Management." *IT Professional* 19(6): 26-32. <https://doi.org/10.1109/MITP.2017.4241463>.

Gattungen und Emotionen in der Lyrik des Realismus und der frühen Moderne

Kröncke, Merten

merten.kroencke@uni-goettingen.de
Universität Göttingen, Deutschland

Konle, Leonard

leonard.konle@uni-wuerzburg.de
Universität Würzburg, Deutschland

Winko, Simone

simone.winko@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Der literarische Wandel vom Realismus zur Moderne ist nach wie vor Gegenstand vielfältiger literaturwissenschaftlicher Debatten. Eine dieser Debatten betrifft die Frage, wie sich die Gestaltung von Emotionen in lyrischen Texten veränderte. Während einige typologisch argumentierende Forscher:innen (z. B. Andreotti 2014) davon ausgehen, dass die moderne Lyrik gegenüber der traditionelleren Lyrik des Realismus zu einer nüchternen, nicht-emotionalen Ausdrucksweise tendiert, argumentieren andere (z. B. Winko 2003) ausgehend von zeitgenössischen Selbstbeschreibungen, dass die Lyrik der frühen Moderne – historisch verstanden – sehr wohl Emotionen gestaltet, wenn auch in modifizierter Weise (vgl. zu dieser Debatte auch Konle u. a. 2022).

Zu bedenken ist, dass die damalige Lyrik keine homogene Einheit bildet. Aussagen darüber, inwiefern sich die Emotionsgestaltung ‘der’ Lyrik veränderte, lassen sich differenzieren. Ein naheliegendes, wichtiges Differenzmerkmal lyrischer Texte – und damit ein potentiell relevanter Einflussfaktor auf die Emotionsgestaltung – ist die Gattung. Ziel dieses Beitrags ist deshalb, zu prüfen, inwiefern sich lyrische Gattungen in unserem Untersuchungszeitraum durch spezifische Emotionsprofile auszeichnen und ob die Entwicklung der Lyrik vom Realismus zur frühen Moderne unter der Perspektive der Emotionsgestaltung, die wir in (Konle u. a. 2022) herausgearbeitet haben, mit der Zugehörigkeit zu bestimmten lyrischen Gattungen zusammenhängt. An diesem Beispiel soll gezeigt werden, wie wichtig es auch für quantitative Untersuchungen ist, literarische Phänomene nicht isoliert zu

betrachten, sondern als Teil eines komplexen Systems zu modellieren, in dem unterschiedliche Faktoren einander beeinflussen. Der Beitrag versteht sich als Schritt hin zu einer solchen multifaktoriellen Modellierung von Literatur und literarischem Wandel.

Ressourcen

Das zu analysierende Korpus besteht aus Texten in Lyrikanthologien aus dem Untersuchungszeitraum, die sich auf Gedichte von Zeitgenoss:innen konzentrieren. Das Teilkorpus ‘Realismus’ umfasst Gedichte aus Anthologien, die zwischen 1850 und den frühen 1880er Jahren publiziert wurden; das Teilkorpus ‘Moderne’ enthält Texte aus Anthologien, die um 1900 erschienen sind und deren Herausgeber:innen die Gedichte aufgrund ihrer Modernität ausgewählt haben. Welche Texte als ‘realistisch’ und welche als ‘modern’ gelten, wird in diesem Beitrag also aus zeitgenössischer Sicht (und nicht aus Sicht heutiger Forscher:innen) modelliert.¹ Da keine der einbezogenen Anthologien nach 1911 erschienen ist, beschränken sich die Analysen auf die *frühe* Moderne.

Tabelle 1: Korpus Statistik

	Anthologien	Gedichte	Wörter
Realismus	8	3367	400k
Moderne	12	2882	320k
Insgesamt	20	6249	720k

Für 1412 Korpustexte wurden die Gattungszugehörigkeit und die Emotionsgestaltung annotiert. Die Gattungsannotation hält fest, ob ein Text bestimmten im Untersuchungszeitraum relevanten thematischen Gattungen (Naturlyrik, Liebeslyrik usw.) sowie nicht-thematischen Gattungen (Ballade, Lied usw.) angehört und ob er situativ bestimmt oder situativ unbestimmt ist. Wenn im Folgenden von ‘Gattungen’ die Rede ist, sind in aller Regel die thematischen Gattungen gemeint, auf die sich dieser Beitrag erst einmal konzentriert. Die Gattungszuordnung ist weder exklusiv noch zwingend: Während der Annotation konnten einem Text genau eine, aber auch keine oder mehrere Gattungen zugewiesen werden.² Die Emotionsannotation zielt darauf ab, die im Text gestalteten Emotionen (und nicht die Leseremotionen) zu erfassen. Genutzt wurde ein Set von 40 diskreten Emotionen, darunter zum Beispiel Liebe, Trauer, Hoffnung, Sehnsucht oder Hass. Einerseits handelt es sich um Emotionen, die in gängigen Emotionstheorien (Ekman 1992; 1999; Plutchik 1980b; 1980a; 2001) als grundlegend angesehen werden, andererseits wurden zusätzliche Emotionen, die in den Korpustexten häufig vorkommen, aufgenommen, um das Emotionsset an das historische Material anzupassen. Die Annotationseinheiten sind Wörter bzw. Wortfolgen.³ Da für viele einzelne Emotionen nicht genügend Annotationen vorliegen, um ihre maschinelle Detektion trainieren zu können, werden die Emotionen nachträglich zu sechs Gruppen zusammengefasst: Liebe, Freude, Trauer, Erregung/Überraschung, Angst und Wut. Die Gruppierung orientiert sich an der Emotionshierarchie in Shaver u. a. (1987).

Tabelle 2: Annotierte Gattungen

Gattung	Liebe	Natur	Philosophie	Religion
Anzahl	350	286	163	100
Gattung	Poetologie	Politik	Kultur	Geschichte
Anzahl	61	21	104	118

Jedes Gedicht wurde zunächst von zwei Annotator:innen annotiert; anschließend haben die beiden Annotator:innen ihre Annotationen miteinander verglichen, die Disagreements diskutiert und eine Konsensannotation erstellt. Auf den Konsensannotationen beruhen alle weiteren Auswertungen. Das Agreement der Einzelannotationen beträgt 0.69 (Krippendorffs Alpha).

Methoden

Um Gattungslabel für das gesamte Korpus herstellen zu können, sollen Classifier trainiert und angewandt werden.⁴ Da Gedichte sehr kurz sind, kommt ein Verfahren auf bag-of-words-Basis nicht in Betracht. Stattdessen wird auf das Fine-Tuning neuronaler Sprachmodelle zurückgegriffen. In einer Vorstudie zur Gattung Liebeslyrik sollen das geeignetste Sprachmodell und Hyperparameter⁵ ermittelt werden. Getestet werden die Modelle gbert-large (Chan, Schweter, und Möller 2020) und gottbert-base (Scheible u. a. 2020). Das beste Ergebnis wird mit gbert-large erzielt (Batchsize 20, Learning Rate 1e-5 und 50 Epochen). Durch eine Anpassung auf das Lyrikkorpus durch fortgesetztes Pretraining⁶ (Gururangan u. a. 2020) kann die Performance weiter gesteigert werden (siehe Tab. 3 ,gbert-large-poetry).

Tabelle 3: Ergebnis der Vorstudie zur Gattung 'Liebeslyrik'. Evaluation durch 5-fold-cross validation.

Modell	Accuracy (std)
gbert-large-poetry	.880 (.022)
gbert-base	.854 (.040)
Gottbert-base	.826 (.032)

Das Modell wird mit den ermittelten Hyperparametern verwendet, um binäre Classifier für alle übrigen Gattungen zu trainieren. Da die Klasse der nicht zur fokussierten Gattung gehörenden Gedichte in jedem Fall größer ist, wird epochenweise Random Undersampling angewandt. Für politische Lyrik wird kein Modell trainiert, da nicht genügend Beispiele vorhanden sind (siehe Tabelle. 2).

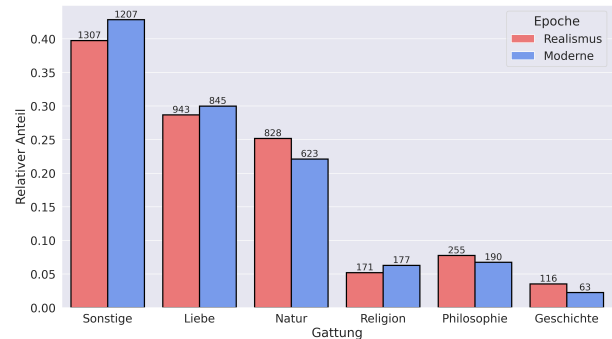
Tabelle 4: Ergebnisse der Evaluation der Classifier für thematische Gattungen (5-fold cross-validation).

Gattung	Liebe	Natur	Poetologie	Geschichte	Politik	Philosophie	Religion	Kultur
Acc	.880	.833	.623	.872	-	.723	.815	.470
Std	.022	.024	.224	.026	-	.028	.059	.169

Tabelle 4 zeigt die Qualität der trainierten Classifier. Da die Performance für die Gattungen Kultur und Poetologie nicht ausreicht, um solide Analysen zu ermöglichen, werden diese im Weiteren nicht behandelt und fallen, wie Politik, der Kategorie 'Sonstige' zu, die daneben diejenigen Gedichte umfasst, die keiner thematischen Gattung zugeordnet wurden.

Für Informationen zur Annotation von Emotionen und Modellen zur Emotionsdetektion siehe (Konle u. a. 2022).

Ergebnisse

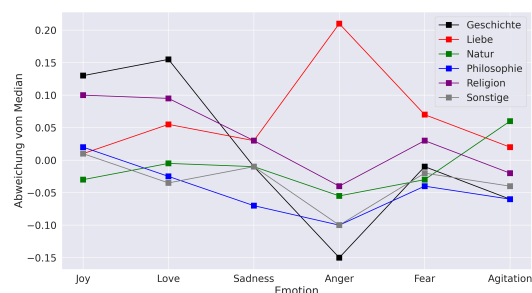


Verteilung von thematischen Gattungen in Realismus und Moderne. Absolute Anzahl über Balken. Die Kategorie 'Sonstige' repräsentiert diejenigen Gedichte, die keiner der übrigen in dieser Abbildung gezeigten thematischen Gattungen zugeordnet wurden.

Verteilung von thematischen Gattungen in Realismus und Moderne. Absolute Anzahl über Balken. Die Kategorie 'Sonstige' repräsentiert diejenigen Gedichte, die keiner der übrigen in dieser Abbildung gezeigten thematischen Gattungen zugeordnet wurden.

60% der untersuchten Gedichte konnten (mindestens) einer der fünf Gattungen Liebeslyrik, Naturlyrik, religiöse Lyrik, philosophische Lyrik oder Geschichtslyrik zugeordnet werden. Liebeslyrik und Naturlyrik sind im Korpus deutlich verbreiteter als religiöse Lyrik, philosophische Lyrik und Geschichtslyrik. Im Epochenvergleich werden nur begrenzte Verschiebungen sichtbar; die Rangfolge der Häufigkeiten bleibt konstant (siehe Abb. 1).

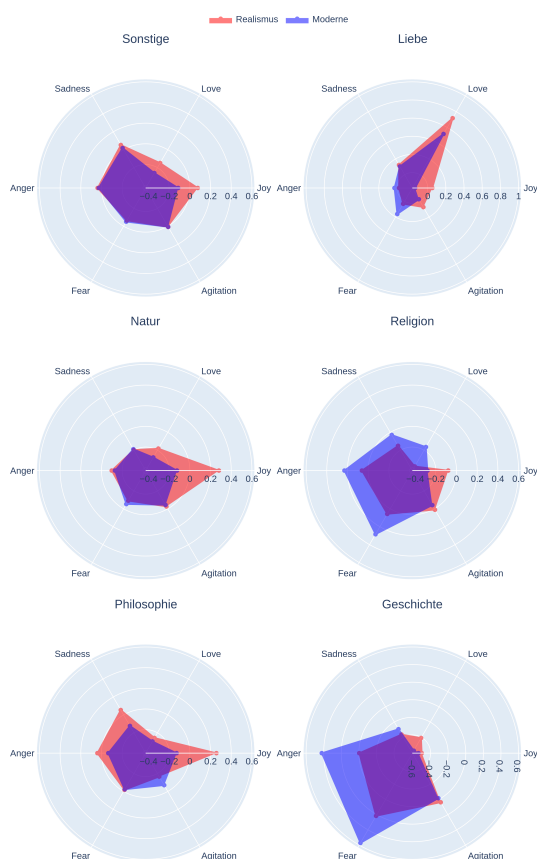
Wie die Gattungen kommen auch die Emotionen im Korpus unterschiedlich oft vor: Emotionen der Gruppen Liebe, Freude und Trauer werden deutlich häufiger gestaltet als Emotionen der Gruppen Erregung/Überraschung, Angst und Wut. Die Häufigkeit positiver Emotionen – Liebe und Freude – nimmt zur Moderne hin ab (Konle u. a. 2022).



Emotionsprofile nach Gattungen. Die Emotionen wurden mit dem Median aller Gedichte normalisiert, um auch Unterschiede in wenig häufigen Emotionen (Anger, Fear, Agitation) sichtbar zu machen. Die 0-Linie markiert die Übereinstimmung mit dem Median aller Gedichte, darüber bedeutet mehr und darunter weniger Emotion.⁷

Die Ergebnisse (Abb. 2) zeigen, dass Gattungen eigenständige Emotionsprofile ausbilden; dies gilt auch für die nicht durch Emotionen definierten Gattungen. Einige Gattungen verhalten sich in Hinsicht auf bestimmte Emotionen ähnlich, z.B. philosophische Gedichte und Geschichtslyrik in Hinsicht auf die Gruppen Freude und Trauer, weichen aber in anderen Emotionen stark voneinander ab. Während z.B. in Liebeslyrik – wenig überraschend – Emotionen der Gruppe Liebe dominant sind, werden in Geschichtslyrik Emotionen der Gruppen Liebe und Freude unter- und Emotionen der Gruppen Wut und Angst überproportional dargestellt.

Ein kontrastiver Blick auf Gattungen unter der Perspektive von Epochen (Abb. 3) zeigt, dass sich (1) die lyrische Emotionsgestaltung in der frühen Moderne gegenüber dem Realismus verändert, und zwar (2) je nach Gattung in unterschiedlicher Weise.



Emotionsprofilen nach Epochen und Gattungen. Die Abbildung zeigt anders als Abb. 2 nicht die Abweichung vom Median, sondern den Mittelwert der Emotionen.

Der summarische Trend über alle Gattungen zu weniger Emotionen, verursacht vor allem durch den Rückgang positiver Emotionen in der Moderne (siehe Konle u.a. 2022), betrifft nicht alle Gattungen in gleicher Weise. Während sich die Ergebnisse zu Liebes-, Natur- und philosophischer Lyrik noch diesem Trend zuordnen lassen,

entwickeln sich Geschichts- und religiöse Lyrik durch Zunahme negativer Emotionen eigenständiger. Religiöse Lyrik läuft dem Trend sogar durch ein vermehrtes Auftreten von Emotionen der Gruppe Liebe entgegen. Diese Befunde erlauben einen differenzierten Blick auf Emotionen, Gattungen und deren Veränderung in der frühen Moderne.

Case Study: Religiöse Lyrik

Anschließend an unsere Ergebnisse stellt sich die Frage nach den Faktoren, welche die gattungsspezifischen Emotionsprofile zu unterschiedlichen Zeitpunkten beeinflussen. Für Geschichtslyrik bietet sich die These an, dass mit der Thematisierung von Konflikten und Feindseligkeiten (Detering und Trilcke 2013) negative Emotionen einhergehen. Ob in der frühen Moderne diese Themen zunehmen oder stärker mit Emotionen verbunden werden, muss geprüft werden.

Die Entwicklung der Emotionen in religiöser Lyrik ist noch komplexer, es nehmen nicht nur negative Emotionen zu, sondern auch die der Gruppe Liebe, bei gleichzeitigem Rückgang der Gruppe Freude.⁸ Diese Beobachtung kann an Einsichten der Forschung anschließen und diese ergänzen. Bekannt ist, dass moderne Lyrik trotz z.T. dezidiert Kritik noch „vielfältig auf religiöse Traditionen bezogen“ bleibt (Detering 2016, 126), dass sie sich u.a. stärker von christlichen Institutionen entfernt und religiöse Motive umdefiniert. Offenbar verbindet sie aber auch mehr und andere Emotionen mit dem Thema als die Lyrik des Realismus. Einen ersten Hinweis auf den Zusammenhang zwischen Motiven und Emotionen gibt Tabelle 5, die u.a. eine Verschiebung der distinktiven Substantive von der Institution zu Personen sowie eine Tendenz zu symbolfähigen Ausdrücken zeigt. Die Verschiebung zum persönlichen Glaubensaspekt könnte die erhöhte Anzahl der Emotionen erklären. Betrachtet man die in der frühen Moderne häufiger werdenden Emotionen (Tab. 6) zeigt sich, dass ein Teil der Zunahme über die Sexualisierung religiöser Inhalte erklärt werden kann. Zur Erklärung der vermehrten negativen Emotionen wären weitere Untersuchungen nötig. Aufschlussreich ist auch, dass sich der Darstellungsmodus religiöser Lyrik insofern ändert, als in Gedichten der frühen Moderne verstärkt Emotionswörter und Wörter vorkommen, die Emotionen körperlich, z.B. gestisch oder mimisch ausdrücken.

Tabelle 5: Distinktive Substantive der Gattung Religion. Ranking nach Keyness. genauer: Simple Maths Parameter (Brezina 2018, S.85).

Realismus	Moderne
Wald	Gott
Kirche	Seele
Kaiser	Nacht
Bischof	Auge
Segen	Mensch
Grab	Tod
Priester	Weib
Glocke	Licht
Mönch	Welt
Heil	Stimme
Werk	Traum
Sieg	Kreuz
Haus	Brust
Friede	Blut
Dom	Volk
Andacht	Erde
Lenz	Herz
Sonntag	König
Mahl	Kind
Ehr	Leib

Tabelle 6: Verschiebung der Emotionen zwischen Realismus und Moderne in religiöser Lyrik (Datengrundlage: Manuell annotierte Gedichte).

Abnehmend	Zunehmend
Freude	Liebe
Ausgeglichenheit	Abneigung
Lust (nicht-sexuell)	Sehnsucht
Aufregung	Leid
Dankbarkeit	Lust (sexuell)

Fazit

Die gleichzeitige Beobachtung von Gattung und Emotion zeigt nicht nur, dass Gattungen wesentlichen Einfluss auf die Verteilung von Emotionen haben, sondern auch, dass der Übergang von Realismus zu früher Moderne innerhalb von Gattungen eigenen Dynamiken folgt. Im Fall religiöser Lyrik und Geschichtsliteratur verlaufen diese Dynamiken sogar in Widerspruch zum Gesamttrend. Im Anschluss ergeben sich drei weitere Perspektiven: Wie sehen die Emotionsprofile der nicht-thematischen Gattungen aus? Welche anderen differenzierenden Faktoren neben der Gattung können wir identifizieren, z.B. Autorschaft, Publikationsort, Intertextualität usw.? Ein wichtiger Aspekt könnte die Zeit sein, bezogen auf kleinere Einheiten als Epochen: Wie entwickeln sich die Lyrik und die einzelnen Gattungen unter der Perspektive der Emotionsgestaltung in der Zeit, etwa von Jahrzehnt zu Jahrzehnt?

Fußnoten

1. Korpus: Winko u. a. 2022; Korpusbeschreibung: Winko u. a. 2022a.
2. Annotationsrichtlinien Lyrische Gattungen: Kröncke u. a. (2022b)
3. Annotationsrichtlinien Emotionsmarker und Emotionen: Kröncke u. a. (2022)
4. Code und Daten: https://github.com/LeKonArD/Gattungen_und_Emotionen_dhd2023
5. Batchsize (5,10,15,20), Learning Rate (1e-5, 5e-5, 1e-4) und Epochen (10,20,50)
6. 10 Epochen über das gesamte Korpus, Batchsize 64, Learning Rate 1e-5
7. Dieses Vorgehen ist mit dem Fokus unseres Aufsatzes auf Gattungen begründet. Die gewählte Normalisierung führt dazu, dass Unterschiede zwischen Textgruppen visuell identifiziert werden können. Da die absoluten Häufigkeiten der Emotionen im Korpus stark schwanken (Liebe kommt 7-fach häufiger als Wut vor), ist es nicht möglich, diese in einer Graphik abzubilden, ohne die Unterschiede der Gattungen zu marginalisieren. Ähnliches gilt für die Art der Normalisierung: Wir haben uns für den Median über alle Gedichte entschieden, da dies die 'neutralste' Variante ist. Normalisierung mit den Gedichten, die keiner Gattung zugeschlagen werden, erscheint wahllos, da wir über diese Gruppe nichts wissen. Eine Normalisierung durch balanciertes Sampling der Gattungen impliziert die Annahme, dass jede Gattung gleich häufig ist und die Verteilungsunterschiede durch einen *Selection Bias* entstanden sind. Dies ist aber in Anbetracht der Größe des Korpus unwahrscheinlich.

8. Ein t-test auf den euklidischen Distanzen der Emotionsprofile religiöser Lyrik in Moderne und Realismus resultiert in einem signifikanten Unterschied ($p < 0.05$).

Bibliographie

- Andreotti, Mario. 2014. *Die Struktur der modernen Literatur: Neue Wege in die Textanalyse. Einführung Epik und Lyrik*. 5. Aufl. Bd. 1127. Wien/Köln/Weimar.
- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. 1. Aufl. Cambridge University Press. <https://doi.org/10.1017/9781316410899>.
- Chan, Branden, Schweter, Stefan, Möller, Timo (2020). German's next language model. In: COLING 2020 – Proceedings of the 28th International Conference on Computational Linguistics. December 8-13, 2020 (Virtual Event), pp. 6788–6796
- Detering, Heinrich. 2016. „Lyrik und Religion“. In *Handbuch Lyrik*. Theorie, Analyse, Geschichte, herausgegeben von Dieter Lamping, 119–28. Stuttgart: Metzler.
- Detering, Heinrich, und Peer Trilcke, Hrsg. 2013. *Geschichtsliteratur. Ein Kompendium*. 2 Bde. Göttingen: Wallstein.
- Ekman, Paul. 1992. „An Argument for Basic Emotions“. *Cognition and Emotion* 6 (3–4): 169–200. <https://doi.org/10.1080/02699939208411068>.
- Ekman, Paul. 1999. „Basic Emotions“. In *Handbook of Cognition and Emotion*, herausgegeben von Tim Dalgleish und Mick J. Power, 45–60. Chichester, UK: John Wiley & Sons, Ltd. <https://doi.org/10.1002/0470013494.ch3>.
- Gururangan, Suchin, Ana Marasovič, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, und Noah A. Smith. 2020. „Don't Stop Pretraining: Adapt Language Models to Domains and Tasks“. arXiv:2004.10964 [cs], Mai. <http://arxiv.org/abs/2004.10964>.
- Konle, Leonard, Merten Kröncke, Fotis Jannidis, und Simone Winko. 2022. „Emotions and Literary Periods“. In *Digital Humanities Conference Abstracts*, 278–81. Tokyo. <https://dh2022.dhii.asia/dh2022bookofabstracts.pdf>.
- Kröncke, Merten, Fotis Jannidis, Leonard Konle, und Simone Winko. 2022a. „Annotationsrichtlinien Lyrische Gattungen“. <https://doi.org/10.5281/zenodo.6021007>.
- Kröncke, Merten, Fotis Jannidis, Leonard Konle, und Simone Winko. 2022b. „Annotationsrichtlinien Emotionsmarker und Emotionen“. <https://doi.org/10.5281/zenodo.6021152>.
- Plutchik, Robert. 1980a. „A General Psychoevolutionary Theory of Emotion“. In *Theories of Emotion*, 3–33. Elsevier. <https://doi.org/10.1016/B978-0-12-558701-3.50007-7>.
- Plutchik, Robert. 1980b. „A Psychoevolutionary Theory of Emotions“. *Social Science Information* 21 (4–5): 529–53. <https://doi.org/10.1177/053901882021004003>.
- Plutchik, Robert. 2001. „The Nature of Emotions“, *American Scientist*, 89 (4): 344–50.
- Scheible, Raphael, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, und Martin Boeker. 2020. „GottBERT: a pure German Language Model“. arXiv. <http://arxiv.org/abs/2012.02110>.
- Shaver, Phillip, Judith Schwartz, Donald Kirson, und Cary O'Connor. 1987. „Emotion Knowledge. Further Exploration of a Prototype Approach“. *Journal of Personality and Social Psychology* 52 (6): 1061–86. <https://doi.org/10.1037/0022-3514.52.6.1061>.

Winko, Simone. 2003. *Kodierte Gefühle: zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900*. Berlin: Erich Schmidt Verlag.

Winko, Simone, Leonard Konle, Merten Kröncke, und Fotis Jannidis. 2022a. „Lyrikanthologien 1850-1910“. <https://doi.org/10.5281/zenodo.6053951>.

Winko, Simone, Leonard Konle, Merten Kröncke, und Fotis Jannidis. 2022b. „Korpusbeschreibung der Lyrikanthologien 1850-1910“. <https://doi.org/10.5281/zenodo.6053972>.

GND und Normdaten für europäische Literatur? Personen und Werke in den multilingualen Korpora von ELTeC

Calvo Tello, José

calvotello@sub.uni-goettingen.de

Niedersächsische Staats- und Universitätsbibliothek Göttingen, Georg-August-Universität Göttingen

Rißler-Pipka, Nanette

nanette.rissler-pipka@gwdg.de

Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG)

Barth, Florian

florian.barth@uni-goettingen.de

Niedersächsische Staats- und Universitätsbibliothek Göttingen, Georg-August-Universität Göttingen

GND, Normdaten und die Digital Humanities

Viele Projekte in den Digital Humanities verwenden Identifier für Entitäten wie Personen, Werke, Körperschaften oder Orte, die auf eindeutige Einträge in Normdaten-Verzeichnissen oder in Knowledge Bases verweisen (Barth 2022 u. a.; Rosenkötter und Fischer 2020; Fischer und Jäschke 2018; Herrmann und Lauer 2018; Dieckmann, Hermes, und Neuefeind 2017). Zu diesen Ressourcen gehören u. a. die Gemeinsame Normdatei (GND), andere Normdaten von Nationalbibliotheken, Wikidata, VIAF, DBpedia, Getty oder CERL. Jede dieser Ressourcen ist nach verschiedenen Kriterien aufgebaut und bietet unterschiedliche Funktionalitäten. Dies bringt Vor- und Nachteile für die Projekte mit sich, die sie nutzen. Zum Beispiel hat Wikidata keinen echten Normie-

rungscharakter im Gegensatz zu Normdaten-Verzeichnissen wie der GND. Wikidata hat jedoch den Vorteil, dass Nutzende selbständig neue Entitäten anlegen können. Dies ist bei den durch Bibliotheken verwalteten Normdaten-Verzeichnissen meist nur auf Antrag möglich (in Zukunft sollen, zumindest für die GND, die Eingabemöglichkeiten durch angepasste Webformulare und Redaktionsumgebungen erweitert werden, vgl. Kett u. a. 2022).

Nichtsdestotrotz ist die Sprache des Forschungsobjekts (z. B. von Textkorpora) und die Wahl der Normdaten-Ressource stark voneinander abhängig. Einige Projekte, die mit deutschsprachigen Texten arbeiten, haben sich für die GND entschieden, um Personen oder Werke zu identifizieren, u. a. die Digitale Bibliothek im TextGrid Repository,¹ das Deutsche Textarchiv² oder die deutschsprachigen Korpora aus DraCor (Fischer u. a. 2019) und ELTeC (Burnard, Schöch, und Odebrecht 2021). Projekte aus der deutschsprachigen Wissenschaftslandschaft, die im Bereich der nicht-deutschen Philologien forschen, entscheiden sich eher für andere Ressourcen, um Personen und Werke zu identifizieren. Die romanistischen Korpora der CLiGS-TextBox (Französisch, Spanisch, Italienisch und Portugiesisch; Schöch u. a. 2019), die spanischen Korpora CoNNSA (Calvo Tello 2021) und CONHA (Henny-Krahmer 2018), und die französischen Korpora in ELTeC und DraCor benutzen für die Identifikation ihrer Entitäten überwiegend Wikidata und VIAF.

Die Bibliotheken und Fachinformationsdienste (FIDs) vieler Philologien in Deutschland verwenden die GND für die Sacherschließung ihrer Titel. Sie reichern umgekehrt die GND mit immer mehr Daten zu fremdsprachigen Autor*innen und deren Werke an. Es liegt nahe, dass die GND weiterhin hauptsächlich Autor*innen und Werke aus dem deutschsprachigen Raum verzeichnet. Das heißt jedoch nicht, dass die GND eine Quelle ist, die sich nur für die Germanistik eignet. Generell ist die GND in Form von Agenturen organisiert, die sich auf viele Bibliotheken und andere Institutionen in Deutschland verteilen und von der DNB koordiniert werden.³ Seit 2020 spielt die GND außerdem eine neue Rolle durch das Projekt „GND für Kulturdaten“ (Rosenkötter und Fischer 2020; Balzer u. a. 2019) und seit 2021 durch die Beteiligung in der NFDI⁴ (Text+, NFDI4Culture).⁵

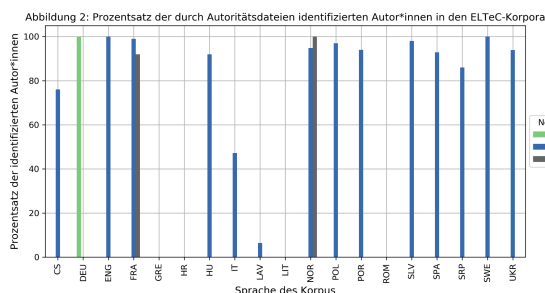
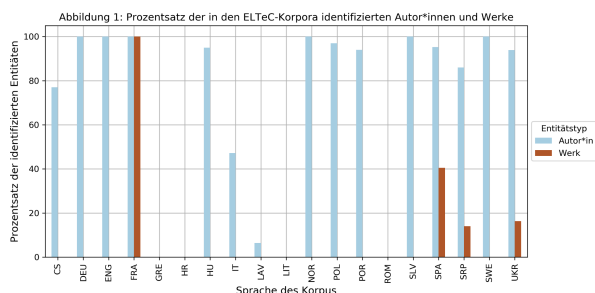
Daher fragen wir uns: Wie stark ist das Ungleichgewicht innerhalb der GND zwischen Einträgen zu deutsch- und fremdsprachiger Literatur?⁶ Können die Anglistik, die Romanistik, die klassischen Philologien, die Slavistik und andere damit rechnen, ihre Entitäten in der GND zu finden?

ELTeC: mehrsprachige, vergleichbare Korpora für europäische Literatur

Unsere Frage beantworten wir anhand der multilingualen Korpora von ELTeC. Dabei handelt es sich um literarische Korpora in verschiedenen europäischen Sprachen, die in der COST-Action Distant Reading erstellt wurden (Schöch u. a. 2021; Odebrecht, Burnard, und Schöch 2021). Das Ziel des Projektes war die Zusammenstel-

lung vergleichbarer Korpora mit 100 Romanen pro Sprache. Aktuell wurde dies für 11 Sprachen erreicht, während für 10 andere Sprachen weniger Romane vorliegen. Außerdem wurde jedes Dokument mit Metadaten zu Autor, Werk, Edition und Text ausgezeichnet. Auf Grundlage dieser Metadaten konnten wir unsere Analysen erstellen.

Die Personen und Werke in ELTeC sind teilweise bereits mit Wikidata, GND oder VIAF eindeutig identifiziert. Die Abbildungen 1 und 2 zeigen im Vergleich die Annotation mit Normdaten für Autor*innen und Werke pro Sprache und die Wahl der Normdatenressource für die Identifikation von Autor*innen.



Für die Sprachen Griechisch (GRE), Kroatisch (HR), Litauisch (LIT) und Rumänisch (ROM) konnten offenbar keine Normdaten verwendet werden. Um die Vergleichbarkeit der Ergebnisse zu gewährleisten, werden diese Sprachen bei den folgenden Analysen ausgeschlossen. Außer für das französische (FR) und in kleinen Teilen für das spanische (SPA), serbische (SRP) und ukrainische (UKR) Korpus wurden keine Werknormdaten eingetragen. Für die Autor*innen wurde überwiegend VIAF genutzt. Nur das französische und norwegische Korpus wurde auch mit Wikidata und das deutsche Korpus als einziges mit GND-IDs versehen.

Methode

Wir gehen in zwei Schritten vor, um ein vollständiges Bild über die mögliche Abdeckung mit Normdaten zu erhalten. Zunächst extrahieren wir die IDs der Autor*innen aus den TEI-Dokumenten der ELTeC-Korpora. Anhand der IDs werden die fehlenden Identifier aus Wikidata, GND und VIAF extrahiert. Das gelingt für die GND über die API von Lobid,⁷ und für Wikidata und VIAF durch

ihre native API. Nach diesem Schritt erhalten alle Autor*innen eindeutige Identifier aus allen drei Ressourcen, falls Mappings gefunden werden konnten.

Im nächsten Schritt werden die Werke mit Rückgriff auf die Autor*innen-ID identifiziert. Auch wenn für vier ELTeC-Korpora Werk-IDs (überwiegend mit VIAF) bereits vom Projekt erfasst wurden, ignorieren wir diese, um die gleiche Methode für alle Korpora anzuwenden. Wir führen drei parallele *Reconciliation*-Prozesse mit den drei Ressourcen für jedes Werk durch. Genauer, werden alle Werke aller Autor*innen abgerufen, um die Ähnlichkeit zwischen dem Titel des Werks in ELTeC und allen Titeln der Werke der jeweiligen Autor*in aus den Normdaten zu vergleichen. Für jede mögliche Paarung wird ein statistischer Wert für die Ähnlichkeit zwischen beiden Titeln berechnet (0 für Titel, die gar keine Gemeinsamkeiten haben, 1 für Titel, die deckungsgleich sind). Der Werktitel mit der höchsten Ähnlichkeit wird ausgewählt. Für die weitere Analyse werden nur die Werke berücksichtigt, deren Wert höher als 0.5 liegt. Für diese drei parallelen und automatischen *Reconciliation*-Prozesse werden die APIs von Lobid, Wikidata und VIAF benutzt.

Ergebnisse

Ausgehend von den bereits in ELTeC identifizierten Autor*innen (vgl. Abb. 1) wird überprüft, ob diese auch in den jeweils anderen Normdatenressourcen (GND, Wikidata, VIAF) vorhanden sind. Daher sind hier Sprachen, für die keine Normdaten in ELTeC existieren (Griechisch, Kroatisch, Litauisch, Rumänisch), nicht berücksichtigt.

Ergebnisse zu Autor*innen

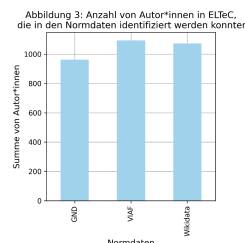
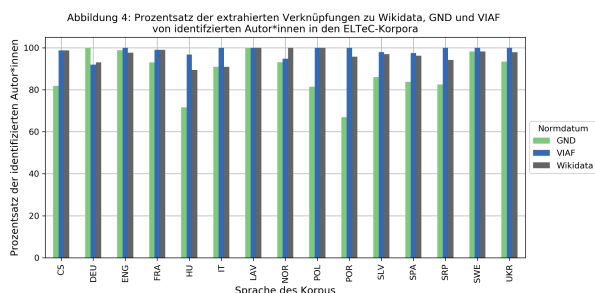


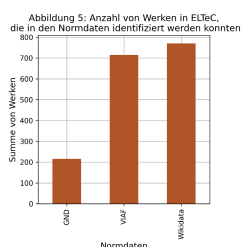
Abbildung 3 zeigt die Summe der Normdaten zu Autor*innen pro Ressource, die in den hier betrachteten Korpora gefunden oder ergänzt werden konnten. Für alle drei Ressourcen ist die Abdeckung hier sehr gut. Die GND liegt im Vergleich nur leicht zurück.

Die Verteilung dieser Daten pro Sprache wird in Abbildung 4 gezeigt. Das Bild entspricht der Zusammenfassung aus Abbildung 3. Erwartungsgemäß hat das deutsche Korpus die höchste Quote in der GND. Das norwegische Korpus kann mehr Treffer mit Wikidata als mit VIAF erzielen. Für Tschechisch und Polnisch erreichen sowohl VIAF als auch Wikidata sehr gute Ergebnisse. Neben dem Deutschen bietet die GND eine gute Abdeckung für Sprachen wie Englisch, Französisch, Italienisch, Norwegisch, Schwedisch und Ukrainisch.⁸

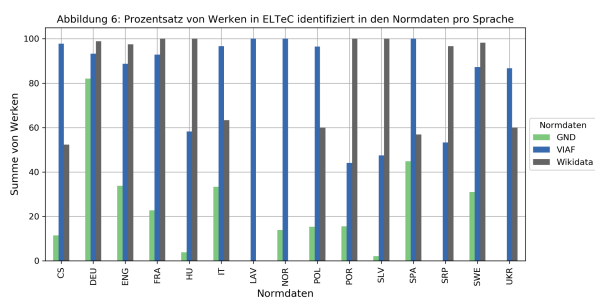


Ergebnisse zu Werken

Das Bild ändert sich erwartungsgemäß, wenn nicht Autor*innen, sondern Werke in den drei Ressourcen gesucht werden (vgl. Abb.5). Während VIAF und Wikidata mehr als 700 Werke aus ELTeC verzeichnen, erreicht die GND nur knapp über 200.



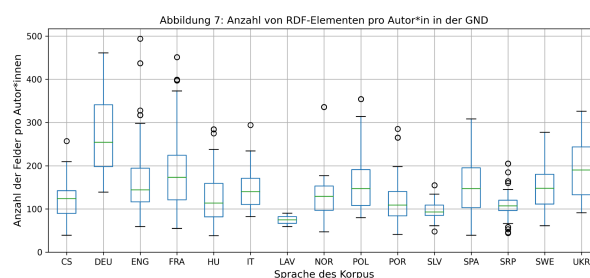
Um diese Zahlen besser zu verstehen, zeigt Abbildung 6, dass nur das deutsche Korpus akzeptable Ergebnisse aus der GND erreicht (80 % der Werke). Alle anderen Sprachen bewegen sich zwischen null und knapp über 40 %. Die Abdeckung von Wikidata oder VIAF ist für viele Sprachen deutlich höher: von 60 % bis zu 100 %.



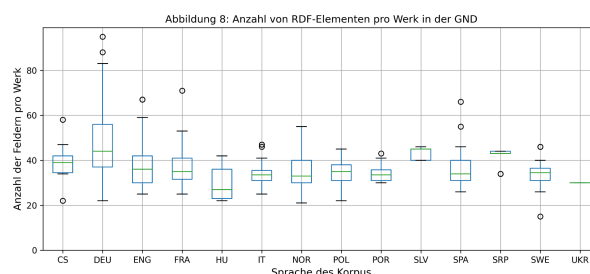
Anzahl der Informationen pro Entität

Entscheidend für eine Bewertung der Ressource ist nicht nur, ob eine Entität vorhanden ist, sondern wie gut sie mit Metadaten beschrieben ist. Für die GND ist zu erwarten, dass Entitäten aus dem deutschsprachigen Raum ausführlicher beschrieben werden als Entitäten aus anderen Regionen. Um dies zu messen, werden die

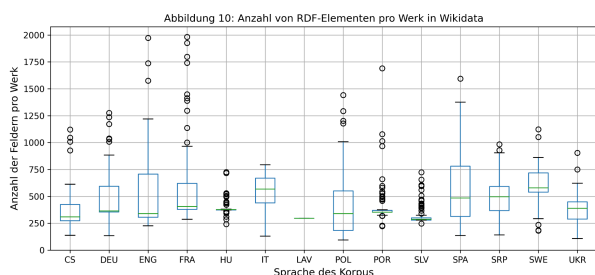
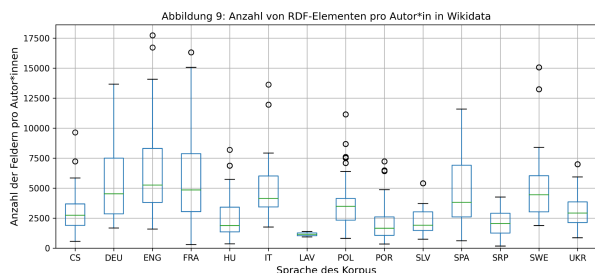
Daten von allen ELTeC-Autor*innen und deren Werke als XML-RDF-Dokumente aus der GND heruntergeladen und die Anzahl der XML-Elemente quantifiziert. Ohne den semantischen Gehalt der Elemente zu bewerten, gehen wir davon aus, dass mehr Elemente auch mehr Informationen pro Entität bedeuten. Abbildung 7 zeigt daher für die GND die Anzahl der Elemente pro Autor*in. Während für Autor*innen aus dem deutschsprachigen Raum 200-330 Elemente vorhanden sind, werden nur 100-240 für andere Sprachen verzeichnet. Auch wenn Sprachen wie Französisch oder Ukrainisch mittlere Werte (bis 240) zeigen, ist der Abstand zwischen diesen Sprachen und dem Deutschen immer noch sehr groß.



Für die Werke (vgl. Abb. 8) erreichen die deutschsprachigen Entitäten wieder deutlich höhere Werte als alle anderen Sprachen in der GND. Hier ist der Unterschied im Vergleich weniger groß, weil insgesamt für Werke weniger Elemente angelegt werden.

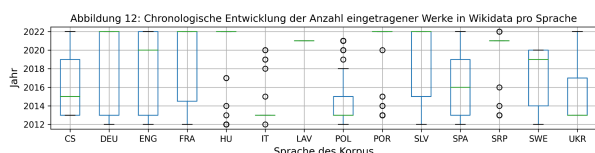
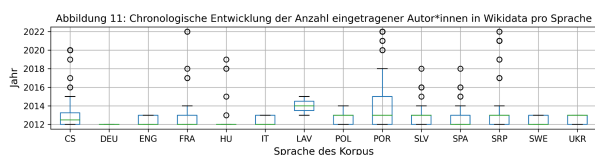


Für einen Vergleich wurden diese Daten auch aus Wikidata extrahiert (Abbildungen 9 und 10). Wir prüfen, ob in Wikidata ähnliche Verzerrungen gegenüber dem Englischen oder anderen Sprachen zu beobachten sind. Jedoch hat in Wikidata keine Sprache einen so klaren Vorsprung im Vergleich zu allen anderen Sprachen wie das Deutsche in der GND.



Anreicherung durch ELTeC

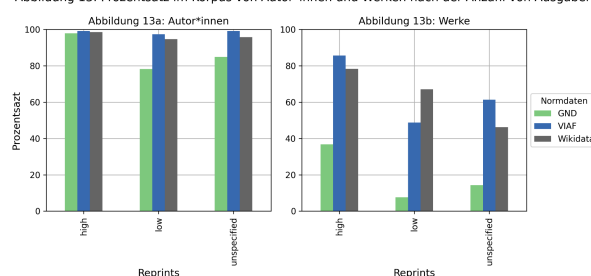
Insgesamt ist ein möglicher Grund für die bessere Abdeckung für Autor*innen und Werke in Wikidata und VIAF ist die Tatsache, dass die Teilnehmenden von ELTeC die Entitäten in Wikidata selbst eingetragen haben (Neßif u. a. 2022). Die Abbildungen 11 und 12 zeigen, dass zwar die Mehrheit der Autor*innen aus ELTeC bereits vor dem Start des Projekts (2017-2018) in Wikidata vorhanden waren, aber viele neue Werkeinträge entstanden. Die Metadaten zu Werken aus den sieben Sprachen, die Neßif u. a. (2022) ausgewählt haben, führen zu einer deutlichen Verbesserung der Abdeckung in den letzten Jahren.



Kanonisierungsgrad

Eine weitere Hypothese ist, dass der Kanonisierungsgrad die Abdeckung in den drei Ressourcen beeinflusst. In den ELTeC Korpora wurde dies anhand des Metadatenfelds "reprints" belegt. Für Autor*innen und Werke, die keine oder wenige "reprints" ("low") haben, zeigt Abbildung 13, dass alle drei Ressourcen eine niedrigere Abdeckung haben. Dabei ist der Unterschied für die GND deutlich größer als bei VIAF und Wikidata. Die Daten deuten darauf hin, dass die GND stärker vom Kanonisierungsgrad beeinflusst ist als die anderen zwei Ressourcen. Besonders niedrig ist die Abdeckung von nicht kanonisierten Werken in der GND (Abb. 13, "low reprints"). Zu beachten ist, dass die Verteilung von solchen Metadaten in den ELTeC Korpora nicht gleichmäßig ist. Die Ergebnisse können dementsprechend allein durch die Zusammensetzung der Korpora und die Metadatenanreicherung beeinflusst sein.

Abbildung 13: Prozentsatz im Korpus von Autor*innen und Werken nach der Anzahl von Ausgaben



Abschluss

Wie gut können nicht-germanistische Projekte aus dem deutschsprachigen Raum mit der GND Autor*innen und Werke identifizieren? Sollten sie lieber auf Wikidata oder VIAF zurückgreifen? Um das zu beantworten, wurden die multilingualen Korpora von ELTeC analysiert. Auch wenn diese Korpora nicht vollständig ausgewogen hinsichtlich Repräsentation der Inhalte und der Persistenz ihrer Identifier sein können, sind sie eine wertvolle Ressource von und für die Community. Weitere ähnliche Evaluationen könnten in Zukunft durchgeführt werden, wenn umfassenderes Vergleichsmaterial identifiziert oder zusammengestellt wird. Das ELTeC Korpus ist zeitlich (19. Jh.), quantitativ und sachlich (100 Sprache) notwendig beschränkt und vor diesem Hintergrund sind auch die vorliegenden Ergebnisse zu betrachten.

Generell zeigt die GND eine gute Abdeckung von Personendaten und ist damit sehr nah an Wikidata oder VIAF. Jedoch fällt die GND deutlich mehr Felder (d.h. mehr Informationen) zu deutschen Autor*innen als zu anderen europäischen Autor*innen (für Personendaten außerhalb der Literatur mag das anders aussehen).

Hinsichtlich der Werknormdaten kann die GND nur für das Deutsche akzeptable Ergebnisse liefern. Auch für andere große Sprachen wie Französisch, Englisch oder Spanisch enthält die GND nur 40 % der enthaltenen Werke in ELTeC. Nicht nur die Abdeckung, sondern auch der Informationsgehalt ist für Werke deutschspra-

chiger Autor*innen höher als für alle anderen Sprachen. Darüber hinaus scheint die GND stärker vom Kanonisierungsgrad abhängig zu sein als VIAF oder Wikidata.

Wenn die GND damit als national-ausgerichtete Normdateninstitution erwartbar schlechter abschneidet, dann wäre zu prüfen, ob die Normdaten anderer Einrichtungen (vor allem von Nationalbibliotheken) eine ähnliche oder sogar stärkere Favorisierung der eigenen Sprache verzeichnen.

Durch die Einbindung der GND (der DNB und GND-Agenturen in anderen Bibliotheken) in die NFDI öffnet sich die GND nicht nur der Community, sondern es werden auch wichtige Diskussionen zu Multilingualität (GNDmul)⁹ und Internationalität geführt. Wir sind zuversichtlich, dass die GND mithilfe der Community den Anteil fremdsprachiger Werke und Autor*innen in Zukunft erhöhen kann. Innerhalb von Text+ versuchen wir, Normdaten und Forschungsdaten zu verknüpfen. Die in dieser Analyse verwendeten Skripte werden auch für die Entwicklung von Pipelines zur Datenanreicherung im TextGrid Repository genutzt. So können die neu identifizierten Personen und Werke aus den ELTeC-Korpora mit den entsprechenden IDs zu den Daten aus dem TextGrid Repository hinzugefügt werden. Umgekehrt konnte damit auch ELTeC neue Daten für die Korpora gewinnen.

Fußnoten

1. <https://textgridrep.org/>.
2. <https://www.deutschestextarchiv.de/>.
3. Vgl. GND-Partner: https://gnd.network/Webs/gnd/DE/UeberGND/Partner/partner_node.html; <https://prezi.com/p/unl16mzwubbs/gndzoom/>.
4. Nationale Forschungsdateninfrastruktur: <https://www.nfdi.de/>.
5. Text+: <https://www.text-plus.org/>; NFDI4Culture: <https://nfdi4culture.de/>.
6. Von den 501.913 Einzelwerken in der GND sind 57.857 Deutsch, 10.476 Englisch und 6.918 Französisch, 5.918 Italienisch, 2.419 Spanisch, 449 Portugiesisch, 390 Rumänisch, aber auch 12.329 Latein, 6.072 Griechisch und der Großteil ohne Sprachangabe: 379.872 (<https://explore.gnd.network/search?f.satzart=Werk&f.land.limmit=80&rows=25>)
7. Vgl. Lobid-API, mit der GND-Einträge abgerufen werden können: <https://lobid.org/gnd/api>.
8. Das lettische Korpus (LAV) zählen wir nicht mit, weil der Balken in Abb. 4 zwar 100% für alle drei Ressourcen anzeigt, aber es sich insgesamt nur um 5 Romane handelt (vgl. Abb. 1).
9. https://gnd.network/Webs/gnd/DE/UeberGND/Partner/partner_node.html

Bibliographie

Barth, Florian, Varachkina, Hanna, Döncke, Tillmann, und Luisa Gödeke. Levels of Non-Fictionality in Fictional Texts. In *Proceedings of ISA-18 Workshop at LREC2022*, 27–32. Marseille, 20 June 2022. <http://www.lrec-conf.org/proceedings/lrec2022/workshops/ISA-18/pdf/2022.isa18-1.4.pdf>.

Balzer, Detlev, Barbara K. Fischer, Jürgen Kett, Susanne Laux, Jens M. Lill, Jutta Lindenthal, Mathias Manecke, Martha Rosenkötter, und Axel Vitzthum. 2019. „Das Projekt ‚GND für Kulturdaten‘ (GND4C)“. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 6 (4): 59–97. <https://doi.org/10.5282/o-bib/2019H4S59-97>.

Burnard, Lou, Christof Schöch, und Carolin Odebrecht. 2021. „In search of comity: TEI for distant reading“. *Journal of the Text Encoding Initiative*, Nr. Issue 14 (März). <https://doi.org/10.4000/jtei.3500>.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Digital Humanities Research 5. Bielefeld: transcript.

Dieckmann, Lisa, Jürgen Hermes, und Claes Neufeind. 2017. „Bild, Beschreibung, (Meta)Text. Automatisierte inhaltliche Erschließung und Annotation kunsthistorischer Daten“. In *Digitale Nachhaltigkeit*, 103–7. Bern: DHd. <https://zenodo.org/record/3684825#.YuoMEBzP1aQ>.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hecht, Christopher Kittel, Carsten Milling, und Peer Trilcke. 2019. „Programmable Corpora - Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor“. In *6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHd 2019, Frankfurt & Mainz, Germany, March 25-29, 2019*, herausgegeben von Patrick Sahle und Patrick Helling. <https://doi.org/10.5281/zenodo.4622061>.

Fischer, Frank, und Robert Jäschke. 2018. „Liebe und Tod in der Deutschen Nationalbibliothek“. In *DHd2018: „Kritik der digitalen Vernunft“*, 261–66. Cologne, Germany: Digital Humanities im deutschsprachigen Raum. <https://hal.archives-ouvertes.fr/hal-01787558>.

Henny-Krahmer, Ulrike. 2018. „Exploration of Sentiments and Genre in Spanish American Novels“. In *Puentes/Bridges*. México DF: ADHO. <https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/>.

Herrmann, J. Berenike, und Gerhard Lauer. 2018. „Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne“. *Osnabrücker Beiträge zur Sprachtheorie*, Nr. 92: 127–56.

Kett, Jürgen, Christoph Kudella, Andrea Rapp, Regine Stein, und Thorsten Trippel. 2022. „Text+ und die GND - Community-Hub und Wissensgraph“. *Zeitschrift für Bibliothekswesen und Bibliographie* 69 (1-2): 37–47. <https://doi.org/10.3196/1864295020691262>.

Nešić, Milica Ikonić, Ranka Stanković, Christof Schöch, und Mihailo Skoric. 2022. „From ELTeC Text Collection Metadata and Named Entities to Linked-Data (and Back)“. In *8th Workshop on Linked Data in Linguistics*, 7–16. Marseille: LREC. <http://www.lrec-conf.org/proceedings/lrec2022/workshops/LDL/pdf/2022.lidl2022-1.2.pdf>.

Odebrecht, Carolin, Lou Burnard, und Christof Schöch. 2021. „European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels.“ Zenodo. <https://doi.org/10.5281/zenodo.4662444>.

Rosenkötter, Martha, und Barbara Fischer. 2020. „Normdaten der Faktenanker für Qualität im semantischen Retrieval. Der Ausbau der Gemeinsamen Normdatei (GND) im Projekt GND für Kulturdaten (GND4C)“. In *Spielräume: Digital Humanities zwischen Modellierung*

und Interpretation , 344–45. Paderborn: DHd. <https://zenodo.org/record/3666690#.YuoNIRzP1aQ>.

Schöch, Christof, Tomaz Erjavec, Roxana Patras, und Diana Santos. 2021. „Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives“. *Modern Languages Open*, Mai. <https://doi.org/10.5281/ZENODO.4742419>.

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer, und Stefanie Popp. 2019. „The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML“. *Journal of the Text Encoding Initiative*, August. <https://doi.org/10.4000/jtei.2085>.

Grenzen der Offenheit: eine digitale Sammlung zur Erforschung historischer Arzneimittelrezepte

Dinger, Patrick

patrick.dinger@uni-muenster.de
Westfälische Wilhelms-Universität Münster,
Deutschland

Horstmann, Jan

jan.horstmann@uni-muenster.de
Westfälische Wilhelms-Universität Münster,
Deutschland

Schellhammer, Stefan

Stefan.Schellhammer@wi.uni-muenster.de
Westfälische Wilhelms-Universität Münster,
Deutschland

Troglauer, Patrick

patrick.troglauer@wi.uni-muenster.de
Westfälische Wilhelms-Universität Münster,
Deutschland

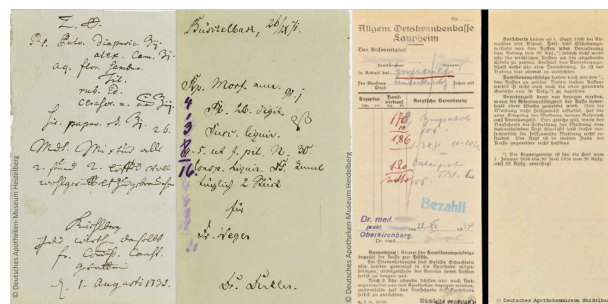
Einleitung

Das BMBF-geförderte Forschungsprojekts „ArlS – Durch das Artefakt zur ‘infra structura’“¹ hatte das Ziel, durch Erschließung von Arzneimittelrezepten einen erkenntnisermöglichenden Zugang zur Entstehung der Gesundheitsinfrastruktur in Deutschland und Österreich zu generieren.² Im Fokus der Forschungsaktivität des Projekts lag das Vorhaben, über historische Rezepte Ein-

blick in die Entwicklungsgeschichte des deutschen Gesundheitssystems zu erhalten. Die auf ihnen zu findenden Spuren bieten Hinweise auf kleinere und größere Veränderungen ihrer historischen Umwelt, wie bspw. die Entstehung der Krankenkassen oder neue Abrechnungsformen. Da auf keine vorhandenen Daten dieser Art zurückgegriffen werden konnte, wurden erstmals historische Arzneimittelrezepte in größerem Umfang digitalisiert und der Versuch unternommen, diese artefakt-individuell zu untersuchen und zu beschreiben. Gleichsam als Nebenprodukt entstanden bei der digitalen Erschließung Scans der Rezeptzettel, welche die Möglichkeit für eine weitere inhaltliche Erfassung im Rahmen von Anschlussforschung bieten. Das Bestreben, die Datenbank inkl. Scans als eine den FAIR-Prinzipien³ entsprechende Datensammlung hierfür zu veröffentlichen, stößt jedoch auf einige Hindernisse.

Das Projekt war eine Kooperation zwischen Wirtschaftsinformatiker*innen und Pharmaziehistoriker*innen der Universitäten Aachen, Münster und Marburg sowie dem Deutschen Apotheken-Museum Heidelberg und wurde über vier Jahre (2018–2022) gefördert.⁴ Trotz des von den Projektpartner*innen vertretenen Anspruchs an Offenheit musste an mehreren Stellen des Projekts aus unterschiedlichen Gründen von diesem Prinzip abgewichen werden. Dies ist sowohl auf projektspezifische Forschungsinteressen als auch rechtliche Einschränkungen zurückzuführen, die im Folgenden erörtert werden.

Obwohl es sich bei dem Arzneimittelrezept um einen jahrhundertalten, millionenfach ausgestellten Gegenstand der Gesundheitsversorgung im alltäglichen Leben handelt, wurde dieser bislang kaum als erhaltenswert wahrgenommen. Im ArIS-Projekt ist nun eine einmalige Datenbank mit digitalisierten Rezeptblättern aus mehreren Sammlungen unterschiedlicher Herkunft mit über 12.200 Datensätzen entstanden.



Beispiele historischer Arzneimittelrezepte (Bildinhaber: Deutsches Apotheken-Museum Heidelberg)

Mit Blick auf den typischen Lebenszyklus von Forschungsdaten⁵ und den Anspruch von Open Science (vgl. Heise 2018) lassen sich bei einem interdisziplinären Projektvorhaben dieser Größenordnung unterschiedliche Herausforderungen beobachten: (1) Der große historische Rahmen, aus dem die Objekte stammen, wirft in der Modellierung die Frage nach einer geeigneten Definition des Terminus „Rezept“ auf. (2) Methodisch zeigt sich bei großen, heterogenen Handschriftenkorpora die Notwendigkeit, traditionelle Erschließungsmethoden durch visuelle Analysen zu ergänzen. (3) Neben

rechtlichen Herausforderungen bei Open Data, wie Urheber- und Verwendungsrechten, sind bei Gesundheitsdaten besondere Schutzbedürfnisse und Fristen zu wahren, die nach einem Konzept für eine dynamische Öffnung von Daten verlangen.

Die verschiedenen Möglichkeiten der inhaltlichen Erschließung bei herausfordernder Rechtslage sind die zentralen Aspekte sowohl des Forschungsprojekts ArIS als auch des vorliegenden Beitrags.

Offene Supportstrukturen für das ArIS-Projekt

Ziel des ArIS-Projekts war es, die Rolle und Bedeutung des Arzneimittelrezepts in der Entstehung des modernen Gesundheitswesens zu untersuchen. Zu diesem Zweck sollte zunächst eine empirische Grundlage geschaffen werden, indem vorhandene Sammlungen von Schriftstücken (meist Zettel) mit Rezepten unterschiedlicher Standorte identifiziert, digitalisiert und virtuell zusammengeführt wurden.

Zur Erschließung, Modellierung, Analyse und Langzeitarchivierung der im Projekt entstehenden Forschungsdaten griffen die Münsteraner Projektpartner aus der Wirtschaftsinformatik auf die Kompetenzstrukturen der eigenen Institution zurück und erarbeiteten zusammen mit der Universitäts- und Landesbibliothek und ihren Service Centern for Data Management⁶ und for Digital Humanities⁷ ein strukturiertes Vorgehen.⁸ Die für die Forschenden an der Universität Münster frei und dauerhaft zur Verfügung stehenden Supportstrukturen ermöglichten sowohl eine enge und nachhaltige Zusammenarbeit als auch eine Orientierung an Standardformaten. Impulse für die Modellierung und den Einsatz von Methoden des maschinellen Lernens (vgl. Bönisch 2022) konnten so frühzeitig und während der gesamten Projektdauer aufgenommen werden.

Zu Beginn lag ein Großteil der physischen Artefakte als ungesichtete Blattsammlungen in Kartons vor, zu denen nur in Teilen Metadaten erfasst waren. Daten auf Artefaktebene oder auch Aussagen über mögliche Ordnungsprinzipien in Bezug auf Kartons oder andere physische Objektklammern gab es i.d.R. nicht. Die Digitalisate bisher nicht archivierter Schriftstücke wurden daher mit Inventarnummern versehen, die Rückschlüsse auf Ort, Lagerform und Blattnummer zulassen. Auf diese Weise sollte die Wiederauffindbarkeit der physischen Blätter an den jeweiligen (musealen) Standorten gewährleistet sein, um einem durch die Digitalisierung möglicherweise entstehenden Informationsverlust (Bindung, Ordnungsprinzipien) entgegenzuwirken. Anschließend wurden die Digitalisate inklusive der bereits existierenden Metadaten in eine easydb-Datenbank importiert.⁹ Die an der ULB Münster betriebene Datenbank erlaubt eine flexible Gestaltung des Datenmodells, sodass auf die Anforderungen des interdisziplinären Forschungs- und Erschließungsprojekts eingegangen werden konnte. Geschaffen wurde so eine standardisierte Datenbank, in welcher die heterogenen Bestände unterschiedlicher Provenienz zusammengeführt, erfasst und verwaltet werden.

Durch die neu gewonnene digitale Verfügbarkeit der Artefakte sowie das flexible Rechtemanagement der ULB Münster konnten Teilsammlungen, wie beispielsweise die des Deutschen Apotheken-Museums in Heidelberg, von ausgewiesenen Fachwissenschaftler*innen erschlossen werden. Zugänglichkeit und Multibenzutzer-Betrieb der Forschungsdatenbank ermöglichten es den Projektpartner*innen des ArIS-Projekts asynchron und ohne Konflikte in den Daten zu erzeugen, ihr fachspezifisches Wissen bei der Erschließung einzubringen.

Standardisierte Metadatenmodellierung und projektspezifische Erweiterungen

Bei der Erschließung digitalisierter Sammlungen der Universität Münster liegt der Fokus auf einer standardisierten Erfassung der Metadaten, der Nutzung etablierter Datenformate wie LIDO oder METS/MODS sowie der Anbindung von Normdaten und kontrollierten Vokabularen (vgl. DFG-Praxisregeln). Nur unter Beachtung dieser Voraussetzungen ist die Entstehung von interoperablen (*interoperabel*) und wiederverwendbaren (*reusable*) Forschungsdaten gewährleistet. Bei der Bildung der umfassenden Datensammlung (vgl. Schöch 2017) wurden Sammlungen von Arzneimittelrezepten mit heterogenem Inventarisierungs- und Kuratierungsstatus zusammengeführt. Daraus ergibt sich die konzeptionelle Frage, ob der standardisierte Erschließungsworkflow den Anforderungen der vorliegenden Sammlung überhaupt gerecht werden kann (vgl. Dörk und Glinka 2018).

In enger Kooperation zwischen dem Teilprojektteam in Münster und der ULB Münster (Service Center for Data Management; SCDM) wurde entschieden, anstelle einer standardisierten Erschließung der einzelnen Rezepte zunächst eine flexible Modellierung der vorhandenen Bild- und Objektdaten anzustreben. Hintergrund ist, dass anders als in unserer Alltagsvorstellung ein Digitalisat bzw. Datensatz nicht zwangsläufig mit einem einzelnen Arzneimittelrezept auf einer Seite eines Blattes gleichzusetzen ist. Das Rezept als abstraktes Konzept, das erst in einem bestimmten sozialen Kontext seine Form und Definition erhält, ist vielmehr unabhängig von einer bestimmten physischen Repräsentation. Heutzutage wird unter einem Arzneimittelrezept die formelle, schriftliche Aufforderung von Ärzt*innen an Apotheker*innen zur Abgabe von Arzneimitteln an eine/n bestimmte/n Patient*in verstanden. Inwiefern dieses Begriffsverständnis und dessen definitorische Merkmale aber auf historische Rezepte übertragbar sind, ist Teil eines pharmaziehistorischen Forschungsdiskurses (vgl. Seidel 1977, 22f.) und soll an dieser Stelle nicht weiter ausgeführt werden. Für den Projektkontext ist jedoch bedeutsam, dass die einzelnen Digitalisate manuell dahingehend überprüft werden mussten, ob sie Rezeptinformationen enthalten.

Für die Modellierung der Daten wurden die durch den externen Dienstleister dokumentierten Informationen zu den Artefakten in die Datenbank übernommen, wie z.B. die Klammerung oder Bindungen der Blätter. Zu diesem Zeitpunkt war jedoch unklar, ob den wenigen ermittelten Informationen ein historisch relevantes Ordnungsprinzip

zu Grunde lag oder ob es sich um zufällige Gegebenheiten handelte. Gleichwohl erschien die Abbildung dieser Parameter als ein sinnvoller erster Schritt, eine Ordnung für die Daten und damit einen Ausgangspunkt für die Analyse zu finden.

Alle vorhandenen Rezeptblätter wurden formal erschlossen, kategorisiert, datiert und zu einer übergreifenden Datenbank zusammengeführt. In einem zweiten Schritt können einzelne Arzneimittelrezepte mit spezifischen Fragestellungen aus dem Bereich der Medizin- und Pharmaziegeschichte nach den etablierten Standards des Sammlungsmanagements¹⁰ erschlossen werden. Zugleich können auf Grundlage der entstandenen Datenbasis zentrale Fragen wie die Entwicklung der Arzneimittelrezepte anhand der identifizierten, übergreifenden Kriterien nachgezeichnet werden.

Inhaltliche Erschließung als Herausforderung und Voraussetzung

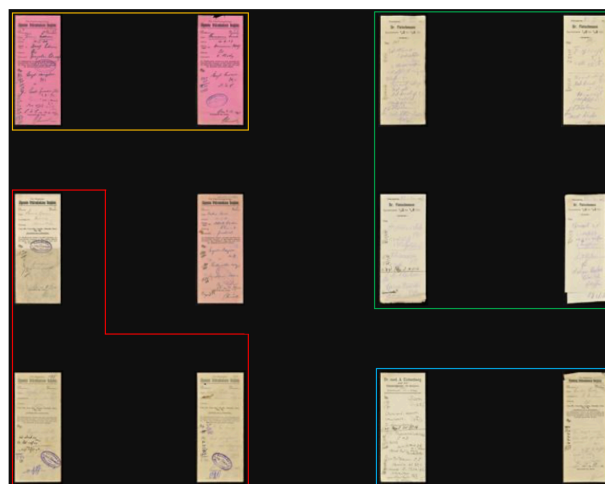
Die Erschließung und Erfassung von grundlegenden Metadaten der losen Zettelsammlung ist ein enormer Gewinn für die Forschung, da hier die Existenz möglicher Quellen nachgewiesen wird. Ähnlich wie bei vergleichbaren Archivalien entsteht so eine Art Zusammenführung, die einen Überblick über bestimmte Informationen eines größeren Zeitraums zulässt, in Bezug auf den Inhalt der Rezepte jedoch noch wenig Aussagekraft hat und damit zwar eine gute, aber keine hinreichende Quelle darstellt. Die Forschungsfragen etwa nach den verwendeten Arzneimitteln/Präparaten beginnen aber erst auf einer feingranulareren Ebene der Modellierung, wenn auch die Inhalte identifiziert werden.

Um die Findbarkeit von Inhalten der Rezepte zu erhöhen (und anschließende Analysen zu ermöglichen), wäre eine vollständige Maschinenlesbarkeit der Rezepte optimal. Die Möglichkeiten der inhaltlichen Erschließung stoßen in diesem Fall jedoch an methodische Grenzen: Die Arzneimittelrezepte sind in der Regel vollständig handschriftlich geschrieben oder zumindest handschriftlich ausgefüllt. Es finden sich zu viele unterschiedliche Handschriften und sprachliche Besonderheiten (Abkürzungen, Apothekerlatein), die nach aktuellem Stand der algorithmischen Möglichkeiten eines Handwritten-Text-Recognition-Ansatzes¹¹ nicht trainiert werden können, sondern händisch transkribiert werden müssten. Hierfür fehlten im Projekt die zeitlichen und personellen Kapazitäten. Bislang wurde daher lediglich ein Subset der Rezepte im Rahmen eines geplanten Webauftritts des Deutschen Apotheken-Museums Heidelberg vollständig transkribiert.

Konfrontiert mit diesen Herausforderungen wurde im Projekt nach alternativen Wegen zur Überschreitung der Objektebene gesucht. Die Interpretation der losen Blattsammlung als Kollektion visueller Artefakte erschien dabei vielversprechend. Mit Methoden der Computer Vision (vgl. z.B. Arnold und Tilton 2019; He et al. 2016; Redmon et al. 2016) lässt sich das Korpus als Ganzes in seiner visuellen Ausprägung analysieren und z.B. mithilfe einer Mustererkennung clustern. Denkbar wäre hier

etwa die Suche nach verwendeten Stempeln, Schriftfarben/Schreibutensilien oder standardisierten Vorstrukturierungen der Rezeptzettel, anhand derer sich eine über die Zeit zunehmende Formalisierung nicht nur der Behandlung, sondern auch der damit einhergehenden Kommunikation der Akteure feststellen ließe.

Eine Herausforderung ist hierbei, dass die Rezeptzettel nicht mit dem Ziel einer automatisierten, visuellen Analyse eingescannt wurden, sodass häufig etwa Kriterien wie geknickte Ecken oder Verfärbungen fälschlicherweise zu einem Clustering führen, das für die inhaltsbezogene Forschungsfrage keine Aussagekraft hat. In einem verhältnismäßig aufwändigen Preprocessing müssten die Scans daher vor einer eingehenden Computer-Vision-Analyse einzeln bereinigt und normiert werden, um anhand inhaltlicher Kriterien aussagekräftig clustern zu können. Erste Schritte wurden hierfür mit dem am DHLab Yale entwickelten Tool PixPlot (vgl. Duhaime 2019) gemacht. Im Gegensatz zu eigentlich intendierten Anwendungsszenarien diente die Software dazu, im Zusammenspiel mit easydb eine Gesamtschau auf die Daten zu kreieren, die mit der Betrachtung einzelner Bilddateien in der Datenbank nicht erreicht werden konnte.



Clusterbildung ausgewählter Arzneimittelrezepte mit PixPlot

Das Tool erlaubt es, die Rezeptzettel nach Ähnlichkeit oder auch anhand einer Zeitachse zu clustern. Durch das Zusammenspiel menschlicher und maschineller Analyse können darüber hinaus die identifizierten Erscheinungsformen der Arzneimittelrezepte bestätigt, auf visueller Ebene evaluiert und Datenanomalien erkannt werden. Obwohl diese Verknüpfung von Datenbank und visueller, software-gestützter Analyse bei der Erschließung der Daten nur erprobt werden konnte, erscheint sie als vielversprechendes Instrument, um traditionelle objektbasierte Methoden der Erschließung zu ergänzen.

Die inhaltliche Erschließung der Arzneimittelrezepte würde nicht nur die Findbarkeit der Daten enorm erhöhen, sie wäre auch Grundvoraussetzung, um die im Folgenden beschriebenen rechtlichen Grenzen der Offenheit überwinden zu können und die Daten vollständig *open access* zugänglich zu machen.

Rechtsebene: Herausforderungen und Grenzen der offenen Datenmodellierung

Das ArlS-Projekt wurde Ende August 2022 beendet. Die entstandenen Daten bilden eine vielversprechende Grundlage für weitere Forschung. Die Objekte wurden im Projekt auf Einzelartefaktebene erschlossen, sodass bestimmt werden konnte, für welche Jahre Arzneimittelrezepte vorliegen und wo die physischen Artefakte jeweils auffindbar sind. Ein Subset der Artefakte konnte bereits vollumfänglich transkribiert, d.h. inhaltlich erschlossen werden.

In Bezug auf die Nutzbarkeit von Daten stehen üblicherweise Fragestellungen zu Urheber- und Verwendungsrechten im Vordergrund. Im vorliegenden Fall sind allerdings weitergehende Schutzrechte zu beachten, wie das im Deutschen Reich definierte besondere Schutzrecht für Patientendaten (vgl. Deutsches Reichsgesetzblatt 1871, § 300). Das genaue Ende von Schutzfristen für in Arzneimittelrezepten vorkommende Daten ist juristisch nicht eindeutig zu klären.¹² Hinzugezogene Experten haben dem Projekt geraten, dass Arzneimittelrezepte bis Ende des Jahres 1871 unbedenklich zugänglich gemacht werden können. Bei Rezepten, die ab dem Jahr 1872 ausgestellt wurden, müssten identifizierende Merkmale von Ärzt*innen oder Patient*innen unkenntlich gemacht werden. Dieser Umstand hatte besonderen Einfluss auf den Projektverlauf und offenbart eine zentrale Herausforderung für offene Forschungsdaten und ihre Langzeitverfügbarmachung. So musste hier zweifelsfrei sichergestellt werden, dass alle vorliegenden Digitalisate jahresgenau datiert wurden. Bei mehreren oder teilweise unleserlichen Datumsangaben wurde grundsätzlich das jüngste Datum als Datenbankeintrag gewählt.

Eine Ausgangslage des Projekts ist somit die Zwickmühle, dass die Güte einer automatisierten Erschließung hoch sein muss, um eine inhaltliche Aussage zu treffen – und bis dahin können auch nicht beispielsweise lediglich die Scans der nicht-transkribierten Zettel öffentlich zur Verfügung gestellt werden, da so potentiell personenrelevante Daten offengelegt würden. Da die aktuellen Methoden hier nicht ausreichende Ergebnisse liefern, können auch personenrelevante Daten nicht leicht bereinigt werden.

Die Existenz von langen Schutzfristen über mehrere Jahrzehnte hat damit zur Folge, dass eine öffentlich zugängliche Archivierung nur mit Schwärzung zentraler Merkmale der Artefakte möglich wäre. Hierfür wäre wiederum die Transkription sämtlicher Arzneimittelrezepte Voraussetzung, um etwa automatisiert nach benannten Entitäten suchen zu können. Alternativ könnte man nur einen eng umrissenen Teil der Sammlung zugänglich machen. Weiterhin ist zu beachten, dass mit fortschreitender Zeit Schutzfristen für bestimmte Jahrgänge der Artefakte aufgelöst werden müssten. Diese Dynamik in der Möglichkeit, Daten zugänglich zu machen, ist eine besondere Herausforderung für Open Data in einer projektbasierten Forschung. Auch dauerhaft bereitstehende Supportstrukturen können hier häufig keine Abhilfe schaffen. Dazu kommen Aspekte des ethischen

Umgangs mit Kulturgut. Neben einer möglichst umfangreichen Umsetzung der FAIR-Prinzipien sollen die CARE-Prinzipien nicht minder Beachtung finden (vgl. Research Data Alliance 2019).

Im Sinne des Projekts wurde ein mehrstufiges Vorgehen gewählt, das dem Museum eine zentrale Rolle einräumt: Die Datenbank wird als „work-in-progress“ zum Projektende an das Deutsche Apotheken-Museum übergeben. Durch die initiale Analyse aller Artefakte ist es möglich, Teile des Datenschatzes zugänglich zu machen und diesen Teil sukzessive auszuweiten. Damit wird nicht nur den Schutzrechten Rechnung getragen, sondern auch sichergestellt, dass die physischen Artefakte weiterhin für Forscher*innen zugänglich bleiben. Vollständig erfasste und transkribierte Artefakte können sukzessive, vorausgesetzt die Schutzfristen entfallen, ebenfalls öffentlich zugänglich gemacht werden.

Niemand kann antizipieren, welche Forschungscommunitys sich in Zukunft mit welchen Fragen an bestimmte Datenschätze wenden. Das ist ein Grundprinzip des Forschens, auf das die FAIR-Prinzipien produktiv antworten. Die öffentliche Zugänglichkeit nützt damit der gesamten Forschung und macht vor allem die erhobenen Daten zukunftsfähig. Unser Verständnis von FAIR und Open Access ist, dass hier der normative Anspruch erhoben wird, zunächst keine Zugangsschranken zu erstellen. Dieser Anspruch muss vor anderen Rechtsgütern verantwortungsvoll ausgehandelt werden. Es gilt daher ganz im Sinne der European Commission (2021, 61) der Grundsatz: „as open as possible, as closed as necessary“.

Wie diese Balance aussehen sollte, ist, wie das Beispiel der Daten der Arzneimittelrezepte zeigt, nicht einfach zu beantworten. Es zeigt sich, dass eine gesellschaftliche Auseinandersetzung mit dieser Frage hilfreich wäre, um gegebenenfalls rechtliche Rahmenbedingungen neu zu bewerten. Für Museen ergibt sich mit diesem Aspekt der Digitalisierung die neue Herausforderung, wie sie einerseits dem Anspruch des Open Access genügen können und andererseits den Schutz der Daten vor Zugriff durch rechtlich Unbefugte gewährleisten. Dieser Beitrag veranschaulicht jene Herausforderung, ohne eine eindeutige Empfehlung aussprechen zu können.

Fußnoten

1. Vgl. <https://www.sprache-der-rezepte.de/> (zugriffen: 28. Juli 2022).
2. CRediTs: Patrick Dinger (Data curation, Writing – original draft, Writing – review & editing), Jan Horstmann (Conceptualization, Writing – original draft, Writing – review & editing), Stefan Schellhammer (Conceptualization, Writing – review & editing), Patrick Troglaier (Data curation, Formal Analysis, Visualization, Writing – review & editing).
3. Vgl. <https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/faire-daten/> (zugriffen: 28. Juli 2022).
4. Die fachwissenschaftliche Perspektive der vorliegenden Einreichung ist durch die Projektpartner im Deutschen Apotheken-Museum Heidelberg und der Universität Marburg gegeben. Auch wenn der Beitrag nur von einem Teil des Teams bestritten wird, bildet ihre Perspektive einen zentralen Grundstein des gesamten

Projekts, seiner Ergebnisse und Vorgehensweisen. So wurden Forschungsergebnisse beispielsweise auf pharmaziehistorischen Kongressen wie der Pharmaziehistorischen Biennale der Deutschen Gesellschaft für Geschichte der Pharmazie 2021 oder in Fachzeitschriften wie der Deutschen Apotheker Zeitung präsentiert und mit Fachwissenschaftler*innen diskutiert (vgl. Avci et al. 2020, Avci et al. 2021).

5. Vgl. <https://www.forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/> (zugegriffen: 28. Juli 2022).

6. Vgl. <https://www.uni-muenster.de/Forschungsdaten/> (zugegriffen: 28. Juli 2022).

7. Vgl. <https://www.uni-muenster.de/DH/scdh> (zugegriffen: 28. Juli 2022).

8. Grundlage des Projekts bilden 12.200 Datensätze (teilweise mit Scans der Vorder- und Rückseite) in unterschiedlichen Ausprägungsformen von sieben Standorten vom 16. bis 21. Jahrhundert, welche durch einen spezialisierten Dienstleister zu Beginn der Projektlaufzeit im TIF-Format digitalisiert wurden.

9. Vgl. <https://www.programmfabrik.de/easydb/> (zugegriffen: 28. Juli 2022) und mit Bezug auf Forschungsdaten z.B. Kloppmann und Kastner 2020.

10. Neben etablierten Metadatenformaten, kontrollierten Vokabularen und Normdaten ist auch der SPEC-TRUM-Standard für die Dokumentation und Verwaltung eines Objekts zu beachten. Zur deutschen Fassung des Standards siehe https://wissenschaftliche-sammlungen.de/de/service-material/materialien/dokumentationsstandard-spectrum-auf-deutsch-2013?pk_campaign=Newsletter-2013-04 (zugegriffen: 28. Juli 2022).

11. Vgl. zu Neuerungen in diesem Bereich z.B. Tomasek, Reul und Wehner 2022.

12. Projekterterne Experten haben darauf aufmerksam gemacht, dass in § 203 StGB nach wie vor nicht zum Ausdruck gebracht wird, wann die nach dem Tod der/des Patient*in weiter bestehen bleibende Geheimhaltungspflicht endet. Das spiegelt sich in den Archivgesetzen wider, die auch nach Ablauf der personenbezogenen Schutzfrist von zehn Jahren nach dem Tode der im Dokument herausgehobenen Personen und der Geheimhaltungsschutzfrist von 60 Jahren seit der Entstehung des Archivguts eine weitere Einschränkung oder sogar Versagung der Benutzung ermöglichen.

Bibliographie

Arnold, Taylor und Lauren Tilton. 2019. "Distant Viewing: Analyzing Large Visual Corpora." *Digital Scholarship in the Humanities* 34 (1). <https://doi.org/10.1093/llc/fqz013>.

Avci, Meral, Kerstin Grothusheitkamp, Patrick Troglauer, Stefan Schellhammer, Christoph Friedrich, Elisabeth Huwer und Barbara Simon. 2020. "Vom analogen zum digitalen Arzneimittelrezept. Eine lange Transformationsgeschichte." *Deutsche Apotheker Zeitung* 43 (22.10.2020), 78–79. Stuttgart.

Avci, Meral, Kerstin Grothusheitkamp, Stefan Schellhammer, Patrick Troglauer. 2021. "Die Rolle der Armen bei der Entstehung des deutschen Gesundheitssystems." Poster. *Pharmaziehistorische Biennale 2021: Heilpflanzen im Wandel der Zeiten* (08.–10.12.2021). Detmold.

URL: https://www.sprache-der-rezepte.de/sites/sprache-der-rezepte.de/files/attachments/poster_final.pdf (zugegriffen: 06. Dezember 2022).

Bönisch, Dominik. 2022. "Training the Archive – Von der maschinellen Exploration musealer Sammlungsdaten zur Curator's Machine." *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum"* (DHd 2022). Potsdam. <https://doi.org/10.5281/zenodo.6327949>.

Deutsches Reichsgesetzblatt. 1871. (24), 127–205. [https://de.wikisource.org/wiki/Strafgesetzbuch_f%C3%BCr_das_Deutsche_Reich_\(1871\)](https://de.wikisource.org/wiki/Strafgesetzbuch_f%C3%BCr_das_Deutsche_Reich_(1871)) (zugegriffen: 28. Juli 2022).

DFG-Praxisregeln "Digitalisierung" [12/16]. https://www.dfg.de/formulare/12_151/ (zugegriffen: 28. Juli 2022).

Dörk, Marian und Katrin Glinka. 2018. "Der Sammlung gerecht werden: Kritisch-generative Methoden zur Konzeption experimenteller Visualisierungen." *DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum"* (DHd 2018). Köln. <https://doi.org/10.5281/zenodo.4622364>.

Duhaime, Douglas. 2019. PixPlot. <https://github.com/YaleDHLab/pix-plot> (zugegriffen: 22. Juli 2022).

European Commission, Directorate-General for Research and Innovation. 2021. "Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)." Version 1.0, Brussels. <https://doi.org/10.2777/935288>.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren und Jian Sun. 2016. "Deep residual learning for image recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. Las Vegas. <https://doi.org/10.1109/CVPR.2016.90>.

Heise, Christian. 2018. *Von Open Access zu Open Science: Zum Wandel digitaler Kulturen der wissenschaftlichen Kommunikation*. Lüneburg: meson press. <https://doi.org/10.14619/1303>.

Kloppmann, Jens und Charlotte Kastner. 2020. "easydb. Flexibles Framework zum Aufbau von Metadaten- und Medienrepositorien. Anwendungsfall: Forschungsdaten." *Programmfabrik GmbH. Digital Summer School 2020 der SUH (UB Hildesheim)*. <https://doi.org/10.5281/zenodo.3937554>.

Research Data Alliance International Indigenous Data Sovereignty Interest Group. (2019). "CARE Principles for Indigenous Data Governance." *The Global Indigenous Data Alliance*. <https://static1.squarespace.com/static/5d3799de845604000199cd24/t/637acb53881a0973324d18b-f/1668991830292/Die+CARE-Prinzipien+f%C3%BCr+indigene+Data+Governance.pdf> (zugegriffen: 24. November 2022).

Redmon, Joseph, Santosh Divvala, Ross Girshick und Ali Farhadi. 2016. "You only look once: unified, real-time object detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788. Las Vegas. <https://doi.org/10.1109/CVPR.2016.91>.

Schöch, Christof. 2017. "Aufbau von Datensammlungen." In: *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein, Stuttgart: Metzler, 223–233. https://doi.org/10.1007/978-3-476-05446-3_16.

Seidel, Ulrich. 1977. "Rezept und Apotheke. Zur Geschichte der Arzneiverordnung vom 13. bis zum 16. Jahrhundert." Naturwissenschaftliche Diss. Marburg.

Tomasek, Stefan, Christian Reul, und Maximilian Wehner. 2022. "Handwritten Text Recognition und Word Mover's Distance als Grundlagen der digitalen Edition 'Die Kindheit Jesu Konrads von Fußesbrunnen'." *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum"* (DHd 2022). Potsdam. <https://doi.org/10.5281/zenodo.6328199>.

How to Open Heritage? Digitale Erschließungskonzepte für Provenienzforschung am Museum für Naturkunde Berlin

Wagner, Sarah

sarah.wagner@mfn.berlin

Museum für Naturkunde Berlin – Leibniz-Institut für Biodiversitätsforschung, Deutschland

Dubova, Alona

alona.dubova@mfn.berlin

Museum für Naturkunde Berlin – Leibniz-Institut für Biodiversitätsforschung, Deutschland

Marquart, Aron

aron.marquart@mfn.berlin

Museum für Naturkunde Berlin – Leibniz-Institut für Biodiversitätsforschung, Deutschland

Das Forschungscluster Open Heritage und seine Zielsetzungen

Das Museum für Naturkunde Berlin (MfN) verwahrt mehr als 30 Millionen Objekte aus Zoologie, Paläontologie, Geologie und Mineralogie aus allen Teilen der Erde. Dass die gesamte Sammlung und das mit ihr verbundene Wissen der Öffentlichkeit zugänglich sein soll, ist Gründungsgedanke und mit dem Zukunftsplan¹ konkreter Auftrag dieses Forschungsmuseums der Leibniz-Gemeinschaft. Neben baulichen Sanierungsmaßnahmen bildet die digitale Erschließung der Sammlung einen Teil dieses Auftrags.

Die Dokumentation der Objekte am MfN fand bislang überwiegend nach biologischen Kriterien statt. Seit ei-

nigen Jahren aber erfahren Informationen zu Herkunft und Erwerbsumständen eine verstärkte Aufmerksamkeit, nicht zuletzt angesichts der zunehmenden gesellschaftlichen Debatte um die Verantwortung der Museen, Unrechtskontexte aufzuarbeiten, Menschen aus den Herkunftsgesellschaften einen Zugang zu ihrem kulturellen Erbe zu ermöglichen und nicht zuletzt die Rückgabe unrechtmäßig erworbener Kultur- und auch Naturgüter neben dem Sammeln, Bewahren, Forschen, Ausstellen und Vermitteln als Teil ihrer Aufgabenbereiche zu verstehen.² Vor diesem Hintergrund wurde das Forschungscluster „Open Heritage – Naturkunde in globalen Kontexten. Sammlung erforschen, Zukunft gestalten“³ am MfN ins Leben gerufen. Kernaufgabe ist es Strategien und Werkzeuge zu entwickeln, um museale Sammlungen zukünftig als nachhaltige und global zugängliche Wissensressourcen bereitzustellen. Dabei stehen Fragen der Öffnung, Mehrstimmigkeit, Ermöglichung von Teilhabe sowie der Zukunft von musealen und archivalischen Räumen im Fokus, ebenso die kritische Reflexion vergangener, gegenwärtiger und zukünftiger Forschungs-, Sammlungs- und Dokumentationspraktiken der Naturkunde und -geschichte. Das Cluster vereint inter- und transdisziplinär ausgerichtete Projekte aus verschiedenen Forschungsbereichen des Museums, die sich mit kultur- und sozialwissenschaftlichen Fragestellungen der Erforschung, Erschließung und Reflexion der Sammlung im globalen Kontext nähern. Die digitale Erschließung und Bereitstellung von Sammlungsinformationen können einen niedrigschwelligen Zugang, Transparenz, neue Ordnungssysteme und Analyseformen, faire und inklusive Kooperationen über Disziplinen und Institutionen und gesellschaftliches Engagement hinweg bedeuten. Dabei müssen Forschungs-, Entscheidungs- und Übersetzungsprozesse zwischen Informations- und Datenwissenschaften, Katalogisierungs- und Ordnungssystemen, Verschlagwortungen, Thesauri sowie Repräsentationsformaten sowohl in digitalen Datenbanken als auch in der Kuratation im analogen und digitalen Raum kritisch reflektiert und transparent gemacht werden (vgl. Odumosu 2020).⁴ Vor allem im Bereich der Provenienzforschung liegt eine weitere Herausforderung darin, Kontexte zu rekonstruieren und dabei Information aus verteiltem Sammlungs- und Schriftgut miteinander zu vernetzen. Digital auslesbare Archivkataloge oder Findbücher, Transkriptionssoftware oder digitale objektbasierte Sammlungsdatenbanken, die Metadaten zu Provenienzen mitberücksichtigen oder in ihren Fokus rücken (Hopp 2018, 40, Sousa/Moser 2020, 86), bilden zwar gegenüber der analogen Recherche vereinfachte Bedingungen und beschleunigen Provenienzforschungen. Dennoch mangelt es an Verknüpfungen zwischen den einzelnen archivierten Dokumenten und den Sammlungsbeständen, an interinstitutionellen Standards oder an nachhaltiger Dokumentation von Sammlungs-, Forschungs- und Archivierungspraktiken. Dies stellt Forschende vor Schwierigkeiten in der Nachvollziehbarkeit und erfordert einen immensen zeitlichen Recherche- und Rekonstruktionsaufwand. Denn Sammlungsobjekte und die dazugehörigen Informationen zu ihrer Beschaffungs-, Nutzungs- und Verlagerungsgeschichte liegen in der Regel nicht nur in einer Sammlung – wie der des MfN – räumlich verstreut vor, sondern weisen Verteilungsnetz-

werke sowie Verbindungen mit anderen Akteur*innen, Institutionen oder Orten auf, die auch immer von Leerstellen durchzogen sind (vgl. Kuster et al. 2019, 106).

Obleich die einzelnen Forschungsprojekte des Clusters Mikroperspektiven des Makrosystems Museum auf der Ebene verschiedener Sammlungsbestände, Sammler*innen oder historischer Sammlungskontexte beleuchten, vereint sie der methodologische Zugriff über die interdisziplinäre, quellenbasierte Rekonstruktion der Sammlungs- und Objektgeschichten. Mit einem Schwerpunkt auf einen Sammler, spezifische Sammlungsbestände oder einen bestimmten Expeditionskontext wird der Versuch unternommen, zugehörige Objektbestände und ihre Provenienzen zu erschließen, die heterogenen Materialien aus Sammlungsgegenständen, dokumentierendem Schrift- oder Bildgut miteinander zu verknüpfen und digital auffindbar zu machen. Zwei dieser Projekte und ihre unterschiedlichen Ansätze werden im Folgenden vorgestellt. Dabei stehen hinter den Projekten des Clusters letztendlich auch verschiedene Vorstellungen, Ziele und Definitionen dahinter, was die Begriffe „Open Heritage“ in ihrer Umsetzung bedeuten könnten. Die Definition von „Open Heritage“ wird bei den folgenden Fallbeispielen im Sinne der Zugänglichkeit und Sichtbarkeit der Quellen aufgegriffen. Mithilfe der Forschungstools soll so Teilhabe an der Nutzung und Auswertung der Quellen ermöglicht werden (vgl. Sousa/Moser 2020, 96).⁵

Semantische Annotation der historischen Jahresberichte des Museums

Eine äußerst ergiebige Quelle zu Personalentwicklungen, Sammlungspraktiken, Bestandszu- und -abgängen sowie räumlichen Vernetzungen und Wegen von Personen oder Objekten bilden statistische Publikationen oder Jahresberichte sammelnder Institutionen.

Das Projekt „Forschungsfokus Provenienz: Digitale Edition der Jahresberichte des Museums für Naturkunde 1887-1915 und 1928-1938“ beschäftigt sich mit der digitalen Erschließung dieser Quellenbestände, insbesondere in Hinblick auf die Zeit der kolonialen Expansion des Deutschen Reiches.⁶ Die mineralogisch-petrographische, die geologisch-paläontologische und die zoologische Sammlung, die in dieser Zeit gemeinsam mit der Generalverwaltung das MfN konstituierten, waren institutionell ein Teil der Friedrich-Wilhelms-Universität, der heutigen Humboldt-Universität zu Berlin (HU). Diese publizierte in den Jahren zwischen 1887 und 1915 sowie 1928 und 1938 jährlich eine inzwischen von der HU-Bibliothek digitalisierte Chronik,⁷ in der jede ihr zugehörige organisatorische Einheit aufgefordert war, einen Bericht über die Aktivitäten und Ereignisse des Vorjahres einzureichen. In den Berichten der vier Abteilungen des MfN lassen sich Informationen zur Organisationsstruktur des Museums, zu personellen Veränderungen, Veröffentlichungen und Lehrveranstaltungen, Raumnutzung, Museumsinfrastruktur und -ordnungen, Nutzung und Zuwachs der einzelnen Museumssammlungen sowie ihrer wissenschaftlichen Auswertung und Bearbeitung finden.

Zentral sind außerdem Listen der Zu- und Abgänge von Sammlungsobjekten durch Tausch, Kauf und v. a. sogenannte Schenkungen durch Forscher oder Kolonialbeamte. Da insbesondere diese Daten eine einzigartige Quelle für die Sammlungsgeschichte des MfN darstellen, wurden speziell diese Aspekte in dem Projekt digital erschlossen.

Das Ziel dieser Digitalisierung ist es, ein sowohl für Menschen als auch Maschinen langfristig abfragbares Repositorium dieser Informationen zu erstellen. Bei der Modellierung der Daten waren vor allem drei Anforderungen relevant: Erstens mussten Textsequenzen und ihre jeweiligen Kontexte repräsentiert werden, die über die Jahre von unterschiedlichen Autoren und damit aus unterschiedlichen Perspektiven verfasst wurden. Deshalb wurde als Methode eine manuelle semantische Annotation der Textteile mit *INCEpTION*, einem für diesen Zweck an der TU Darmstadt entwickelten Tool, ausgewählt.⁸ Zweitens sollte die Möglichkeit bestehen, einerseits die semantischen Entitäten durch zusätzliche Quellen, sowohl aus digital noch unerschlossenem Archivmaterial als auch aus Repositorien wie Wikidata, inhaltlich zu bereichern, andererseits die strukturierten Daten im Sinne eines Linked Open Data Ansatzes in fremde Datensätze barrierefrei zu inkludieren – die Modellierung in einem Wissensgraphen lag damit nahe. Schließlich mussten die in den Jahresberichten angesprochenen Themenbereiche adäquat abgebildet werden können. Die von dem International Council for Documentation (CIDOC) des International Council of Museums als ISO-Standard entwickelte Ontologie CIDOC CRM bot sich als ISO-Standard mit ereigniszentriertem Dokumentationsansatz dafür an.⁹ Mithilfe dieses Ansatzes konnten aus den Jahresberichten über 12.000 individuelle Transaktionen von Objekten an das Museum rekonstruiert werden. Nahezu 80% davon konnten über 2.300 Sammler*innen sowie über 1.200 einzigartige Ursprungsorte der Objekte zugeordnet werden.

Die digitale, textbasierte Rekonstruktion der Berliner Kunstkammer

Ein weiteres Projekt des Clusters, das sich mit der textbasierten Wiedergewinnung von Sammlungs- und Objektinformation befasst, ist das DFG-Projekt „Das Fenster zur Natur und Kunst. Eine historisch-kritische Aufarbeitung der Brandenburgisch-Preußischen Kunstkammer“.¹⁰ Die Naturalien der Berliner Kunstkammer – eine Sammlung, die zwischen 1600 und 1875 an verschiedenen Orten auf der Spreeinsel existierte und deren Bestände sich fortwährend neu formierten – gingen 1810 in den Besitz der neu gegründeten Friedrich-Wilhelms-Universität (HU) über, zu der das Museum für Naturkunde bis 2009 gehörte. Damit bilden Objekte der enzyklopädisch angelegten Kunstkammer der preußischen Kurfürsten und Könige einen Grundstock des MfN. Neben einer Buchpublikation (vgl. Becker et al. 2023) entstand als Ergebnis des Projekts eine virtuelle Forschungsumgebung,¹¹ in der die wichtigsten Archivalien zur Berliner

Kunstkammer im Zeitraum von 1603 bis 1812 (ca. 25 Quellen) und die darin überlieferten Objekte recherchiert werden können. Ziel war es, neben der Erforschung einzelner Objektwege die Bestände ausgehend von Archivalien zu rekonstruieren. Dabei entstanden knapp 2000 digitale Objekteinträge. Einige Objekte haben sich zwar heute noch erhalten, so in der Sammlung des MfN, der HU oder der Staatlichen Museen zu Berlin. Viele von ihnen sind jedoch nur noch in historischen Quellen wie Inventaren, Museumsführern oder Reisebeschreibungen nachweisbar. Diese Tatsache wird durch die textbasierte Rekonstruktion berücksichtigt, denn so kann vor allem die Herkunft der Information zu Objekten nachvollziehbar gemacht werden. Alle aus den Quellen gewonnenen Objektinformationen werden beim Objekteintrag gebündelt, so etwa ihre Bezeichnung, das Material, aus dem sie bestehen, Motive, die sie zeigen, Personen, die mit ihnen in Verbindung stehen wie Hersteller oder Erfinder, die Art ihrer Präsentation, ihr Standort in den Sammlungsräumen oder auch Angaben zu ihrer Herkunft. Indem versucht wurde, die Objekte durch die verschiedenen Quellen hindurch immer wieder zu identifizieren, können divergente Angaben aus den Quellen nun verglichen werden. So lässt sich genau verfolgen, wie sich beispielsweise die Bezeichnung eines Objekts im Laufe der Zeit verändert, sein Standort oder sein Ort in der Systematik der Sammlung wechselt. Dieser Ansatz bietet die Möglichkeit, Schlussfolgerungen zum Bedeutungswandel eines Objekts oder Veränderungen in der Sammlungspraxis und -logik zu ziehen.

Bei der digitalen Erschließung kam die open source Software *WissKI*¹² als Grundgerüst sowie ein – wie auch bei der HU-Chronik – auf dem CIDOC CRM basierendes Datenmodell zum Einsatz (vgl. Wagner 2020, Wagner 2023). *WissKI* ist auf die standardbasierte Dokumentation heterogener Materialien ausgerichtet, erlaubt es, individuelle Sachverhalte zu modellieren und miteinander in Beziehung zu setzen, und bildet eine ideale Grundlage für Linked Open Data und damit die Basis für die digitale Vernetzung und Bereitstellung von Information aus dem Bereich kulturellen Erbes nach den FAIR-Prinzipien (Wilkinson et al. 2016). Das entwickelte Datenmodell, bei dem ausgehend von Schriftgut bzw. dessen Transkription Objekte und Sammlungen referenziert sowie ihnen zugewiesene Eigenschaften und Kontexte im Wortlaut – und zusätzlich mit Normdaten angereichert – dokumentiert werden, wird nun mit Blick auf seine Tragfähigkeit auf weitere Provenienzforschungsprojekte des Clusters ausgeweitet.

Herausforderungen und Potenziale für die Provenienzforschung

Die hier vorgestellten Erschließungsprojekte zeigen Herausforderungen und Potenziale der Provenienzforschung auf. Beide Projekte speichern die gewonnene Information in auf die jeweiligen Sachverhalte angepasste Wissensgraphen, die unterschiedliche Einstiegspunkte in die erfassten Daten ermöglichen und auf eine Anbindung an externe Informationsressourcen ausge-

legt sind. Die Erschließungsprojekte stellen daher Daten zu Quellen zur Verfügung, die in bereits existente Forschungsdateninfrastrukturen von NFDI4Culture oder Text+ eingebettet werden und so zu einer Wissens- und Datenvernetzung beitragen können (Fuhrmeister/Hopp 2019, 220). Die Annotation der Jahresberichte des MfN bietet die Möglichkeit, die in der Schriftquelle genannten semantischen Objekte, beispielsweise Ereignisse, Personen, Bestände oder Regionen, als untereinander mittelbar und unmittelbar verschränkte Netzwerke zu erfassen und eingängig visuell aufzubereiten. Hiermit lassen sich zum Beispiel geografische Translokationen gesammelter Güter aufzeigen. Diese Annotationsform geht somit über die Indexsuche von Personen- oder Ortsnamen und deren Verknüpfung mit GND-Einträgen hinaus, wie bei Transkriptionsprojekten wie der Transkribus¹³-basierten Plattform des Projekts „Die Rezesse der niederdeutschen Städtetage“.¹⁴

Auch die virtuelle Forschungsumgebung zur Berliner Kunstkammer bietet verschiedene Rechercheeinstiege. So wurde für Objekte und Quellen jeweils eine eigene Suche mit spezifischen Suchfacetten eingerichtet, wie dies in ähnlicher Weise durch die Objektkategorien und Filterfunktion der Datenbanken von PAESE (vgl. Andratschke/Müller 2021) und des BASA-Museums¹⁵ umgesetzt wurde. Daneben existieren Zugriffe über Personen, Motive, Objektarten oder auch Herkunftsorte, die wiederum mit Schriftgut und Objekten vernetzt und damit kontextualisiert sind. Auf diese Weise wird die Sammlungskonstitution des MfN über die einzelnen Objektgeschichten in Vernetzung zu anderen Objektkonvoluten, Schrift- und Bildquellen nachvollziehbar gemacht.

Die vorgestellten digitalen Erschließungsmethoden bieten Ansätze für erforderliche Strategien der digitalen Provenienzforschung für eine „mögliche[...] ‘Sichtbarmachung’ von räumlichen und zeitlichen Abläufen des Kulturguttransfers“ (Hopp 2018, 42)¹⁶ – auch für naturkundliche Sammlungen, die in diesem Diskurs bislang unterrepräsentiert sind.¹⁷ Die Rekonstruktion und Sichtbarmachung der historischen Kontexte der Sammlungsbestände des MfN liegen in der Verantwortung des Museums, insbesondere wenn diese aus Gewaltkontexten des Imperialismus oder Kolonialismus stammen, unter machtasymmetrischen Erwerbsbedingungen oder im Zuge problematischer Sammlungs- und Forschungspraktiken in das Museum gelangt sind. Die Transparenz und Sichtbarmachung kolonialer Verflechtungen innerhalb der Bestandskonstitution naturkundlicher Sammlungen setzt daher die Aufarbeitung der Provenienzen der Sammlungsbestände und ihre möglichst niedrigschwellige Bereitstellung durch Digitalisierungsmaßnahmen voraus. Dabei erfordert der koloniale Kontext der Sammlungspraktiken und die Sensibilität der betreffenden Sammlungsgegenstände auch eine Reflexion und sensible Herangehensweisen in der Forschungs- und Digitalisierungspraxis, die Forschende vor spezifische Epistemologien sowie Logistiken stellt, ethische und politische Standards benötigt, aber auch einen kritischen sowie politischen Reflexionsprozess anregen kann.¹⁸

Fußnoten

1. https://www.museumfuernaturkunde.berlin/sites/default/files/mfn_zukunftsplan_digital.pdf (zugegriffen: 5. Dezember 2022).
2. Siehe dazu den „Leitfaden zum Umgang mit Sammlungsgut aus kolonialen Kontexten“ des Deutschen Museumsbundes, <https://www.museumsbund.de/wp-content/uploads/2021/03/mb-leitfaden-web-210228-02.pdf> (zugegriffen: 5. Dezember 2022), die „Ersten Eckpunkte zum Umgang mit Sammlungsgut aus kolonialen Kontexten“, https://www.kmk.org/fileadmin/pdf/PresseUndAktuelles/2019/2019-03-25_Erste-Eckpunkte-Sammlungsgut-koloniale-Kontexte_final.pdf (zugegriffen: 5. Dezember 2022) oder die „3 Wege-Strategie“ für die Erfassung und digitale Veröffentlichung von Sammlungsgut aus kolonialen Kontexten in Deutschland, https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2020/201014_Kontaktstelle-Sammlungsgut_Konzept_3-Wege-Strategie.pdf (zugegriffen: 5. Dezember 2022).
3. <https://www.museumfuernaturkunde.berlin/de/wissenschaft/forschungscluster-openheritage> (zugegriffen: 5. Dezember 2022).
4. Brigitta Kuster, Britta Lange und Petra Löffler rücken in einem Dreiergespräch dabei in den Fokus, dass auch die Praktiken der wissenschaftlichen Arbeit, der Nutzung oder der Entscheidungsfindungen im Archiv ebenfalls als Teile von diesen in einem „Meta-Archiv“ dokumentiert werden müssten. Da bei der Digitalisierung von Sammlungen auch eine Reevaluierung darüber stattfinden würde, was der Öffentlichkeit zur Verfügung gestellt werde und was nicht, müssten auch entstehende Leerstellen sichtbar gemacht werden. (Kuster et al. 2019, 104-111).
5. Die genannte Integration von Mehrstimmigkeit wurde in diesen Projekten noch nicht umgesetzt. Im Rahmen des Forschungsclusters beschäftigen sich jedoch andere Projekte wie „Civic digitisation“ und „Academic solidarities“ konkret mit Partizipationsmöglichkeiten und deren Gestaltungsweisen im digitalen Raum, für welche auch die vorgestellten virtuellen Forschungsumgebungen weiterentwickelt werden könnten.
6. <https://www.museumfuernaturkunde.berlin/de/wissenschaft/forschungsfokus-provenienz> sowie <https://github.com/marquart/naturkundemuseum-annotation> (zugegriffen: 5. Dezember 2022).
7. <http://www.digi-hub.de/viewer/resolver?urn=urn:nbn:de:kobv:11-d-6653534> (zugegriffen: 5. Dezember 2022).
8. <https://inception-project.github.io> (zugegriffen: 5. Dezember 2022).
9. <https://cidoc-crm.org> (zugegriffen: 5. Dezember 2022).
10. <https://www.museumfuernaturkunde.berlin/de/wissenschaft/das-fenster-zur-natur-und-kunst> (zugegriffen: 5. Dezember 2022).
11. <https://berliner-kunst-kammer.de> (Launch: 15. Dezember 2022).
12. <https://wiss-ki.eu/> (zugegriffen: 5. Dezember 2022).
13. <https://readcoop.eu/de/transkribus/> (zugegriffen: 5. Dezember 2022).

14. <https://transkribus.eu/r/rezesse-niederdeutscher-staedtetage/#/> (zugegriffen: 5. Dezember 2022).
15. https://kosmos.uni-bonn.de/wisski_basa/ und <https://www.postcolonial-provenance-research.com/datenbank/> (zugegriffen: 5. Dezember 2022).
16. Den Ansatz der Sichtbarmachung und der „größtmögliche[n] Transparenz“ über die Herkunft der dokumentierten Objekte strebt auch die bereits genannte PAESE-Verbunddatenbank an (vgl. Andratschke/Müller 2021, 3).
17. Die kommentierte Online-Edition der fünf Reisetagebücher Hans Posses vom Germanischen Nationalmuseum bildet ein herausragendes Beispiel einer virtuellen Forschungsumgebung und der kommentierten Annotation von Quellenmaterial, <https://editionhansposse.gnm.de/> (zugegriffen: 5. Dezember 2022).
18. Ein Artikel zu solcherart Fragen und Herausforderungen ist von Kolleg*innen des Clusters Open Heritage derzeit in Begutachtung (vgl. Kaiser et al. 2023).

Bibliographie

- Andratschke, Claudia, Müller, Lars. 2021. *Einführung in die PAESE-Datenbank*. https://www.postcolonial-provenance-research.com/wp-content/uploads/2022/03/PAESE-Datenbank_Einf%C3%BChrung_final.pdf (zugegriffen: 5. Dezember 2022).
- Becker, Marcus, Eva Dolezel, Meike Knittel, Diana Stört und Sarah Wagner. 2023. *Die Berliner Kunstskammer. Sammlungsgeschichte in Objektbiografien vom 16. bis 21. Jahrhundert*. Petersberg: Imhof.
- Fuhrmeister, Christian, Hopp, Meike. 2019. "Rethinking Provenance Research." *Getty Research Journal* 11.2019: 213-231.
- Hopp, Meike. 2018. "Provenienzrecherche und digitale Forschungsinfrastrukturen in Deutschland: Bedürfnisse, Desiderate, Tendenzen." In *...(k)ein Ende in Sicht. 20 Jahre Kunstrückgabegesetz in Österreich*, hg. von Eva Blimlinger und Heinz Schödl, 35-59. Wien: Böhlau Verlag.
- Kaiser, Katja, Heumann, Ina, Nadim, Tahani, Keysar, Hagit, Petersen, Mareike, Korun, Meryem, Berger, Frederik. 2023. "Utopias of Mass Digitisation and the Colonial Realities of Natural History Collections." *Journal of Natural Science Collections* (in Vorbereitung).
- Kuster, Brigitta, Lange, Britta, Löffler, Petra. 2019. "Archive der Zukunft? Ein Gespräch über Sammlungspolitik, koloniale Archive und die Dekolonisierung des Wissens." *Zeitschrift für Medienwissenschaft*, 20.1: 96-111.
- Odumosu, Temi. 2020. "The Crying Child. On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons." *Current Anthropology* 61.22: 289-302.
- Sousa, Jason, Moser, Ariane. 2020. "Data and Databases in Provenance Research." In *Provenance Research Today. Principles, Practice, Problems*, hg. v. Arthur Tompkins, 85-96. London: Lund Humphries.
- Wagner, Sarah. 2020. "Unsichtbares sichtbar machen. Semantische Modellierung interpretativer Vorgänge am Beispiel der historischen Bestandsaufnahme der Brandenburgisch-Preußischen Kunstskammern." In *Digital Humanities im deutschsprachigen Raum 2020. Spielräume. Digital Humanities zwischen Modellierung und Interpretation*, hg. von Christof Schöch, 238-240.

Wagner, Sarah. 2023. "Vom Schloss ins Internet. Die virtuelle Forschungsumgebung zur Berliner Kunstskammer." In *Die Berliner Kunstskammer. Sammlungsgeschichte in Objektbiografien vom 16. bis 21. Jahrhundert*, hg. v. Marcus Becker, Eva Dolezel, Meike Knittel, Diana Stört und Sarah Wagner, 16-21. Petersberg: Imhof.

Wilkinson, Mark, Dumontier, Michel, Aalbersberg, IJsbrand et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18> (zugeschrieben: 5. Dezember 2022).

Increasing the visibility of Tyrol's cultural heritage through historical newspapers – the triple-open approach of the Zeit.shift project

Lyding, Verena

verena.lyding@eurac.edu

Eurac Research Bolzano/Bozen, Italien

Franzini, Greta

greta.franzini@eurac.edu

Eurac Research Bolzano/Bozen, Italien

Introduction

Tyrolean tangible cultural heritage is manifested in various artefacts and documents spread across different regions in Austria and northern Italy, spanning the territory of historical Tyrol as it was delineated in the early 20th century. The written testimonies are preserved in historical institutions, private collections and public libraries. Efforts to digitize historical Tyrolean documents to safeguard their long-term preservation began about two decades ago with the Austrian Literature Online (ALO) project (Egger and Mühlberger, 2000) and Europeana Newspapers (Pekárek and Willems, 2012). However, the digitization of all relevant historical documents is still far from complete and digitization processes are often pursued by institutions with limited coordination between them.

The fact that many historical documents are only available as paper copies limits their access to expert groups and the public alike.¹ It follows that historical text sources are known and used by only a few people when they could be valuable sources of information for larger parts of so-

ciety, be those relating to the educational sector, public culture or research.

This is where the Zeit.shift project comes in. Using historical newspapers as a use case, this interdisciplinary and interregional project aims at increasing the visibility and use of historical text sources by bringing together dispersed items in digital form, making them accessible and promoting their value and use through educational and participatory campaigns.

Project consortium and project goals

Zeit.shift is funded by the European Regional Development Fund and Interreg V-A Italia - Austria 2014-2020. The project started in October 2020 and will run until June 2023. It is a cooperation between two libraries, the Dr. Friedrich Teßmann Library (Bolzano, Italy, project lead) and the University and State Library of Tyrol (Innsbruck, Austria), together with the research Institute for Applied Linguistics at Eurac Research (Bolzano, Italy). The library partners contribute their data and expertise in collecting, cataloguing, digitizing and publishing historical text sources. The research institute contributes its expertise in computational linguistics, Digital Humanities, crowdsourcing and participatory approaches for the (semi-)automatic processing of textual resources with the aim of enhancing their informational content, their searchability and thus their overall use. Seven cultural institutions support the project as associated partners.²

The Zeit.shift projects seeks to preserve, develop and communicate the cultural and textual heritage of the historical region of Tyrol by, firstly, digitizing and bringing together Tyrolean historical newspapers from the early 20th century (mostly written in German Fraktur); secondly, by upvaluing the data, improving their searchability and making them accessible via a dedicated and free online portal; and, thirdly, by disseminating the data to be used by various stakeholder groups through free educational materials, workshops and participatory activities. The activities are promoted through campaigns, and complemented by free educational workshops and an e-learning course.

Research and development

The aims of Zeit.shift to extend access, visibility and use of historical text documents call for an open research and development approach. Indeed, the project fosters open procedures and open access in three distinct but interconnected areas of the project implementation: (1) openness in collaboration and knowledge sharing, (2) openness of data and tools, and (3) openness in education and participation.

Open collaboration and knowledge sharing

Collaboration openness mostly concerns the work of the two library partners, who join forces in digitization efforts which were previously carried out in parallel with little interaction. This includes sharing experiences on good practices, exchanging information on the cataloguing work, agreeing on naming conventions and harmonizing workflows. In addition, both data and digitization costs are shared between the partners and the digitized data will be published through a joint online portal.

The collaboration with the research partner also builds on openness in data and knowledge exchange, but it is less challenging as in this typical research collaboration the competences of the partners are complementary and thus responsibilities clearly defined.

Open data and tools

All data that is produced within the project is and will be openly available. The digital copies of the historical newspapers and their metadata are published under Creative Commons Attribution licenses. The computational linguistic processing and enhancement of the newspaper data is based on non-commercial open tools; all scripts, toolchains and other code developed within Zeit.shift are and will be made available via Eurac Research's GitLab repository.³ Data collected through the participatory activities is and will also be accessible under open licenses.

Open education and participation

Zeit.shift pursues an active and open approach towards citizen participation and dissemination of project results. One of its goals is to engage the public and raise interest for the historical data that is digitized within the project. This is done by developing initiatives that invite the local population to interact with the data and contribute to their processing and annotation (see §4.2.). These initiatives are promoted through specific campaigns, including stands at conferences and transfer-oriented events such as the *Long Night of Research*, as well as radio and TV broadcasts, announcements in local print media, flyers, stickers and a prize competition (ibid.).

Additionally, Zeit.shift is investing in the formation of multipliers for different stakeholder groups, ranging from professional archivists and library staff to hobby historians ('Chronisten' in German), teachers and teacher trainers. The course materials are jointly elaborated by the partners and are openly shared with the community via multiple institutional repositories. They introduce users to both the basics of digital and automated procedures in text processing and historical studies, and to the participatory activities offered by the project. In addition, the project has published a free Massive Open Online Course (MOOC) to show students and interested lay citizens how historical daily newspapers from North, East and South Tyrol can be used as online sources for research, to illustrate the potential of Digital Humanities and to provide insights into some computational linguistic techniques that facilitate work with historical newspapers.⁴

Project results

Two years into the project, we can report the following results.

Digitized historical newspapers

Some 47.631 issues (amounting to 423.782 pages) from 41 newspapers published between 1880 to 1950 in the historical region Tyrol have been digitized and are currently being made available through the online portals of the Teßmann⁵ and Innsbruck University Libraries⁶. The digitized data includes OCR'd text in ALTO-XML format, aligned at the page level with scanned 400 dpi resolution color images.

The data is automatically annotated for Named Entities using the spaCy named entity tagger (Honnicbal and Montani, 2017) augmented with a trimmed list of local toponyms ('Flurnamen' in German⁷). Each newspaper page is automatically tagged with one or more topic labels distinguishing between local news, global news, announcements, advertisements, and serial stories. The final, joint newspaper portal, which is still under development, will make use of these annotations to enhance search functionality.

Participatory activities

Two participatory activities have been developed and are available online. Both are open to all participants alike but were designed for specific user groups.

The first activity (*macro-task*), targets German-speaking specialists (librarians, chroniclers, historians, etc.) and invites them to geo- and semantically tag historical advertisements automatically extracted from the Zeit.shift newspaper collection using the Historypin platform (Fig. 1).⁸ Currently, the Zeit.shift collection in Historypin contains some 7,000 adverts from ten different newspapers and, as of December 2022, 27 people have (geo)tagged 351 adverts with a total of 1,668 unique tags. As a by-product of the project, the Zeit.shift team produced the German translation of the Historypin platform to the benefit of the German-speaking community.⁹



Figure 1. An advert from the *Schwazer Lokalanzeiger* about cough sweets in Historypin. The pink and black pin on the map marks the pharmacy's location. Source: <https://tinyurl.com/mzkzzhwz>

The second activity is *Ötzi!*, a custom web game inviting participants between 11-14 years of age to speed-type words extracted from the newspapers, which is designed to help them learn this script whilst contributing to OCR post-correction efforts (*micro-task*).¹⁰ In the game, alpine animals walk in the direction of Ötzi the Iceman¹¹ looking to harm him while Fraktur words appear on the screen; players must type the words correctly as fast as possible to fend off the animals and thus preserve Ötzi's health (Fig. 2). Knowledge of German is not necessary to play the game but certainly an advantage.

Ötzi! is released under an MIT license and can be played on both desktop and mobile devices. It is made up of two software components.¹² The *frontend* implements user interaction, that is, the gameplay; it is written in Javascript using Phaser¹³ and is distributed via Itch.io¹⁴, the *de facto* standard platform for indie games. The *backend* implements all data flow (providing words to the frontend and receiving game data via an API) and data analysis; it is written in Javascript using Express¹⁵ and is hosted on Eurac Research infrastructure. *Ötzi!* sources assets (audio files and graphics) published under open licenses only.

Transcriptions anonymously typed by players are collected to test the efficacy of the game as an OCR manual post-correction tool. As of December 2022, the game has been played by 1,754 unique devices and out of the 6,909 words typed thus far, 890 confirm the OCR output, 442 provide corrections and 5,577 are pending automatic evaluation.¹⁶

In September 2022, the project launched a prize competition for middle schools in South Tyrol to advertise the game to the younger population.¹⁷ The competition and advertising campaign, which ran between 12th September and 31st October 2022, was very successful in recruiting players from both the target group and the general public, with 1,820 games initiated, 729 games completed, 88.72 hours of total game time (an average 3 minutes per initiated game and 7 minutes per completed game) distributed across 914 unique devices.

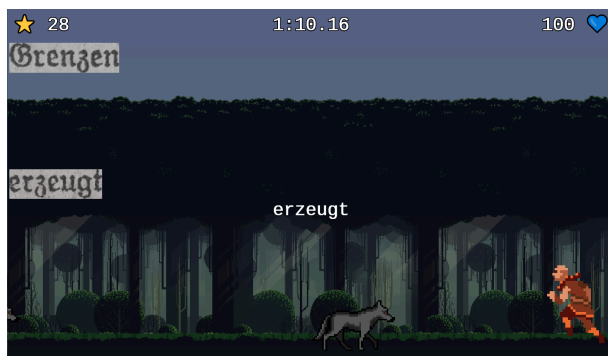


Figure 2. *Ötzi!* screenshot. The player types Fraktur words as they appear on the screen. At the top of the screen, points accumulated are shown in the left corner, elapsed game time in the center and health status points in the right corner.

Dissemination and training

Multiple training and dissemination activities have been carried out in both South Tyrol and North Tyrol. To date, the project has conducted nine training workshops for the Historypin activity for a rough total of 100 participants, and another 200 people have been informed about the project activities through presentations at local stakeholder institutions, including Tiroler Landesmuseum Ferdinandeum and Standtarchiv Bozen. The MOOC contains six modules and is followed by 63 participants. Flyers and posters have been distributed by all partners in their vicinity. Several press releases have been published¹⁸, and scientific publications have been presented at three events: a demo session at the Engaging Citizen Science Conference 2022 (CitSci2022)¹⁹ in Aarhus (Denmark) (Franzini et al., forthcoming), and two poster presentations at the annual conference of the *Associazione per l'Informatica Umanistica e la Cultura Digitale* (AIUCD)²⁰ in Lecce (Italy) (Franzini et al., 2022) and at the 7th Austrian Citizen Science Conference²¹ in Dornbirn (Austria).

Joint newspaper portal

The design phase of the Zeit.shift newspaper portal addressed the requirements and technical needs of the three project partners. The portal is currently being implemented by an ICT service provider and will go live in the spring of 2023.

Conclusions and future work

In this article we describe how an open approach is helping to increase the visibility and support the use of historical resources within the Zeit.shift project. Indeed, we believe that the wider promotion and use of cultural heritage data should move beyond the mere provision of open data to foster the engagement of citizens through education and participatory initiatives, as well as the collaboration among cultural institutions in relation to data, procedures and networks.

At the same time, historical sources are often difficult to process automatically, which makes manual processing by researchers and wider stakeholder groups highly necessary. Follow-up work on the project might focus on developing additional participatory activities to enhance the historical data sources in a playful and educational fashion. This could also include further research on how any data collected from a non-expert audience can be processed and aggregated to improve the original sources in a reliable way. These activities would serve a twofold purpose: improve the quality of digital copies of fragile historical items for long-term preservation, and increase their visibility and use for educational purposes and public cultural awareness.

Acknowledgements

Zeit.shift has received funding from the European Regional Development Fund under the Interreg Italia - Ös-

sterreich 2014-2020 Programme (Project no. ITAT3030). We thank the conference reviewers for their helpful suggestions and are especially grateful to our contributing citizens.

Fußnoten

1. Access to these typically fragile and valuable materials is often regulated by strict protocols, such as assisted visits by appointment.
2. Euregio Tirol-Südtirol-Trentino, Abteilung Tiroler Landesarchiv der Tiroler Landesregierung, Tiroler Landesmuseen, Südtiroler Kulturinstitut, Tiroler Bildungsforum, Südtiroler Landesarchiv der Autonomen Provinz Bozen-Südtirol, Bibliotheksverband Südtirol.
3. <https://github.com/commul>
4. <https://imoox.at/course/zeitshift2022>
5. <https://digital.tessmann.it/>
6. <https://diglib.uibk.ac.at/obvuiibz>
7. <https://www.natura.museum/de/magazine/flurnamen-suedtirols/>
8. <https://www.historypin.org/en/zeit-shift>
9. <https://about.historypin.org/2021/10/13/historypin-in-german/>
10. <https://eurac.itich.io/oetzit>
11. <https://www.iceman.it/en/the-iceman/>
12. The code is available at <https://gitlab.inf.unibz.it/commul/oetzit>
13. <https://phaser.io/>
14. <https://itch.io/>
15. <https://express.js.com/>
16. In our current experimental setup, an OCR'd word is confirmed or corrected if transcribed by at least three different players.
17. Press releases of the 2022 prize competition are available at <https://news.provinz.bz.it/de/news/otzit-startschuss-fur-online-gewinnspiel-fur-kinder-und-jugendliche>; <https://www.suedtirolnews.it/unterhaltung/kultur/online-gewinnspiel-fuer-kinder-und-jugendliche-startet-am-12-september>; <https://www.lavocedibolzano.it/la-biblioteca-provinciale-tessmann-organizza-sino-al-30-ottobre-il-gioco-a-premi-online-otzit/>
18. <https://all4ling.eurac.edu/projects/zeitshift/press/>
19. <https://conferences.au.dk/citsci2022>
20. <http://conference.unisalento.it/ocs/index.php/aiucd2022/aiucd2022>
21. <https://www.oesterreichische-citizen-science-konferenz-2022.com/>

Bibliographie

Egger, Alexander and Günther Mühlberger. 2000. "Austrian. ALO oder die virtuelle Bibliothek der österreichischen Literatur." *Bibliotheksdienst* 34(6): 958-967. <https://doi.org/10.1515/bd.2000.34.6.958>

Franzini, Greta, Egon W. Stemle, Verena Lyding, Andrea Abel, Johannes Andresen, Karin Pircher, Silvia Gstrein, Barbara Laner, Johanna Walcher, Maritta Horwath, Christian Koessler. 2022. "Citizen Humanities in Tyrol: A case study on historical newspapers." In *Proceedings of*

the Eleventh AIUCD Annual Conference (AIUCD 2022), edited by Ciraci, Fabio, Giulia Miglietta and Carola Gatto, 236-238. Lecce, 1-3 June. ISBN: 978-88-9425-356-6. DOI: <http://doi.org/10.6092/unibo/amsacta/6848>

Franzini, Greta, Paolo Brasolin, Verena Lyding, Egon Stemle, Andrea Abel, Johannes Andresen. Forthcoming. "Zeit.shift: Driving Citizens to Tyrolean Historical Newspapers." In *Proceedings of the Engaging Citizen Science Conference 2022 (CitSci2022)* edited by Nielsen, Kristian, Gitte Kragh, and Lori Nash, Proceedings of Science.

Honnibal, Matthew and Ines Montani. 2017. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing."

Pekárek, Aleš and Marieke Willems. 2012. "The European Newspapers – A Gateway to European Newspapers Online." In *Progress in Cultural Heritage Preservation*, edited by Ioannides, Marinos, Dieter Fritsch, Johanna Leissner, Rob Davies, Fabio Remondino, and Rossella Caffo, 654-659. EuroMed 2012. Lecture Notes in Computer Science, vol 7616. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34234-9_68

Internationale Autor*innen zu Gast in der DDR: Die Einreisekartei des Schriftstellerverbandes und ihre digitale Aufbereitung

Fischer, Frank

fr.fischer@fu-berlin.de
Freie Universität Berlin

Illmer, Viktor Jonathan

v.illmer@fu-berlin.de
Freie Universität Berlin

Regeler, Lukas Nils

lukas.regeler@fu-berlin.de
Freie Universität Berlin

Müller-Tamm, Jutta

muellert@zedat.fu-berlin.de
Freie Universität Berlin

von Berenberg-Gossler, Luise

l.von.berenberg-gossler@fu-berlin.de
Freie Universität Berlin

Diehr, Franziska

diehrf@rki.de

Robert Koch-Institut

1. Forschungshintergrund

Marcel Reich-Ranicki kam 1955 und 1956. Der sowjetische Autor Michail Scholochow besuchte die DDR 1964, zwei Jahre bevor er den Nobelpreis erhielt; ähnlich der guatemaltekeische Schriftsteller und spätere Nobelpreisträger Miguel Asturias, der 1965 nach Ostberlin reiste. Friederike Mayröcker folgte einer Einladung im Mai 1987. Andere kamen wiederholt, wie der ungarische Dichter Gábor Hajnal, der sich zwischen 1957 und 1986 dreizehn Mal in Ostberlin aufhielt. Eingeladen hatte jeweils der Deutsche Schriftstellerverband (DSV), über den der Großteil der internationalen literarischen Kontakte in der DDR organisiert wurde. Tausende Daten zur Einladungs- politik des Verbandes sind in einer Kartei in der Akademie der Künste in Berlin hinterlegt, deren Bestand im Rahmen der Archivarbeiten für das Forschungsprojekt »Writing Berlin« digitalisiert wurde.¹

»Writing Berlin« ist Teil des Exzellenzclusters »Temporal Communities. Doing Literature in a Global Perspective« (EXC 2020) und befasst sich mit den facettenreichen Aktivitäten zur Förderung des internationalen literarischen Austauschs in der geteilten Stadt nach dem Bau der Berliner Mauer. Ein besonderes Augenmerk liegt dabei auf den Auswahlprozessen und den kulturpolitischen Implikationen dieser Aktivitäten, ihrem Niederschlag in literarischen Texten sowie auf der Frage, inwiefern die sich verändernde politische Gemengelage Biografien und die soziale Stellung der betreffenden Autor*innen beeinflusste. Die Internationalisierung der Berliner Literaturszene ist bislang nur in einigen wenigen Fallstudien untersucht worden, vor allem im Hinblick auf die Netzwerktätigkeit einzelner Schriftsteller*innen (vgl. Böttiger 2005, Berbig 2005). Der institutionalisierte Austausch, der einen Großteil der internationalen Kontakte im Osten der Stadt ausmachte, war bislang noch nicht Gegenstand weitergehender Studien – zwar liegen allgemeine Untersuchungen zum Schriftstellerverband der DDR vor, diese erwähnen die politisch so relevante Auslandsarbeit der Organisation jedoch bestenfalls beiläufig (vgl. zum DSV allgemein Pampierri 2004, Walther 2006, Michael et al. 1997) und betrachten lediglich einen sehr eingeschränkten Zeitraum (vgl. insbesondere zu den 1950er-Jahren Degen 2011, Gansel 1997). Die Einreisekartei des DSV, der wichtigsten nichtstaatlichen Literaturinstitution im Ostteil der Stadt, erlaubt es nun, die internationalen Kontakte und ihre Konjunkturen insbesondere in der spannungsgeladenen Zeit während des Bestehens der Berliner Mauer zu erforschen: den Verlauf dieser Aktivitäten insgesamt, die länderbezogene Einladungspolitik, die Umstände individueller Aufenthalte und ihre politische Rolle für das Herkunftsland. Sie ermöglicht auch, Literaturkontakte weniger um besonders hervorstechende Einzelpersonen zentriert zu denken und dabei gerade auch Autor*innen zu berücksichtigen, die

durch Kanonisierungsprozesse der Vor- und Nachwendzeit ggf. in Vergessenheit geraten sind.

2. Digitalisierung und Anreicherung mit OpenRefine

Zunächst wurden die Einreisekarteien im Archiv des DSV transkribiert. Als Grundlage dafür wurden die nach Ländern und Autor*innennamen geordneten Karteien verwendet. Für jede*n einreisende*n Autor*in existiert so mindestens ein separates Blatt, auf dem die verschiedenen Aufenthalte vermerkt sind. Die Mitarbeiter*innen der Auslandsabteilung des DSV ergänzten ggf. noch biografische Informationen oder auch ein Presse- oder Passfoto. Über die Jahrzehnte änderte sich vielfach die Art der Aufzeichnung, ein Großteil der etwa 3.000 Karteien orientiert sich jedoch an dem in Tabelle 1 wiedergegebenen Schema, das am Beispiel des kubanischen Dichters Nicolás Guillén in Abbildung 1 illustriert werden soll.

Name, ggf. Vorname	Land
[ggf. kurze biografische Information]	
Datum	Aufenthaltsgrund
Datum	Aufenthaltsgrund
...	...

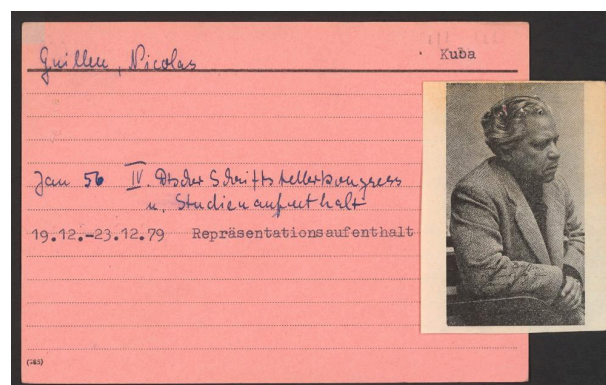


Abbildung 1: Beispiel für die Datengrundlage anhand der Karte zum kubanischen Autor Nicolás Guillén. Archiv des Schriftstellerverbands der DDR, Literaturarchiv der Akademie der Künste Berlin, Signatur SV 2848. (© Archiv der Akademie der Künste, Veröffentlichung mit freundlicher Genehmigung.)

Qualität und Umfang der angegebenen Informationen variieren stark: Besonders in den 1970er-Jahren etwa wurden teils nur die Nachnamen der Autor*innen notiert, ggf. mit Zusätzen wie »Herr«/»Frau«, der Abkürzung »Gen.« (= Genosse) oder mit akademischen Titeln. Auch die Zeiträume der Aufenthalte sind nicht einheitlich angegeben, gelegentlich finden sich nur Jahres- oder Monatsangaben. Zudem handelt es sich überwiegend um per Hand getätigte Vermerke; nicht immer ist dabei die Schrift der Mitarbeiter*innen leicht zu entziffern. Auch erfolgten viele Angaben offenbar nicht auf Grundlage von Ausweisdokumenten, sondern durch Hörensagen. Gerade bei Autor*innen aus dem Globalen Süden etwa kann die Transkription der Namen inkorrekt sein oder sie wurde unsystematisch an deutsche Schreib-

gewohnheiten angepasst. Der nordkoreanische Autor Chông Tök-ch'öl wird beispielsweise mit »Dok Tschol Dschong« transkribiert, der afghanische Dichter Vâšif Bâkhtari taucht in den Karteien als »Wassef Bachterie« auf, der indische Schriftsteller Harivansh Rai Bachchan unter der Schreibweise »Harbans Rai Bachhan«. Auch Namen von Autor*innen aus sozialistischen Bruderstaaten, etwa der ČSSR, Bulgarien oder Rumänien sind häufig fehlerhaft notiert.

Um die Informationen aus der nach Autor*innennamen sortierten Einreisekartei zu komplettieren, wurden auch die chronologischen Karteien herangezogen sowie etwa punktuell weitere Akten aus dem Archiv des DSV, etwa die zu etlichen Aufenthalten vorhandenen Freundschaftsverträge, Korrespondenzen und Zeitpläne. Als Ergebnis dieser Transkriptionsarbeit entstand eine Excel-Tabelle mit insgesamt 3.709 Einträgen. Die Tabelle enthält Informationen zum Zeitraum des jeweiligen Aufenthalts, den Autor*innen (Name und Staatsangehörigkeit), zu beteiligten Institutionen sowie Angaben zum Anlass bzw. Einladungsgrund.

Mit OpenRefine (Version 3.5.0) wurden die in der Excel-Tabelle enthaltenen Daten vereinheitlicht. So konnten Einträge, die zwar denselben Anlass betrafen, aber unterschiedlich verschriftlicht waren, zusammengeführt werden.

In einem weiteren Schritt wurden die teils in problematischer Weise notierten Autor*innennamen über OpenRefine aufbereitet und mit Normdatensätzen verknüpft. Dadurch wurde zum einen die Verifizierung bzw. Identifizierung der in der Kartei verzeichneten Einträge vereinfacht; zum anderen konnten aus den verknüpften Datenbanken weitere Informationen zu den Autor*innen importiert werden.

Ein erstes umfangreiches Reconciling erfolgte mit dem Virtual International Authority File (VIAF). Als Grundlage hierfür diente der von Jeff Chiu über Codefork bereitgestellte Reconciliation Service (Version 3.0.5, <https://refine.codefork.com/>), der einen erfolgreichen Abgleich von zunächst etwa 25 % der Einträge ermöglichte. Nach manuellen Suchstrategien in der Datenbank, etwa dem Ausprobieren verschiedener Schreibweisen und dem Abgleich mit im VIAF hinterlegten biobibliografischen Daten, konnte die Trefferquote auf fast 90 % erhöht werden.

Ein weiteres Reconciling wurde – über das in OpenRefine integrierte Tool – mit Wikidata vorgenommen. Auch hier konnte durch einige Nachjustierungen eine hohe Trefferquote von 75 % erzielt werden. Der erfolgreiche Abgleich ermöglichte nun den Import weiterer Informationen aus Wikidata, etwa Angaben zu Sprachen, Parteizugehörigkeit oder Geschlecht der Autor*innen. Zudem konnten weitere Identifier über Wikidata importiert und somit Schnittstellen zur Gemeinsamen Normdatei der Deutschen Nationalbibliothek (GND) und zum WorldCat geschaffen werden, wodurch nun auch bibliografische Informationen zu den eingeladenen Autor*innen recherchierbar sind.

Jeder einzelne der hier dargestellten Schritte stellt eine Interpretationsleistung der Daten dar, die ihrerseits wieder nur heuristisch erfolgen, unvollständig und fehlerbehaftet sein kann. Bei dem über OpenRefine bereinigten und abgeglichenen Datensatz handelt es sich somit nur um eine mögliche Lesart der ursprünglichen Einreisekar-

teien, die der fortwährenden Überprüfung und Modifizierung bedarf.

3. Schnittstelle, Web-App und Modellierung unvollständiger Datumsangaben

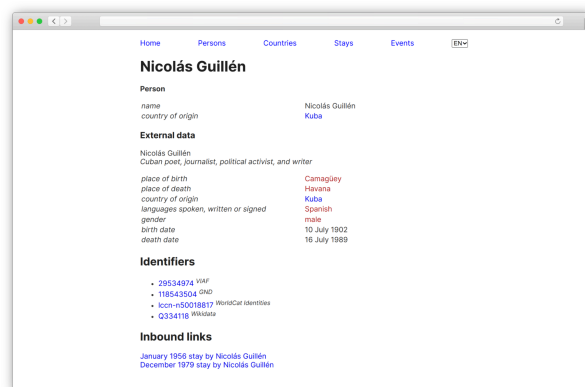


Abbildung 2: Prototypische Ansicht der Web-App für den Eintrag zu Nicolás Guillén.

Um den digitalisierten und erschlossenen Datensatz besser erforschbar zu machen, wurde für die Daten eine Schnittstelle geschaffen, über die sie maßgeschneidert ausgeliefert werden können. Diese Schnittstelle dient auch als Grundlage für die Web-App, die einen explorativen Zugang zum Datenmaterial ermöglichen soll (Abbildung 2). Die App wurde in TypeScript mit dem SvelteKit-Framework geschrieben (vgl. <https://kit.svelte.dev/>), welches das Front-End-Framework Svelte (vgl. <https://svelte.dev/>) in ein umfangreiches Web-Framework mit serverseitigem Rendering (SSR) integriert. Svelte agiert, anders als das vergleichbare React, nicht während der Laufzeit, sondern als Compiler. Zum Build-Zeitpunkt werden alle HTML-Vorlagen zu nativen JavaScript-Funktionen kompiliert. Ein Vorteil ist der Verzicht auf ein virtuelles DOM (Document Object Model) und die Berechnung der damit einhergehenden Deltas (vgl. Harris 2018). SvelteKit, das sich derzeit noch in der Beta befindet, gilt als Svelte-Pendant zu Reacts Next.js. Damit läuft Svelte auch auf dem Server und ist in der Lage, Seiten dort entweder zum Build- oder Anfragezeitpunkt vorzurendern und anspruchsvolle Datenoperationen auszuführen, während clientseitig die volle Flexibilität eines reaktiven Frameworks erhalten bleibt (vgl. <https://kit.svelte.dev/docs/introduction>). Im vorliegenden Projekt sind das etwa Daten aus Wikidata, die zur Anreicherung des bestehenden Datensatzes gebündelt serverseitig abgerufen und transformiert werden.

Wegen der teils unvollständigen Datumsangaben haben wir auf das Extended Date/Time Format (EDTF) gesetzt. Dieses 2019 von der International Organization for Standardization als Erweiterung zu ISO-8601 gedachte Datumsformat erlaubt es unter anderem, verschiedene Arten von Ungewissheit formalisiert auszudrücken. Für die Zwecke dieses Projekts besonders fruchtbar ist die Einbeziehung von »unspecified digits« (Library of Con-

gress 2019), die unbekannte Teile eines Datumsformats explizieren: Ein nicht spezifizierter Tag im Februar 1972 kann etwa als »1972-02-XX« dargestellt werden, der gleiche Fall bezogen auf einen Tag im Jahr 1986 als »1986-XX-XX«. Darüber hinaus muss die Ungewissheit nicht zwingend von den niedrigstwertigen Stellen herrühren – auch »XXXX-09-24« oder sogar »19XX-05-XX« sind gültige EDTF-Werte. Zwar existiert eine JavaScript-Bibliothek zum Parsen von EDTF-Datumsangaben (vgl. Keil 2022), nicht jedoch zur menschenlesbaren Darstellung. Die Logik zur sprachenübergreifenden Darstellung unvollständiger Angaben wurde deshalb eigens in TypeScript implementiert.

4. Statistische Narrative und Ausblick

Mit den vorliegenden Daten kann ein spezifischer Aspekt des literarischen Lebens in der DDR nun zum ersten Mal auch statistisch ausgewertet werden. Durch chronologische Verlaufsdiagramme zeichnen sich Einladungstendenzen ab, die sich unter anderem politisch deuten lassen.

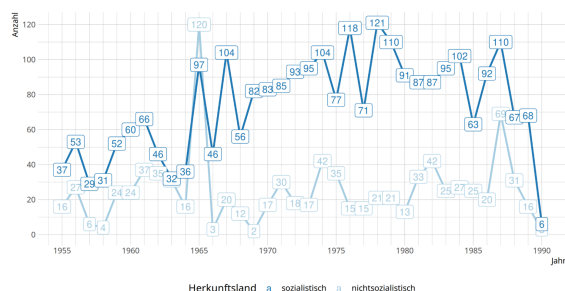


Abbildung 3: Einlaufs-frequenz von Autor*innen nach Staatsform.

So zeigt etwa Abbildung 3, dass fast durchgängig signifikant mehr Autor*innen aus sozialistischen Bruderstaaten eingeladen wurden (vgl. dazu Müller-Tamm 2021). Eine Ausnahme bildet das Jahr 1965 mit dem Internationalen Schriftstellertreffen im Mai, das in Berlin und Weimar stattfand (dokumentiert in: Deutscher Schriftstellerverband 1965).

Ein Blick auf süd- und westeuropäische Länder zeigt nur sporadische Besuche, mit der Ausnahme Frankreichs, dessen breit aufgestellte Linke teils verstärkt mit dem Schriftstellerverband der DDR kooperierte (vgl. Fabre-Renault 2015). Auch mit Autor*innen aus dem englischsprachigen Ausland, vor allem den USA und Australien, gab es noch in den 1960er-Jahren einen vergleichsweise regen Austausch, der in den 1970er-Jahren allerdings vollends zum Erliegen kam.

Dokumentieren lässt sich auch ein hohes Interesse des DSV an Autor*innen aus den sich als neutral verstehenden Staaten Finnland und Schweden, die in den 1960er-Jahren von der SED zu Schwerpunktländern auslandspropagandistischer Aktivitäten erkoren wurden:²

Im Laufe der Jahre ergingen vom Schriftstellerverband etwa 100 Einladungen in das recht dünn besiedelte Nord-europa, unter anderem an die Nobelpreisträger Halldór Laxness (Island), Eyvind Johnson (Schweden) sowie die finnische Star-Autorin und PEN-Präsidentin Eeva Joen-pelto.

Autor*innen aus Ostblockstaaten waren jedoch weit-aus regelmäßiger bei literarischen Terminen in Ostberlin zu Gast. Hier zeigen die Daten, dass sowjetische Besucher*innen stets in der Überzahl waren, ein Beleg für die Quotenregelung, die der Einladungspolitik zugrunde lag.

Die von uns angebotene Schnittstelle ermöglicht viele weitere statistische Anfragen. Ihre Funktion ist aber nicht auf die projektbezogene Auswertung beschränkt. Vielmehr kann der von uns erstellte, semantisch angereicherte Datensatz auch langfristig eine Funktion im wachsenden Ökosystem der digitalen Literaturwissenschaft übernehmen und bietet sich für den Austausch mit komplementären Projekten wie der »Forschungsplattform Literarisches Feld DDR« an (vgl. <https://ddr-literatur.de/>). Reisedaten von Autor*innen können bio-grafische Datenbanken ergänzen. Gleichzeitig konnten wir den Datensatz durch das Ausstatten mit Normdaten breiter kontextualisieren und mit Zusatzinformationen speisen. Insofern bietet der Datensatz jenseits der sta-tistischen Auswertung auch die Möglichkeit einer enzyklopädischen Nutzung für Fallstudien: Interessierte Wis-senschaftler*innen unterschiedlichster Disziplinen und Philologien³ können ihn als Einstiegspunkt für Erkundungen von bi- und multilateralen Literaturkontakten nutzen, indem sie sich innerhalb der Web-Anwendung durch Länderlisten und Personeneinträge bewegen. Auf diese Weise können sie etwa nachvollziehen, welche Autor*innen aus dem Ausland zu welchen Anlässen und Zeitpunkten über den Schriftstellerverband in die DDR entsandt wurden, in welchen Delegationen sie durch die DDR reisten, wem sie bei Kongressen, Rundtischgesprächen oder Studienaufenthalten gegenüber-saßen. Über die Verknüpfung mit Wikidata und VIAF ist es dabei auch möglich, Bezüge zu biografischen Eckdaten (Parteilzugehörigkeit, Ehrungen und Literaturpreise), Veröffentlichungen und Übersetzungen im In- und Ausland herzustellen. In Planung ist zudem ein analoges Vorhaben zu den internationalen, institutionalisierten Literaturkontakten im Westen der Stadt (DAAD, Literarisches Colloquium Berlin etc.), für das bereits etliche Daten vor-liegen. Wenngleich sich der Literaturaustausch in Ost und West schon allein durch unterschiedliche politische Strukturen nur bedingt vergleichen lässt, wäre es damit auch möglich, Korrespondenzen, (personelle) Überschneidungen und Konkurrenzsituationen zur Zeit der Berliner Mauer aus unterschiedlichen Perspektiven zu beleuchten.

Fördernachweis

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder innerhalb des Exzellenzclusters Temporal Communities: Doing Literature in a Global Perspective – EXC 2020 – Projekt-ID 390608380.

Fußnoten

1. Archiv des Schriftstellerverbandes der DDR, Literaturarchiv der Akademie der Künste Berlin, Signaturen SV 2831, 2837, 2838, 2839, 2848. Das Projekt steht im intensiven Austausch mit Gabriele Radecke und den Mitarbeiter*innen am Literaturarchiv der Akademie der Künste. Wir möchten uns an dieser Stelle sehr herzlich für die enge Zusammenarbeit bedanken.
2. Das Selbstverständnis des DSV in Bezug auf die allgemeinen Bestrebungen zur Auslandspropaganda in der DDR erläutert der Sekretär und stellvertretende Vorsitzende des Verbandes, vgl. Koch 1964.
3. »Writing Berlin« versteht sich als kollaboratives Forschungsprojekt, an dem etliche Wissenschaftler*innen u.a. aus der Germanistik, Anglistik, Romanistik oder Slawistik beteiligt sind. In dieser Weise veranstaltet das Projekt immer wieder Tagungen und veröffentlicht Sammelbände, in denen neben der globalen Perspektive auch Fallstudien und der Austausch unterschiedlicher Perspektiven im Vordergrund stehen, vgl. dazu Müller-Tamm 2021; Klengel et al. 2023.

Bibliographie

- Berbig, Roland (Hg.). 2005. *Stille Post. Inoffizielle Schriftstellerkontakte zwischen West und Ost. Von Christa Wolf über Günter Grass bis Wolf Biermann*. Berlin: Ch. Links.
- Böttiger, Helmut. 2005. *Elefantenrunden. Walter Höllerer und die Erfindung des Literaturbetriebs*. Berlin: Literaturhaus.
- Degen, Andreas (Hg.). 2011. *Szenen Berliner Literatur. 1955–1965*. Berlin: Matthes & Seitz.
- Deutscher Schriftstellerverband (Hg.). 1965. *Internationales Schriftstellertreffen Berlin und Weimar, 14.–22. Mai 1965. Protokoll*. Berlin: Aufbau.
- Fabre-Renault, Catherine. 2015. »Frankreich, ein Sonderfall? Zu Frankreichs Beziehungen zur DDR und ihrem Einfluss auf die Rezeption der DDR-Literatur.« In: Matthias Aumüller, Erika Becker (Hg.): *Zwischen literarischer Ästhetik und sozialistischer Ideologie. Zur internationalen Rezeption und Evaluation der DDR-Literatur*. Berlin. S. 18–24.
- Gansel, Carsten. 1997. »Deutschland einig Vaterland? Der Deutsche Schriftstellerverband und seine Westarbeit in den fünfziger Jahren.« In: Mark Lemstedt, Siegfried Lokatis (Hg.): *Das Loch in der Mauer. Der innerdeutsche Literaturaustausch*. Wiesbaden: Harrassowitz. S. 261–278.
- Harris, Rich. 2018. »Virtual DOM Is Pure Overhead.« In: *Svelte Blog*, 27. Dezember 2018. (<https://svelte.dev/blog/virtual-dom-is-pure-overhead>)
- Keil, Sylvester. 2022. »EDTF Parser for JavaScript.« In: GitHub. Letzter Zugriff am 15. Dezember 2022. (<https://github.com/inukshuk/edtf.js>)
- Klengel, Susanne et al. (Hg.). 2023. *Berlin International. Literaturszenen in der geteilten Stadt (1970–1989)*. Berlin/Boston: De Gruyter. [Im Druck.]
- Koch, Hans. 1964. »Zu einigen Fragen unserer Auslandsarbeit.« In: *Neue Deutsche Literatur*, 5/1964. S. 153–163.

Library of Congress. 2019. »Extended Date/Time Format (EDTF) Specification.« February 4, 2019. Letzter Zugriff am 15. Dezember 2022. (<http://www.loc.gov/standards/datetime/>)

Michael, Klaus, Margret Pötsch und Peter Walther. 1997. »Geschichte, Struktur und Arbeitsweise des Schriftstellerverbands der DDR. Erste Ergebnisse eines Forschungsprojektes.« In: *Zeitschrift des Forschungsverbundes SED-Staat*, 3/1997. S. 58–69.

Müller-Tamm, Jutta. 2021. »Das geteilte Berlin als Katalysator der Internationalisierung des Literaturbetriebs.« In: Dies. (Hg.): *Berliner Weltliteraturen. Internationale literarische Beziehungen in Ost und West nach dem Mauerbau*. Berlin/Boston: De Gruyter. S. 1–39.

Pamperrien, Sabine. 2004. *Versuch am untauglichen Objekt. Der Schriftstellerverband der DDR im Dienst der sozialistischen Ideologie*. Frankfurt/M.: Lang.

Walther, Joachim. 2006. *Sicherungsbereich Literatur. Schriftsteller und Staatssicherheit in der Deutschen Demokratischen Republik*. 2. Aufl. Berlin: Ch. Links.

Knowledge Graph-basierte Forschungsdaten- integration in NFDI4Culture

Tietz, Tabea

tabea.tietz@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für
Informationsinfrastruktur, Deutschland; Karlsruher
Institut für Technologie, Institut AIFB, Deutschland

Bruns, Oleksandra

oleksandra.bruns@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für
Informationsinfrastruktur, Deutschland; Karlsruher
Institut für Technologie, Institut AIFB, Deutschland

Fliegl, Heike

heike.fliegl@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für
Informationsinfrastruktur, Deutschland

Posthumus, Etienne

eposthumus@gmail.com
FIZ Karlsruhe – Leibniz Institut für
Informationsinfrastruktur, Deutschland

Schrade, Torsten

Torsten.Schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz,
Deutschland

Sack, Harald

harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für
Informationsinfrastruktur, Deutschland; Karlsruher
Institut für Technologie, Institut AIFB, Deutschland

Einleitung

Die Nationale Forschungsdateninfrastruktur (NFDI)¹ wurde im Jahr 2020 mit dem Ziel ins Leben gerufen, die Datenbestände aus Wissenschaft und Forschung für das deutsche Wissenschaftssystem systematisch zu erschließen, zu vernetzen und nutzbar zu machen. NFDI4Culture startete im Oktober 2020 als Konsortium der ersten Förderrunde und widmet sich Forschungsdaten zu materiellen und immateriellen Kulturgütern (Altenhöner et al. 2020). Die Interessengemeinschaft von NFDI4Culture reicht von Architektur-, Kunst- und Musik- bis hin zu Theater-, Tanz-, Film- und Medienwissenschaft und besteht aus über 70 beteiligten Organisationen². Dieses multidisziplinäre Konsortium produziert und verarbeitet eine große Menge heterogener Daten und unterhält Repositorien und Services, die für das deutsche Wissenschaftssystem und auch darüber hinaus in Kunst und Kultur von großer Bedeutung sind. Zum Datenspektrum von NFDI4Culture gehören 2D-Digitalisate von Gemälden, Fotografien und Zeichnungen ebenso wie digitale 3D-Modelle kulturhistorisch bedeutender Gebäude, Denkmäler oder audio-visuelle Daten von Musik-, Film und Bühnenaufführungen (Bicher et al., 2022). In den Datendomänen des Konsortiums werden bisher jedoch nur vereinzelt einheitliche offene Standards und Datenmodelle genutzt. Forschungsdaten liegen oftmals in sogenannten Datensilos vor, die weder von außen gefunden noch wiederverwendet werden können, da auch die Zugangsmöglichkeiten uneinheitlich und kompliziert sind. Zudem sind die Lizenzen zur Nutzung der Ressourcen oft nicht sofort ersichtlich oder vollständig geklärt, was deren Nachnutzung zusätzlich erschwert. Weiterhin stellen insbesondere auch die mit digitalen Kulturgütern zusammenhängenden, häufig komplexen datenrechtlichen und datenethischen Aspekte eine Herausforderung dar. Ziel von NFDI4Culture ist es daher, eine bedarfsgerechte Infrastruktur für die Forschungsdaten der Interessengemeinschaft zu schaffen, die den F.A.I.R. Prinzipien folgt und somit das Auffinden, den Zugang, die Nutzbarkeit und Interoperabilität der Ressourcen für alle sicherstellt (Wilkinson et al., 2016).

Die Implementierung der F.A.I.R. Prinzipien erfolgt im Konsortium über die Bereitstellung und Nutzung von domänenspezifischen Ontologien, die Entwicklung von Knowledge Graphs und die Verknüpfung einer Vielzahl von strukturierten Datenbanken und Knowledge Graphs untereinander. Ein einheitlicher und intuitiver Zugriff auf die dezentral vorliegenden Forschungsdatenressourcen des Konsortiums wird über eine zentrale Plattform, das "Culture Information Portal"³ sichergestellt.

In diesem Beitrag wird der aktuelle Stand und die weiteren Planungen der technisch übergreifenden Task Area 5 von NFDI4Culture zur Knowledge Graph-basierten In-

tegration von Forschungsdaten materieller und immaterieller Kulturgütern vorgestellt. Dies beinhaltet eine Diskussion aktueller technischer und domänenspezifischer Herausforderungen, die Vorstellung der NFDICO Ontologie und des NFDI4Culture Knowledge Graphen sowie eine Darstellung zur Implementation des Culture Information Portals.

Herausforderungen der Datenintegration in NFDI4Culture

Die von NFDI4Culture in den Blick genommene Forschungslandschaft ist durch eine starke Diversität gekennzeichnet. Sie umfasst nicht nur eine Vielzahl von Forschungsdisziplinen, sondern auch unterschiedlichste Organisationen, darunter Universitätsinstitute, Kunst- und Musikhochschulen, Akademien, Galerien, Bibliotheken, Archive, Museen und einzelne Forscher*innen. Dementsprechend sind die Forschungsprozesse- und ressourcen, die auffindbar, interoperabel und wiederverwendbar gemacht werden müssen, ebenfalls heterogen und liegen nicht nur in divergierenden Standards und Formaten vor, sondern auch in unterschiedlich aufbereiteten Zuständen. Kollektionen sind nicht immer vollständig digitalisiert und erschlossen, weshalb oft nur wenige beschreibende Metadaten zur Verfügung stehen. Andere Datensätze liegen bereits vollumfänglich als Linked Open Data vor und können problemlos mit dem NFDI4Culture Knowledge Graph verknüpft werden. Die Implementierung der F.A.I.R. Prinzipien ist ein Hauptziel des Konsortiums. Daher werden dedizierte Ontologien zur Verfügung gestellt und Maßnahmen getroffen, um alle Akteur*innen dabei zu unterstützen, eigene Daten und Ressourcen langfristig und nachhaltig selbst in Linked Open Data zu transformieren.

Ein wissenschaftsgeleitetes Forschungsdatenmanagement erfordert die aktive Teilnahme aller. Das Konsortium sieht daher umfangreiche Beteiligungsmöglichkeiten für die Nutzenden aller involvierten Fachdisziplinen, aber auch für Kunst- und Kulturschaffende unterschiedlichster Tätigkeitsbereiche und Vertreter*innen der Zivilgesellschaft vor. Es zielt darauf ab, das breite Spektrum der verschiedenen Akteur*innen im Bereich des Kulturerbes zu repräsentieren. Unter anderem ist vorgesehen, dass die Fachgemeinschaft selbst den Knowledge Graph mit eigenen Ressourcen erweitert und dessen Inhalte pflegt. Bedingung dafür ist eine technische Infrastruktur, die es ermöglicht, Ressourcen intuitiv und ohne tiefe technische Kenntnisse zu kuratieren, hinzuzufügen, zu verknüpfen und zu durchsuchen. Diese Infrastruktur wird eine semantische expressive Repräsentation der Daten ermöglichen, um eine vollumfängliche Umsetzung der F.A.I.R. Prinzipien zu gewährleisten.

NFDICO und der NFDI4Culture Knowledge Graph

Ein Ziel von Task Area 5 in NFDI4Culture ist die Bereitstellung von Ontologien, um die Forschungsdaten in NFDI4Culture standardisiert und formal zu repräsen-

tieren und miteinander verknüpfen zu können. In einem “bottom-up” Ansatz nach der sogenannten “Waterfalls” Methode (Keet, 2020) und im kontinuierlichen Austausch mit den Domänenexpert*innen des Konsortiums wurde dazu die NFDICO Ontologie entwickelt. Sie verknüpft Datensätze, Forschungsprojekte, Services, Repositorien, Institutionen und Forschungsdisziplinen und dient als Grundlage für den NFDI4Culture Knowledge Graph. Als Modellierungsgrundlage dienten die durch die NFDI4Culture Community übermittelten Beschreibungen ihrer Forschungsressourcen. Die Klasse `nfdico:Contribution`⁴ (Abb. 1) repräsentiert die Beiträge der Fachgemeinschaften und ordnet die Arten der Beiträge (z.B. Datenportal, Datensatz, Kollektion, Software, Infrastruktur, Service) unter.

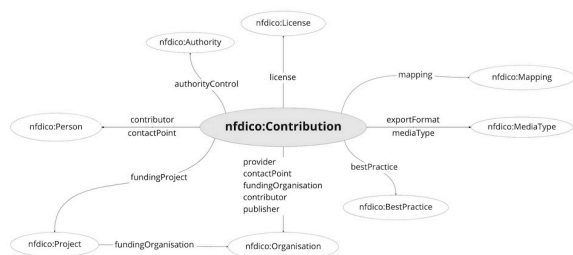


Abbildung 1: Modellierung der Klasse `nfdico:Contribution`

Instanzen können beispielsweise durch Medientypen, Lizenzangaben, zugehörige Personen und Institutionen und Projekte beschrieben werden. Ein Modellierungsbeispiel für die Klasse `nfdico:Service`⁵ als Unterklasse von `nfdico:Contribution` ist in Abb. 2 dargestellt. Die Klasse wird beispielsweise durch die akademische Disziplin, Medientypen, verwendete Normen und Ontologien spezifiziert.

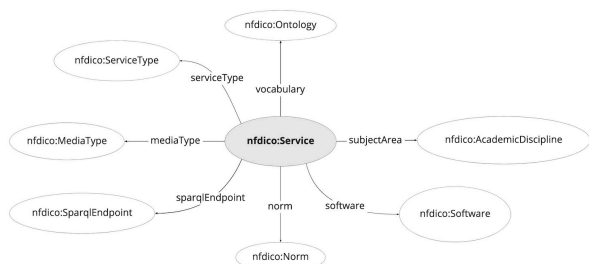


Abbildung 2: Modellierung der Klasse `nfdico:Service`

NFDICO besteht in der Version 1.1 aus 36 Klassen und 60 Objektattributen. Die spezifisch für den Anwendungsfall in NFDI4Culture definierten Klassen (Prefix `nfdico`) wurden den “best-practices” in der Ontologieentwicklung folgend zur Sicherstellung hoher semantischer Expressivität und Interoperabilität mit 24 bereits existierenden Ontologien verknüpft, darunter `frapo`⁶, `fabio`⁷, `void`⁸ und `schema`⁹ (Keet, 2020). Die Ontologie ist seit Juni 2022 öffentlich verfügbar und vollständig dokumentiert¹⁰. NFDICO folgt einem generischen Modellierungsansatz und repräsentiert die Beiträge der Fachgemeinschaft

nicht ausschließlich für NFDI4Culture Domänen, sondern ermöglicht die Nachnutzung durch weitere NFDI Konsortien anderer Fachrichtungen. So dient NFDICO beispielsweise als Grundlage der Basisontologie des NFDI Konsortiums NFDI-MatWerk¹¹. Gemäß der Waterfalls Methodologie der Ontologieentwicklung wird NFDICO in Absprache mit Domänenexpert*innen iterativ und modular erweitert.

Wie bereits o.g., dient NFDICO als Grundlage für den NFDI4Culture Knowledge Graph. Eine erste Version des Graphen ist publiziert und kann über einen öffentlichen SPARQL Endpunkt¹² abgefragt werden. Der inhaltliche Aufbau des Knowledge Graphen erfolgte auf Basis der Forschungsressourcen und beschreibenden Metadaten, die von der NFDI4Culture Community übermittelt wurden. Alle Beiträge wurden normalisiert und via RDFLib¹³ in den Knowledge Graph integriert. Außerdem wurden (soweit möglich) Verknüpfungen der Entitäten aus dem Knowledge Graph zu Wikidata¹⁴ und der Gemeinsamen Normdatei (GND)¹⁵ hergestellt. Das heißt, die Daten im NFDI4Culture Knowledge Graph sind von Anfang an anschlussfähig und können mittels föderierter SPARQL Abfragen erweitert werden. So können für alle in NFDI4Culture beteiligten Organisationen über die jeweiligen Wikidata-Verknüpfungen zusätzliche Informationen, wie zum Beispiel der Typ der Organisation (z.B. Bibliotheken¹⁶) in die Ergebnisanzeige mit einbezogen werden, obwohl diese Informationen im NFDI4Culture Knowledge Graph nicht explizit enthalten sind.

Der NFDI4Culture Knowledge Graph enthält aktuell 1796 Entitäten, darunter 173 Organisationen, 156 Forschungsressourcen, die mit ca. 10.000 RDF-Tripeln beschrieben werden. Der Knowledge Graph wird kontinuierlich erweitert und schrittweise für Beiträge durch die Community geöffnet, sodass alle Akteur*innen in Zukunft selbst eigene Ressourcen verknüpfen und öffentlich zugänglich machen können. Die Interaktion der Community mit dem Knowledge Graph erfolgt über eine einheitliche und intuitive Schnittstelle, das im Folgenden beschriebene Culture Information Portal.

Culture Information Portal und Integration des Knowledge Graph

Mit dem Culture Information Portal verfolgt NFDI4Culture das Ziel, einen einheitlichen, intuitiven und zentralen Einstiegspunkt auf die dezentral gespeicherten Forschungsdaten der Community und alle weiteren Dienste des Konsortiums (wie z.B. ein übergreifendes Helpdesk, nachnutzbare Guidelines zu allen Bereichen des Forschungsdatenmanagements oder das konsortiumsweite Identitäts- und Zugriffsmanagement für die gemeinsamen Kommunikations- und Kollaborationswerkzeuge) bereitzustellen. Forschungsdaten sollen durch die Community selbst hinzugefügt, kuratiert und auffindbar gemacht werden. Weiterhin informiert das Culture Information Portal über aktuelle Veranstaltungen und Neuigkeiten des Konsortiums, sowie über beteiligte Akteure und Projektfortschritte. Alle Informationen im Por-

tal sollen außerdem im Sinne der F.A.I.R. Prinzipien als Linked Open Data vorliegen und eine gute Anschlussfähigkeit an europäische Informationsinfrastrukturen wie die European Open Science Cloud gewährleistet sein. Das Portal wurde aus diesem Grund als webbasiertes Forschungsinformationssystem (Current Research Information System / CRIS) auf Basis des Open Source Content Management Systems TYPO3¹⁷ realisiert. Das Datenmodell des Portals orientiert sich am CERIF-Standard der euroCRIS¹⁸, wodurch gleichzeitig eine sehr gute Kompatibilität zu NFDICO gegeben ist. Für die CRIS Implementierung wurde die TYPO3-Extension "Academy" nachgenutzt und weiterentwickelt¹⁹.

Zur Integration des Culture Knowledge Graphen wurden mit dem Ziel einer förderierten Informationsinfrastruktur zunächst existierende Systeme auf die oben genannten Kriterien überprüft, darunter Wikibase²⁰ und WissKI²¹. Wikibase ist eine freie Software zur kollaborativen Kuratierung von Datenbanken mit der Möglichkeit, Daten im Sinne von LOD zu strukturieren, zu verknüpfen und abzufragen. Wikibase bietet allerdings nicht die Möglichkeit, externe, bereits existierende Ontologien zu integrieren, was die semantische Expressivität der Daten stark begrenzt. Außerdem ist es mit Wikibase nicht möglich, eigene Eingabemasken zu entwickeln, was die intuitive Interaktion mit dem Knowledge Graph einschränkt. WissKI ist eine webbasierte Software zum Sammeln, Strukturieren und Teilen von forschungsbezogenen Daten. Feste Grundlage von WissKI bildet dabei die Optimierung der Software auf CIDOC-CRM, was sich für den hier beschriebenen Anwendungsfall als zu einschränkend gezeigt hat. Die Modellierung in CIDOC-CRM folgt einem Ereignis-basierten Paradigma. Dadurch gerät die Modellierung einfacher Fakten oft sehr komplex und behindert den typischen Anwendungsfall in NFDI4Culture. Durch CIDOC werden einfache SPARQL-Abfragen zu Personen und Organisationen oft hochkomplex und daher ineffizient. Aufgrund der großen Menge an Daten, die in NFDI4Culture erwartet werden, ist eine Abfrage-Effizienz allerdings relevant. Der Zugang zu NFDI4Culture Daten soll für alle Nutzer*innen so einfach wie möglich gestaltet werden, eine CIDOC-CRM-basierte Modellierung schafft aufgrund ihrer Komplexität zusätzliche Barrieren. Dennoch ist CIDOC ein wichtiger und gängiger Bestandteil der GLAM Community. Daher wird ein Mapping des NFDI4Culture Datenmodells nach CIDOC durchgeführt, um einen CIDOC-basierten Export zu gewährleisten.

Als beste Lösung zur Gewährleistung der semantischen Expressivität der erfassten Ressourcen und Metadaten mittels NFDICO und weiterer Ontologien einerseits bei gleichzeitiger Umsetzbarkeit der benötigten Kurationsmechanismen andererseits stellte sich die direkte Implementierung der benötigten Funktionalitäten in TYPO3 heraus. Die TYPO3 Extension "LOD"²² bietet hierbei einen unmittelbar über der relationalen Datenbank des CMS realisierten "Semantic Layer" mit einem konfigurierbaren IRI-Generator sowie IRI-Resolver für alle Datensätze sowie einem RDF-Serialisierer für alle Datenbankinhalte. Alle im Portal bzw. Culture CRIS erfassten Ressourcen werden dabei über eine standardisierte LOD API unter Verwendung des Hydra Core Vocabulary²³ in verschiedenen RDF Serialisierungen (z.B.

RDFa, Turtle, JSON-LD u.a.) veröffentlicht²⁴. Hierdurch können die Daten für den Culture Knowledge Graph bereits jetzt im Kreis der Mitarbeitenden des Konsortiums dezentral kuratiert und kontinuierlich erweitert werden. Die über die LOD API des CMS publizierten Daten werden mittels Ingest-Routinen in den eigentlichen Knowledge Graph integriert. Zum Einsatz kommt oxigraph²⁵ als nativer RDF-Store mit einem pythonbasierten Wrapper²⁶ für ein leichtgewichtiges Deployment des öffentlichen SPARQL-Endpoints, der über ein grafisches Interface wiederum direkt in das Culture Information Portal integriert ist.

Weiteres Vorgehen und Zusammenfassung

NFDI4Culture öffnet Forschungsdatensilos und schafft einheitliche und intuitive Zugriffsmöglichkeit auf Forschungsdaten. Dieses Vorhaben wird in der Task Area 5 durch die Bereitstellung dedizierter Ontologien, die Implementation und Verknüpfung von Knowledge Graphen und die Umsetzung des Culture Information Portals erreicht. Akteur*innen der Community können mit ihrer Beteiligung am Vorhaben ihre Daten also auffindbar, interoperabel und wiederverwendbar machen, die Zitierfähigkeit eigener Ressourcen gewährleisten und die multidisziplinäre Verknüpfungen der Daten für ihre Forschung ausnutzen. Alle hier präsentierten Beiträge sind öffentlich verfügbar und nutzbar. Im weiteren Vorgehen wird die Ontologie bedarfsorientiert erweitert und das Culture Information Portal schrittweise für die Community geöffnet werden, um mit den Inhalten im Knowledge Graph zu interagieren und den Knowledge Graph mit Forschungsdaten anzureichern und diese zu kuratieren. Die technische Infrastruktur wird stetig verbessert, um unter anderem teil-automatisierte Qualitätskontrollen der Daten zu umzusetzen (z.B. durch SHACL²⁷) und die Integration größerer Datenmengen über den SPARQL Endpunkt zu ermöglichen.

Vergleichbare und verwandte Initiativen werden mit der "Linked Data Platform Finland"²⁸ und dem "Dutch Digital Heritage Network"²⁹ umgesetzt. Auch NFDI4Culture ist eine nationale Initiative, gleichsam ist allen Beteiligten die Bedeutung einer globalen Vernetzung bewusst, denn auch deutsche Kulturdaten und deutsches kulturelles Erbe sind international und von internationalem Interesse. In NFDI4Culture wurden bereits Maßnahmen implementiert, um auch Organisationen und Forschende außerhalb des Konsortiums zu involvieren. Beispielsweise dienen dazu das Steering Board³⁰ und die "Linked Open Data Working Group". Technisch ist eine Vernetzung mit anderen Plattformen und die Nutzung international anerkannter Standards ebenso elementar. Der NFDI4Culture Knowledge Graph enthält bereits Verknüpfungen zu Wikidata, die mit Hilfe der Nutzer*innen stetig erweitert werden. Zudem besteht auf der NFDI4Culture Ontologie-Ebene bereits ein Mapping mit dem "Common European Research Information Format (CERIF)". Diese Verknüpfungen werden in Zukunft weiter ausgebaut werden.

Unter Umsetzung der F.A.I.R. Prinzipien und mit der Beteiligung der gesamten Community wird in NFDI4Culture eine Forschungsinfrastruktur geschaffen, die langfristig und nachhaltig Forschungsdaten auffindbar, zugreifbar und nutzbar für alle macht.

Fußnoten

1. <https://www.nfdi.de/>
2. <https://nfdi4culture.de/>
3. <https://nfdi4culture.de/>
4. <https://nfdi4culture.de/ontology#Contribution>
5. <https://nfdi4culture.de/ontology#Service>
6. <http://www.sparontologies.net/ontologies/frapo>
7. <http://www.sparontologies.net/ontologies/fabio>
8. <https://www.w3.org/TR/void/>
9. <https://schema.org/>
10. <https://nfdi4culture.de/ontology>
11. <https://nfdi-matwerk.de/>
12. <https://nfdi4culture.de/de/resources/knowledge-graph.html>
13. <https://github.com/RDFLib/rdfliib>
14. <https://www.wikidata.org/>
15. <https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd.html>
16. SPARQL Query
17. <https://typo3.org/cms>
18. <https://eurocris.org/>. CERIF ist ein von der Europäischen Union empfohlenes, standardisiertes Datenmodell, vgl. <https://openaire-guidelines-for-cris-managers.readthedocs.io/en/v1.1.1/introduction.html>
19. <https://github.com/digicademy/academy>
20. <https://wikiba.se/>
21. <https://wiss-ki.eu/>
22. <https://github.com/digicademy/lod>
23. <https://www.hydra-cg.com/spec/latest/core/>
24. <https://nfdi4culture.de/resource/>
25. <https://github.com/oxigraph/oxigraph>
26. <https://github.com/epoz/shmarql>
27. <https://www.w3.org/TR/shacl/>
28. <https://www.ldf.fi/>
29. <https://netwerkdigitaalerfgoed.nl/en/>
30. <https://nfdi4culture.de/about-us/consortium.html#c256>

Bibliographie

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

Bicher, Katrin et al. 2022. "Digitalisierung des Kulturellen und digitale Arbeitskultur im Forschungsverbund NFDI4Culture. Community-Arbeit an, durch und mit fachspezifischen Datenkorpora und Elementen der FDM-Infrastruktur". Zeitschrift für Bibliothekswesen und Bibliographie, Jahrgang 69, Heft 1-2, S. 26-36. <http://dx.doi.org/10.3196/1864295020691258>

Keet, Maria. 2020. "Methodologies for Ontology Development." An Introduction to Ontology Engineering, Engineering Library.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3, no. 1: 1-9. <https://doi.org/10.1038/sdata.2016.18>

Knowledge Graph Design in der Forschungspraxis. Beschreibung, Interpretation und Kontextualisierung heraldischer Quellen mit der Digital Heraldry Ontology

Schneider, Philipp

philipp.schneider.1@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Hiltmann, Torsten

torsten.hiltmann@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Der Vortrag gibt einen aktuellen Überblick zur Modellierung und Implementierung der *Digital Heraldry Ontology*. Diese entsteht im Rahmen der Entwicklung eines Knowledge Graphen zur Erfassung, Analyse und historischen Kontextualisierung von heraldischen Darstellungen des Mittelalters und der Frühen Neuzeit unter Berücksichtigung divergierender Beschreibungen und Interpretationsperspektiven.

Knowledge Graphen in der historischen Forschung

Historische Daten unterliegen einer hohen Komplexität. Sie sind in der Regel lückenhaft, ambivalent, multiperspektivisch, vage, z.T. widersprüchlich sowie stark an eine konkrete Zeit und einen konkreten Kontext gebunden. Gleichzeitig gibt es ein wachsendes Bewusstsein dafür, Forschungsdaten nachnutzbar und interoperabel zur Verfügung zu stellen (Cremer 2021). Eine zunehmend genutzte Technologie, um mit diesen Herausforderungen umzugehen, sind Knowledge Graphen auf Grundlage von Semantic Web Technologien. Hierbei werden Daten in Form von Graphdatenbanken zur Verfügung gestellt, die durch Ontologien im Sinne einer formalisier-

ten *shared conceptualisation* innerhalb einer Fachcommunity strukturiert (Studer 1998) und nach den Prinzipien von Linked Open Data interoperabel und flexibel veröffentlicht werden. Durch diese Interoperabilität kann die Vernetztheit und Interdependenz von historischem Wissen berücksichtigt werden.

Dabei ist ein großer Vorteil von Knowledge Graphen, dass sie historische Daten in einer Weise abbilden, die sowohl menschen- als auch maschinenlesbar ist. Sie werden daher als *symbolische KI* verstanden, die, im Gegensatz zu Formen der *sub-symbolischen KI*, wie maschinellem Lernen, eine explizite und von Domänenexpert:innen modellierte Repräsentation von Wissen darstellt (Sack 2021). Dadurch lassen sich die allgemeinen Herausforderungen historischer Daten leichter adressieren.

Knowledge Graphen besitzen in vielerlei Hinsicht für die historischen Wissenschaften innovatives Potenzial, das aktuell noch nicht völlig ausgeschöpft ist: Das betrifft die epistemologischen Anforderungen an die Modellierung historischer Daten mit Ontologien (Pierazzo 2019), die Möglichkeit der Kombinierbarkeit und gemeinsamen Auswertbarkeit interoperabler verschiedener Datensätze durch Linked Data (Vogeler 2020) sowie den Einsatz von *reasoning engines*, die durch logische Inferenzen neues Wissen aus bereits vorhandenen Daten ableiten können (Hogan 2021; Ehrlinger 2016).

Die Bedeutung von Knowledge Graphen ist daher in den Digital Humanities in den letzten Jahren stark angewachsen und wird in Zukunft wohl noch weiter zunehmen. So stellen etwa große Infrastrukturprojekte, wie NFDI4Memory,¹ NFDI4Culture,² Virtual Record Treasury of Ireland³ oder Biblissima,⁴ diese Technologie in den Mittelpunkt. Aber auch auf spezifische Themengebiete fokussierte Forschungsprojekte nutzen Knowledge Graphen zur Strukturierung und Bereitstellung ihrer Daten sowie zur Erkenntnisgewinnung.⁵

Die Bedeutung von Wappen für die kulturhistorische Forschung

Warum ist diese Technologie für die Erforschung historischer Heraldik des Mittelalters und der Frühen Neuzeit erforderlich? In diesen Epochen waren Wappen der am häufigsten gebrauchte und in Europa meist verbreitete Träger visueller Kommunikation (Hiltmann 2019; Hablot 2017). Verwendet wurden sie von fast allen sozialen Schichten und Gruppen. Wappen dienten dabei neben der Identifikation ihrer Besitzer:innen vor allem verschiedensten komplexen Kommunikationsakten, wie der Markierung von Besitz, Herrschaft, Jurisdiktion, Gruppenzugehörigkeit, Verwandtschaft, oder Ehre (Paravicini 1998). Damit erlaubt ihre Analyse wichtige Einblicke in vormoderne Gesellschaftsstrukturen und Kultur (Hofman 2021) – sie sind also als eigene geschichtswissenschaftliche (visuelle) Quelle zu betrachten.

Das betrifft auch den Aspekt der historischen Quellenkritik. Wappen wurden auf den unterschiedlichsten Materialien angebracht – von Handschriften über Urkunden

oder Siegel und Münzen bis hin zu Wandmalereien, Teppichen und Möbeln – um nur einige Beispiele zu nennen (Biewer 2003). Dabei konnte dasselbe Wappen – abhängig von diesen konkreten materiellen Kontexten aber auch vom jeweiligen historischen Verwendungskontext – an unterschiedlichen Stellen eine unterschiedliche Bedeutung tragen. Diesen Kontexten muss daher bei der Beschreibung und geschichtswissenschaftlichen Analyse von Wappen Rechnung getragen werden.

Trotz ihrer historischen Bedeutung haben heraldische Quellen in den Geschichtswissenschaften nur eine überschaubare Berücksichtigung genossen. Zurückzuführen ist das nicht zuletzt auf den Umfang ihrer Überlieferung, ihre Komplexität sowie die beschriebene Ambiguität (Hiltmann 2019). Auch die Menge an überlieferten Wappen ist eine Herausforderung. Dies betrifft zum einen die Menge an Wappen auf unterschiedlichen Objekten, zum anderen die Menge verschiedener Wappenbilder, die überhaupt (z.T. auf mehreren Objekten) verwendet wurden. Hier gibt es Schätzungen von ca. einer Million verschiedener Wappenbilder die allein im Mittelalter im Umlauf waren (Pastoreau 2018: 42). Um einen leichteren Zugang zu dieser Form historischer Quellen zu ermöglichen, wird im Rahmen des Projekts *Die Performanz der Wappen*⁶ an der Humboldt-Universität zu Berlin ein Knowledge Graph entwickelt, der einerseits Wappenbeschreibungen zur Verfügung stellt sowie andererseits Aufschluss darüber gibt, wo, wann und von wem diese Wappen verwendet wurden – ebenso, welche Bedeutung diese Wappen in den jeweiligen Verwendungskontexten vermutlich getragen haben. Damit soll der Knowledge Graph einerseits als datenhaltende Infrastruktur genutzt werden können, die diese Art historischer Quellen für die historische Forschung erstmals umfangreich und nachhaltig verfügbar macht. Vor allem aber soll der Knowledge Graph zum anderen dafür genutzt werden, konkrete geschichtswissenschaftliche Forschungsfragen beantwortbar zu machen. Innerhalb des Projekts betrifft das insbesondere Fragen zur Entwicklung und Diversifizierung heraldischer Kommunikation über die Zeit (Hiltmann 2019).

Methodologie zur Entwicklung einer Ontologie für die historische Forschung

Diese beiden Anforderungen – die Bereitstellung einer flexiblen und nachhaltigen Infrastruktur sowie die Berücksichtigung von Forschungsfragen – mussten bei der Entwicklung der Ontologie und der gewählten Methodologie beachtet werden. Sie orientierte sich an aktuellen Best Practices des *Knowledge Engineering*.

Diese umfassen erstens insbesondere ein durch Fragen getriebenes Design. Die grundlegenden Konzepte einer Ontologie werden hierbei auf Grundlage von sogenannten *competency questions* entwickelt (Shimizu 2020). Hierbei handelt es sich um natürlichsprachliche Forschungsfragen, die mit einer fertigen Ontologie beantwortbar sein müssen. Competency questions geben den inhaltlichen Rahmen vor, was zu modellieren ist. Der vorgestellte Knowledge Graph soll etwa nach verschiede-

nen Aspekten durchsuchbar sein, die miteinander kombinierbar sein sollen. Gesucht werden soll sowohl nach einzelnen Wappenbeschreibungen als auch nach Verwendungskontexten bestimmter Wappen, ebenso danach welche Wappen auf einem bestimmten Objekt abgebildet sind, sowie welche(s) Wappen zu einer bestimmten Zeit von einer bestimmten Person verwendet wurde(n). Ein hieraus abgeleitetes Beispiel für eine *competency question* wäre somit z.B.:

Auf welchen Wandmalereien und in welchen Buchhandschriften, die sich alle auf einen Zeitraum zwischen 1400 und 1500 datieren lassen, findet sich eine identische Sequenz an Wappenbildern, die in der gleichen Reihenfolge auftreten, und welchen Entitäten sind die einzelnen Wappen in den jeweiligen Kontexten zugeordnet?

Diese Vielfalt unterschiedlicher zu modellierender und ineinandergreifender Aspekte verweist zugleich auf das zweite Kernthema des modernen *Knowledge Engineering* mit Ontologien: Wichtig ist auch eine möglichst starke Modularisierung der Architektur eines Knowledge Graphen. Dies erfolgt in der Regel über sogenannte *Ontology Design Patterns*. Hierbei handelt es sich um kleinere Teilontologien, die für sich genommen einen bestimmten Sachverhalt abbilden (etwa die Modellierung von menschlichen Personen). Dies ermöglicht es, Teilprobleme getrennt zu modellieren und dadurch eine flexiblere Nachnutzbarkeit und Anschlussfähigkeit der *Ontology Design Patterns* in anderen Kontexten zu gewährleisten (Shimizu 2020).

Aus geschichtswissenschaftlicher Sicht sind weitere zusätzliche Anforderungen zu erfüllen, die als Grundprinzipien bei der Ontologieentwicklung gedient haben. Hierzu zählt insbesondere eine möglichst große Nähe zu den modellierten Quellen, wodurch Erkenntnisse der historischen Grundwissenschaften in Datenmodelle übersetzt werden. Als zentral wurde dabei auch die Trennung zwischen Beschreibung und Interpretation einer Quelle im Datenmodell erachtet (Beretta 2021; Theissen-Lipp 2020). Historische Quellenkritik muss also inhärenter Bestandteil eines Modellierungsprozesses sein. Aus der Sicht des *Knowledge Engineering* bedeutet dies, dass die Umsetzung theoretischer epistemologischer Überlegungen zur Repräsentation historischer Daten in konkrete Ontologien (Beretta 2021; Eide 2019) eine Forschungsleistung des Projekts darstellt.

Semantic Web Technologien zur Beschreibung und Kontextualisierung visueller Quellen

Im Vortrag wird ein Überblick zum aktuellen Entwicklungsstand der für diesen Knowledge Graph erforderlichen Ontologie gegeben. Um die unterschiedlichen Aspekte heraldischer Kommunikation zu erfassen sind insgesamt fünf Teilontologien mit jeweils eigenen Funktionen erforderlich:

- *heraldry*: System zur Beschreibung von Wappen
- *blazon*: Bereitstellung von Wappen als konzeptuellen Bildern

- *representations*: Beschreibung konkreter materieller Repräsentationen dieser konzeptuellen Bilder
- *objects*: Objekte zu denen diese Repräsentationen gehören
- *entities*: Personen, Gruppen oder andere Entitäten, die durch Wappen repräsentiert werden können

Die erste Ontologie (*heraldry*) dient der Beschreibung von Wappen, basierend auf der Praxis und Terminologie des Blasonierens (Hiltmann 2022; Hiltmann 2020). Entscheidend ist dabei, dass es sich bei Wappen trotz ihrer Visualität nicht um Quellen handelt, die sich ausschließlich als Bilder darstellen lassen – vielmehr sind Wappen *konzeptuelle, strukturelle* Bilder (Pastoureaux 1979). Konkret bedeutet dies, dass jedes Wappen eine Kombination aus bestimmten, in unterschiedlichen Wappen immer wieder verwendeten Figuren, Farben und geometrischen Mustern darstellt. Sowohl diese visuellen „Bausteine“ als auch die Art wie sie auf einem Wappen angeordnet sind, werden durch eine halbwegs formale Terminologie und Fachsprache erfasst (*blason* bzw. *blazon*) (vgl. Abbildung 1). Diese heraldischen Konzepte werden in der Ontologie formalisiert und erlauben so eine sowohl von Menschen als auch von Maschinen lesbare, verstehbare und analysierbare Beschreibung von Wappen. Auf diese Weise können beliebige heraldische Kombinationen erstellt werden, die dann als Linked Data eindeutig referenziert werden können. Grundlage für die Klassen und Properties dieser Teilontologie waren daher existierende heraldische Konzepte.

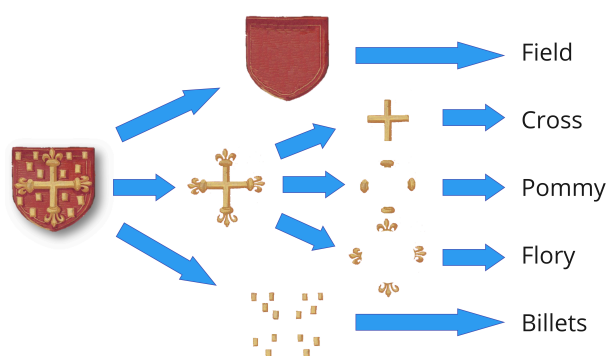


Abbildung 1: Aufteilung heraldischer Begriffe

Eine zweite Ontologie (*blazon*) speichert die unterschiedlichen Beschreibungen von Wappen, unabhängig von ihrem jeweiligen konkreten Verwendungskontext. Dadurch wird jede erdenkliche Wappenbeschreibung mit einer URI eindeutig referenzierbar.

Eine weitere Ontologie (*representation*) modelliert das eindeutige Auftreten einer Wappenbeschreibung in einem konkreten materiellen Kontext auf einem bestimmten Objekt. Dieses kann z.B. eine Handschrift, eine Wandmalerei, oder ein Siegel sein. Auf diese Weise lässt sich disambiguieren in welchem konkreten Kontext ein bestimmtes Wappen verwendet worden ist.

Diese materiellen Kontexte werden in einer weiteren Ontologie (*objects*) repräsentiert. Hier erfolgten ihre quellenkritische Beschreibung und historische Kontextualisierung mit objektspezifischen Metadaten. Dazu zählt beispielsweise die Datierung des jeweiligen Objekts,

aber auch sein Verwendungszeitraum und -kontext, mögliche Auftraggeber:innen, oder der heutige Aufbewahrungsort. Einige dieser Objektmetadaten unterscheiden sich dabei bei verschiedenen Quellentypen – so erfordert die Beschreibung eines Siegels etwa andere Metadaten als die Beschreibung einer Wandmalerei.

Eine letzte Ontologie (*entities*) erfasst schließlich die verschiedenen Entitäten, d.h. Familien, Personen, Institutionen oder auch abstrakte Konzepte, denen in einem bestimmten Kontext eine bestimmte Wappenbeschreibung zugeordnet werden kann. Hierdurch wird also abgebildet, was durch ein Wappen kommuniziert werden kann. Die Auslagerung dieser Aspekte in eine eigene Teilontologie ist dabei notwendig, um modularisierte Klassenstrukturen zu den wappentragenden Entitäten repräsentieren zu können. So lassen sich hier etwa Verwandtschaftsverhältnisse oder territoriale Bezüge abbilden, wodurch flexiblere Abfragen an den Knowledge Graphen möglich werden.

Objects und *entities* beschreiben somit materielle Objekte und wappentragende Entitäten in einem spezifisch heraldischen Kontext. Gleichzeitig ist jedoch auch für diese Teilontologien eine Verknüpfung zu bereits bestehenden (Norm-)Datensätzen im Sinne von Linked Open Data geplant. Über bereits etablierte top-level-Ontologien wie LIDO oder CIDOC CRM können so interoperable Verknüpfungen hergestellt und existierende Metadaten zu wappentragenden *objects* und *entities* nachgenutzt werden.

Mit dem beschriebenen Ansatz können heraldische Darstellungen auf unterschiedlichen Quellen und die dazugehörigen Informationen mit anderen Quellenkorpora kombiniert und gemeinsam analysiert werden.

schrift. Diese konkrete Abbildung in einer Quelle wird durch eine Entität im Knowledge Graphen repräsentiert (durch die URI bei (a) in Abbildung 2). Für dieses Wappenbild existieren zwei divergierende Beschreibungen. Diese beiden unterschiedlichen Wappenbeschreibungen werden jeweils durch ein Beschreibungs-Event (b) mit der Abbildung des Wappens in der Handschrift im Knowledge Graphen verbunden. Die Beschreibungs-Events können dabei durch weitere Metadaten, wie die Entität, die die jeweilige Beschreibung dem Wappenbild in der Quelle zugewiesen hat, angereichert werden. Die eigentlichen Wappenbeschreibungen werden ebenfalls jeweils durch eine eigene Entität repräsentiert (c). Konkret wird hier die Blume im abgebildeten Wappen einmal als „cinquefoil“ und einmal als „rose“ beschrieben.

Den üblichen Herausforderungen bei der Modellierung historischer Daten wie Ambiguität, Vagheit, Unvollständigkeit wird durch diese Repräsentation von Multiperspektivität in der Modellierung der Ontologie(n) Rechnung getragen.

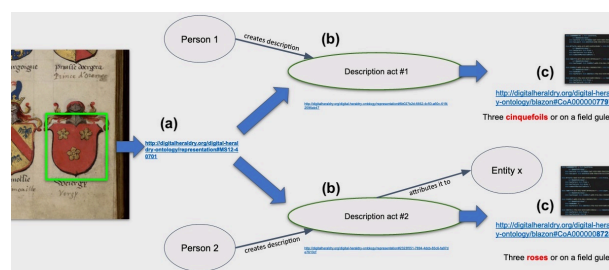


Abbildung 2: Modellierung von Multiperspektivität bei der Beschreibung von Wappen

Multiperspektivität von Beschreibungen und Interpretationen

Darüber hinaus wird jede Form einer Zuordnung event-basiert modelliert – sei es die Zuordnung einer Wappenbeschreibung zu einem konkreten Wappenbild, das auf einem historischen Objekt abgebildet ist oder die Zuweisung eines Wappenbildes in einem bestimmten historischen Kontext zu einer Person oder anderen Entität. Das bedeutet, dass auch mehrere widersprüchliche Zuweisungen innerhalb desselben Kontextes existieren können. Hintergrund ist, dass solche Zuweisungen als historiographische Interpretation verstanden werden müssen, die zunächst als gleichwertig zueinander zu betrachten sind. Eine Zuweisung eines Wappens zu einem Wappenträger ist keinesfalls immer eindeutig und auch die Beschreibung eines Wappenbildes lässt sich auf Grund materieller Fragmentarität historischer Quellen nicht immer eindeutig durchführen. Die Entität, die für diese Zuweisung verantwortlich ist, kann dabei etwa eine moderne Forscher:in sein, aber auch die Umschrift eines Siegels oder der Text in einem mittelalterlichen Manuskript.

Abbildung 2 zeigt einen vereinfachten Überblick zur Funktionsweise dieser event-basierten multiperspektivischen Beschreibung eines Wappenbildes in einer Hand-

Fußnoten

1. <https://4memory.de/> (Zugriff am 03.08.2022).
2. <https://nfdi4culture.de/> (Zugriff am 03.08.2022).
3. <https://beyond2022.ie/> (Zugriff am 03.08.2022).
4. <https://www.biblissima.fr/> (Zugriff am 03.08.2022).
5. Einige ausgewählte Beispiele sind Dutch Golden Agents, <https://www.goldenagents.org/> (Zugriff am 03.08.2022); Sphaera, <https://sphaera.mpiwg-berlin.mpg.de/> (Zugriff am 03.08.2022); Polifonia, <https://polifonia-project.eu/> (Zugriff am 03.08.2022).
6. <https://www.geschichte.hu-berlin.de/de/bereiche-und-lehrstuehle/digital-history/forschung/die-performanz-der-wappen-zur-entwicklung-von-funktion-und-bedeutung-heraldischer-kommunikation-in-der-mittelalterlichen-kultur-12-15-jahrhundert> (Zugriff am 02.08.2022).

Bibliographie

- Beretta, Francesco. „A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)“. *Semantic Web* 12, Nr. 2 (2021): 279–94. <https://doi.org/10.3233/SW-200416>.
- Biewer, Ludwig. „Wappen als Träger von Kommunikation im Mittelalter: Einige ausgewählte Beispiele“. In *Me-*

dien der Kommunikation im Mittelalter, herausgegeben von Karl-Heinz Spieß, 139–154. Stuttgart, 2003.

Cremer, Fabian, Silvia Daniel, Marina Lemaire, Katrin Moeller, Matthias Razum, und Arnost Stanzel. „Data Meets History: A Research Data Management Strategy for the Historically Oriented Humanities“. In *Band 21 Cultural Sovereignty beyond the Modern State: Space, Objects, and Media*, herausgegeben von Gregor Feindt, Bernhard Gissibl, und Johannes Paulmann, 155–78. Berlin / Boston: De Gruyter Oldenbourg, 2021. <https://doi.org/10.1515/9783110679151-009>.

Ehrlinger, Lisa, und Wolfram Wöb. „Towards a Definition of Knowledge Graphs“. In *SEMANTICS 2016: Posters and Demos Track*, 4. Leipzig, 2016.

Eide, Øyvind, und Christian-Emil Smith Ore. „Ontologies and data modeling“. In *The Shape of Data in the Digital Humanities. Modeling Texts and Text-based Resources*, herausgegeben von Julia Flanders und Fotis Jannidis, 178–196. Digital Research in the Arts and Humanities. Abingdon, 2019.

Hablot, Laurent. „Heraldic imagery, definition and principles“. In *The Routledge Companion to Medieval Iconography*, herausgegeben von Colum Hourihane, 386–398. New York, 2017.

Hiltmann, Torsten. „Zwischen Grundwissenschaft, Kulturgeschichte und digitalen Methoden. Zum aktuellen Stand der Heraldik“. *Archiv für Diplomatik* 65 (2019): 287–319.

Hiltmann, Torsten, und Thomas Riechert. „Digital Heraldry. The State of the Art and New Approaches Based on Semantic Web Technologies“. In *L'édition en ligne de documents d'archives médiévaux*, 102–25. Turnhout, 2020.

Hiltmann, Torsten, und Philipp Schneider. „Digital Heraldry Ontology“. Ontology Specification Draft, 2022. <http://digitalheraldry.org/digital-heraldry-ontology/heraldry/0.1.0>.

Hofman, Elmar. *Armorial in Medieval Manuscripts. Collections of Coats of Arms as Means of Communication and Historical Sources in France and the Holy Roman Empire (13th - Early 16th Centuries)*. Bd. 4. Heraldic Studies. Ostfildern, 2021.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, u. a. „Knowledge Graphs“. *ACM Computing Surveys* 54, Nr. 4 (2021): 71:1–71:37. <https://doi.org/10.1145/3447772>.

Paravicini, Werner. „Gruppe und Person. Repräsentation durch Wappen im späteren Mittelalter“. In *Die Repräsentation der Gruppen. Texte - Bilder - Objekte*, herausgegeben von Otto Gerhard Oexle und Andrea von Hülsen-Esch, 141:327–390. Veröffentlichungen des Max-Planck-Instituts für Geschichte. Göttingen, 1998.

Pastoureau, Michel. *Traité d'héraldique*. Grands manuels Picard. Paris, 1979.

Pastoureau, Michel. *L'Art héraldique au Moyen Age*. Paris: Le Seuil, 2008.

Pierazzo, Elena. „How Subjective Is Your Model?“ In *The Shape of Data in the Digital Humanities: Modeling Texts and Text-Based Resources*, herausgegeben von Julia Flanders und Fotis Jannidis, 1. Aufl., 117–32. Digital Research in the Arts and Humanities. London, New York: Routledge, Taylor & Francis Group, 2019. <https://www.taylorfrancis.com/books/9781315552941>.

Sack, Harald. „Hybride Künstliche Intelligenz in der automatisierten Inhaltsschließung“. In *Qualität in der Inhaltsschließung*, herausgegeben von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl, und Hans Schürmann, 70:387–406. Bibliotheks- und Informationspraxis. Berlin, 2021. <https://doi.org/10.1515/9783110691597-019>.

Shimizu, Cogan, Pascal Hitzler, und Adila Krishnadi. „Modular Ontology Modeling: A Tutorial“. In *Applications and Practices in Ontology Design, Extraction, and Reasoning*, herausgegeben von Giuseppe Cota, Marilena Daquino, und Gian Luca Pozzato, 1. Aufl. Bd. 49. Studies on the Semantic Web. Amsterdam: IOS Press, 2020. <https://doi.org/10.3233/SSW200032>.

Studer, Rudi, V. Richard Benjamins, und Dieter Fensel. „Knowledge Engineering: Principles and Methods“. *Data & Knowledge Engineering* 25, Nr. 1 (1998): 161–97. [https://doi.org/10.1016/S0169-023X\(97\)00056-6](https://doi.org/10.1016/S0169-023X(97)00056-6).

Theissen-Lipp, Johannes, Lars Gleim, und Stefan Decker. „Towards Reusability in the Semantic Web: Decoupling Naming, Validation, and Reasoning“. In *11th Workshop on Ontology Design and Patterns @ The 19th International Semantic Web Conference*. Athens, 2020.

Vogeler, Georg. „Das Semantic Web als Giant Global Kontext?“ In *Rekontextualisierung als Forschungsparadigma des Digitalen*, herausgegeben von Simon Meier, Gabriel Viehhauser, und Patrick Sahle, 14:71–80. Schriften des Instituts für Dokumentologie und Editorik. Nordstedt, 2020.

Konflikte als Theorie, Modell und Text – Ein kategorientheoretischer Zugang zur Operationalisierung von Konflikten

Gerstorfer, Dominik

dominik.gerstorfer@tu-darmstadt.de
TU-Darmstadt, Deutschland

Gius, Evelyn

evelyn.gius@tu-darmstadt.de
TU-Darmstadt, Deutschland

Einleitung

Für eine theoriegeleitete Analyse von Texten muss man eine Operationalisierung finden, die die theoretischen Konzepte im Text identifizierbar und damit messbar macht. Dies gilt in nicht-digitalen Forschungskontexten gleichermaßen wie in den Digital Humanities, allerdings ist in letzteren das Problem virulenter. Dies hat Moretti

(2013) bereits prominent festgestellt. Jenseits von konkreten Fragestellungen (wie etwa von Moretti selbst oder in Fischer & Trilcke 2016) fehlen allerdings generell umsetzbare Vorschläge.

In diesem Beitrag wollen wir an einem literaturwissenschaftlichen Anwendungsfall zeigen, wie die angewandte Kategorientheorie Operationalisierung als einen deutlich(er) definierten Workflow ermöglicht, mit dem man von einer theoretischen Grundlage über ein Modell zur Textanalyse – und zurück – kommen kann. Durch den theorieorientierten Fokus der operationalisierten Konzepte, ihre höhere Granularität sowie ihre Kompositionalität bietet der Workflow zudem eine Reihe von Vorteilen, die computationale Analysen zugleich besser und einfacher machen können.

Angewandte Kategorientheorie

Um den Anforderungen sowohl der geisteswissenschaftlichen als auch der informatischen Komponenten der Digital Humanities gerecht zu werden, ist es nötig, einen möglichst flexible und abstrakte – d.h. inhaltsagnostische – Grundlage zu finden. Wir greifen bei unserem Unterfangen auf die angewandte mathematische Kategorientheorie zurück, da sie die Möglichkeit bietet, sowohl die Explikation und Modellierung geisteswissenschaftlicher Fragestellungen zu unterstützen als auch, die Anschlussfähigkeit an informatische Methoden zu gewährleisten (Ehrig 2001).

Die Kernidee der mathematischen Kategorientheorie ist es, beliebige Strukturen als Sammlungen von Objekten und ihren wechselseitigen Beziehungen zueinander zu charakterisieren. Im einfachsten Fall besteht eine Kategorie C aus einer Klasse von Objekten Ob_C und einer Menge von Beziehungen, oder Morphismen, Mor_C zwischen je zwei Objekten $A, B \in Ob_C$. Des Weiteren wird eine individuelle Beziehung als Morphismus $f \in Mor_C(A, B)$ bezeichnet und das Objekt A Quelle (oder *domain*) und B Ziel (oder *codomain*) genannt. Objekte und Morphismen lassen sich nun mithilfe von Pfeilen darstellen $f: A \rightarrow B$. Die Darstellung komplexer Strukturen wird durch Komposition von Morphismen erreichen, d.h. durch Anfügen weiterer Objekte und Pfeile $f: A \rightarrow B \rightarrow C$. Die so konstruierten Verknüpfungen lassen sich diagrammatisch als gerichtete Graphen darstellen.

Durch wechselseitiges Ersetzen – markiert durch einen Asterisk (*) – von Morphismen oder Objekten durch Strukturen können Abstraktionsebenen integriert oder Modellierungen mit mehr Details angereichert werden:

- Mehr Details: Ein Objekt oder ein Pfeil wird rekursiv durch weitere Pfeile und Objekte ersetzt:
 - Ersetzung des Objekts $B: A \rightarrow B \Rightarrow A \rightarrow [* \rightarrow *]$
 - Ersetzung des Pfeils zwischen A und $B: A \rightarrow B \Rightarrow A[* \rightarrow *]B$
- Mehr Abstraktion: Ganze Teilstrukturen durch Objekten oder Pfeile ersetzt werden:
 - Ersetzung durch Objekt: $A \rightarrow B \Rightarrow *$
 - Ersetzung durch Pfeil: $A \rightarrow B \rightarrow C \Rightarrow A \rightarrow *C$

Der ursprüngliche Zweck der mathematischen Kategorientheorie war der Vergleich mathematischer Theo-

rien (Mac Lane 2010), um so strukturelle Ähnlichkeiten zwischen unterschiedlichen mathematischen Disziplinen zu entdecken und sie zu vereinheitlichen. Um dies zu ermöglichen, werden Objekte rein formal, d.h. unter Absehung des konkreten Inhalts, betrachtet, was es uns erlaubt, auch nicht-mathematische Gegenstände zu modellieren. Hinzu kommt, dass dieser hohe Abstraktionsgrad die Abbildbarkeit verschiedener Theorien – oder in unserem Fall: Konzepte – aufeinander und damit auch ihren Vergleich ermöglicht.

Wir verwenden Elemente der angewandten Kategorientheorie als Gerüst, um klar und präzise darzustellen, worüber wir sprechen. Diese Darstellung ist nicht reduktionistisch in dem Sinne, dass wir behaupten, Literaturtheorie könne letztlich durch mathematische Strukturen ersetzt werden. Vielmehr zielen wir darauf ab, literarische Konzepte zu explizieren (vgl. Carnap 1945; Dutilh Novaes 2017) und zu operationalisieren. Mit Operationalisierung ist hier lediglich der Prozess oder Workflow gemeint, um Konflikte in literarischen Texten zu identifizieren und zu analysieren (vgl. Pichler & Reiter 2021). Diese Form der Operationalisieren soll nicht mit Operationalismus verwechselt werden, d. h. der streng positivistischen Vorstellung von Operationalisierung, die Bedeutung mit empirischen Operationen gleichsetzt.

Ein Framework für die Operationalisierung von Analysekonzepten

Ziel dieses Beitrags ist es, ein Framework zur Operationalisierung von Analysekonzepten am Beispiel von literarischen Konflikten zu skizzieren. Dieses Framework besteht im wesentlichen aus drei Formalisierungs- bzw. Abbildungsschritten. Hierzu entwickeln wir einen der mathematischen Kategorientheorie entlehnten Formalismus, der es erlaubt drei Ebenen zu integrieren: 1. Theorie, 2. Modell und 3. Text. Wir zeigen die Schritte im Folgenden am Beispiel einer Analyse von Konflikten in literarischen Texten.

Theorie

Die Aufgabe einer Theorie literarischer Konflikte ist es, festzulegen, unter welchen Bedingungen ein Konflikt vorliegt. Nun gibt es zahlreiche Möglichkeiten, Konfliktkonzepte in der literaturwissenschaftlichen Textanalyse zu nutzen. Für diesen Beitrag fokussieren wir uns auf eine Analyse von Konflikten zwischen Figuren und nutzen ein Konzept aus den Sozialwissenschaften, welches bereits in Gius (2015) im Rahmen einer narratologischen Analyse erprobt wurde. Im Sinne des *Principle of Minimal Departure* gehen wir davon aus, dass das Konzept des realen sozialen Konflikts auch in fiktionalen Welten eine gewisse Gültigkeit hat. Entsprechend halten wir uns an die Definition von Glasl (2011):

Sozialer Konflikt ist eine Interaktion
 - zwischen Akteuren (Individuen, Gruppen, Organisationen usw.),
 - wobei wenigstens ein Akteur

- eine Differenz bzw. Unvereinbarkeiten im Wahrnehmen und im Denken bzw. Vorstellen und im Fühlen und im Wollen
- mit dem anderen Akteur (den anderen Akteuren) in der Art erlebt,
- dass beim Verwirklichen dessen, was der Akteur denkt, fühlt oder will eine Beeinträchtigung
- durch einen anderen Akteur (die anderen Akteuren) erfolge. (Glasl 2011, 17)

Um diese informelle Definition zu operationalisieren und einen Konflikt K formell zu definieren, betrachten wir die darin verwendeten Entitäten und die Beziehungen zwischen ihnen. Glasl definiert die Klasse Akteur, die aus Individuen, Gruppen, Organisationen oder anderen Objekten bestehen kann, und legt fest, dass es mindestens zwei Akteure gibt. Wir definieren folglich die Objekte eines Konflikts Ob_K als die Menge der Akteure: $A = \{x | x = \text{Individuum} \vee x = \text{Gruppe} \vee x = \text{Organisation} \vee \# \}$ und $|A| \geq 2$.

Glasl charakterisiert zwei Beziehungen zwischen den Akteuren, erlebte Unvereinbarkeit u und Beeinträchtigung b , sodass $u, b \in Mor_K$. Zwischen zwei Akteuren $A_1, A_2 \in Ob_K$ gibt es dabei eine mindestens eine erlebte Unvereinbarkeit $u: A_1 \rightarrow A_2$, wobei:

$$u = \begin{cases} \text{Wahrnehmen} \\ \text{Denken bzw. Vorstellen} \\ \text{Fühlen} \\ \text{Wollen} \end{cases}$$

Zwischen den zwei Akteuren gibt es mindestens eine Beeinträchtigung des Verwirklichens $b: A_2 \rightarrow A_1$, wobei:

$$b = \begin{cases} \text{Denken} \\ \text{Fühlen} \\ \text{Wollen} \end{cases}$$

Des Weiteren soll sich die Unvereinbarkeit u auf die Beeinträchtigung b beziehen, sodass $u \Rightarrow b$.

Ein Konflikt K kann nun diagrammatisch wie in Abb. 1 dargestellt werden.

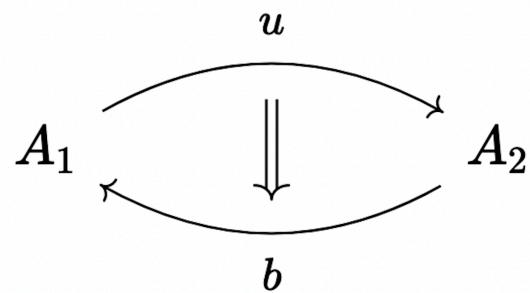


Abb. 1: Minimaler einseitiger Konflikt

An diesem Diagramm ist eine für die literaturwissenschaftliche Analyse schwerwiegende Einschränkung der Theorie Glasls zu sehen: Die Beobachtung des Konflikts ist in der Beziehung zwischen den Akteuren versteckt, d.h. Erzähl- bzw. Beobachtungsinstanzen kommen nicht explizit vor. Dies ist dem Umstand geschuldet, dass Glasl Konfliktbegriff aus den Sozialwissenschaften stammt und entsprechend, anders als in literarischen Texten, die Konfliktparteien zu ihrer Einschätzung befragt werden können. Da es das Ziel unserer Operationalisierung ist, den Konfliktbegriff so zu explizieren, dass auch komplexe Erzählsituationen erfasst werden können, erweitern wir Ob_K um Beobachtungsinstanz B , die in einer Wahrnehmungsbeziehung w zu A_1 und A_2 steht und die Geschehnisse erzählerisch vermittelt (s. Abb. 2):

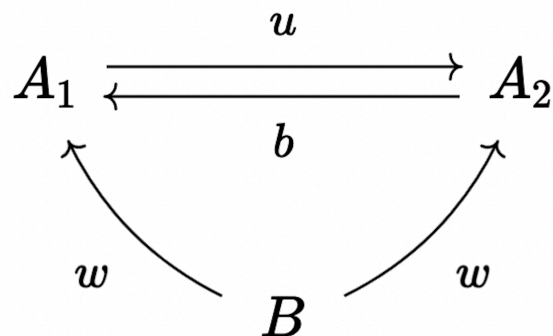


Abb. 2: Minimaler, um Beobachter erweiterter, Konflikt

Im paradigmatischen Fall des von A_1 wahrgenommenen Konflikts sind A_1 und B identisch. Die Erweiterung der Theorie um beliebig viele weitere Beobachter:innen (als erzählte Figuren, Erzählinstanzen) erlaubt es, auch komplexe Narrationen zu erfassen bzw. die Darstellung der Konflikthaftigkeit durch andere Erzählinstanzen als die Konfliktbeteiligten zu berücksichtigen.

Der erweiterte Konfliktbegriff umfasst nunmehr als Objekte Ob_K die Klasse der Akteure $A_1, A_2, \#, A_n$ und die

Klasse der Beobachter $B_1, \#, B_n$ sowie die Beziehungen $u: A \rightarrow A \in \text{Ob}_K$, $b: A \rightarrow A \in \text{Ob}_K$ und $w: B \rightarrow A \in \text{Ob}_K$. Zusammen mit der Möglichkeit, dass Akteure auch Vermittler sein können, ergeben sich vielfältige Modellierungsmöglichkeiten.

Modell

Ein Modell eines Konflikts liegt dann vor, wenn sich in literarischen Texten Kandidaten für Objekte und Morphismen finden, sodass es eine strukturerhaltende Abbildung auf den oben explizierten, erweiterten Konfliktbegriff K gibt. Eine naheliegende Modellierung ist die Abbildung von Elementen der fiktiven Welt, etwa durch das Einsetzen von Figurenkonstellationen ($F_1, \#, F_n$) als Akteure und von Erzählinstanzen (B_I) als Beobachtungsinstanzen. Eine Figurenkonstellation kann wieder als Tripel aus zwei Figuren und einer Relation konstruiert werden. So stehen in Kleists *Der Zweikampf* Herzog Wilhelm von Breysach und sein Bruder, Graf Jakob, in der Beziehung Erbfolgestreit zueinander. Dieser Erbfolgestreit weist alle Merkmale nach der Definition von Glasl auf und wird primär durch eine unbeteiligte Erzählinstanz wiedergegeben (es gibt auch Passagen, in denen die Konfliktbeteiligten selbst über den Konflikt erzählen, insgesamt überwiegt jedoch eindeutig die Darstellung durch die Erzählinstanz). Eine Analyse dieses Verhältnisses besteht in der inhaltlichen Zuordnung der Teile der Figurenkonstellation und der Vermittlungsinstanz zu den Objekten und Morphismen der Theorie:

$A_1 \rightarrow \text{Jakob}$

$A_2 \rightarrow \text{Breysach}$

$B \rightarrow \text{Unbeteiligte Erzählinstanz}$

$u \rightarrow \text{Unvereinbarer Herrschaftsanspruch}$

$b \rightarrow \text{Veranlassung der \#nderung der Erbfolgeregelung}$

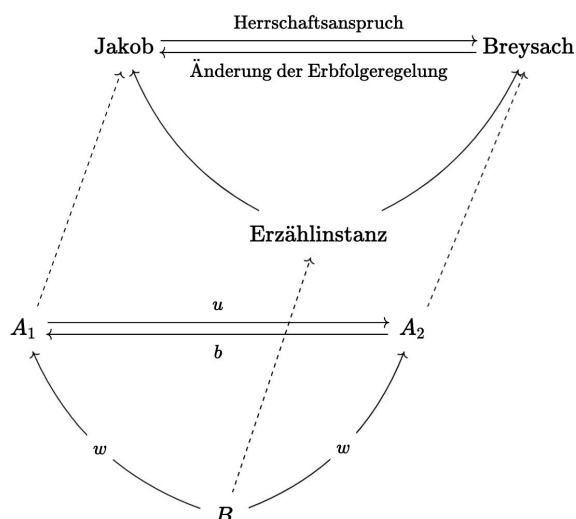


Abb. 3: Diagramm Erbfolgestreit in Kleists "Der Kampf"

Figurenkonstellationen und Beobachtungs- bzw. Erzählinstanzen können aufgrund ihrer Kombinierbarkeit

sehr flexibel modelliert werden; ebenso lassen sich weitere Konflikttheorien integrieren, die auf andere Aspekte abstellen, und weitere Objekte, wie gesellschaftlichen Wandel, oder andere Relationstypen zwischen den Objekten einführen. So könnte man Dahrendorfs (1972) Unterscheidung zwischen latenten und manifesten Konflikten durch zusätzliche Bedingungen integrieren, u.a. dadurch, dass beide oder keine der Konfliktparteien den Konflikt wahrnehmen. Oder man kann Definitionen psychischer Konflikte integrieren, indem man u.a. die Identität $A_1 = A_2$ annimmt. Man könnte auch das Modell auf extratextuelle Beobachtungsinstanzen erweitern und etwa die Einschätzung der Konflikthaftigkeit nicht anhand von deren Vermittlung, sondern anhand der Einschätzung von Leser:innen modellieren.

Text

Auf der Textebene gilt es jene Textphänomene zu bestimmen, welche auf die Figurenkonstellationen und Erzählinstanzen abgebildet werden können. Diese sollen schließlich in der Analyse manuell oder automatisch identifiziert werden. Es geht also um die Bestimmungen von am Konflikt beteiligten Figuren und ihn wahrnehmenden Instanzen und die weiteren, zu Unvereinbarkeiten bzw. wahrgenommener Einschränkung führenden Aspekte. Für jeden Aspekt muss in einer Analyse festgelegt werden, wie er im Text realisiert werden kann. Die Operationalisierung auf der Textebene hängt auch vom anvisierten Untersuchungsmodus ab und kann auf sehr unterschiedliche Weisen erfolgen. So kann es sinnvoll sein, die Figuren als per NER-Analyse erkennbare Personen-Entitäten zu fassen – oder diese entsprechend zu erweitern –, wenn man eine automatisierte Analyse anstrebt. In einer manuellen Analyse ist es vermutlich eher möglich, Figuren als komplexere Phänomene zu fassen und auch Charakterzüge u.ä. zu annotieren. Bei der wahrgenommenen Unvereinbarkeit könnte der Fokus auf repräsentierte Prozesse des Denkens und Fühlens etc. liegen, die zumindest teilweise automatisch erkannt werden können. Oder es könnte eine Sentimentanalyse in Abhängigkeit von den Erzählinstanzen und den beteiligten Figuren durchgeführt werden, bei welcher automatische und manuelle Verfahren kombiniert werden können. Eventuell bietet es sich auch an, Indikatoren für Konflikthaftigkeit herauszuarbeiten.

Unabhängig von den gewählten Phänomenen und ihrer Bestimmung im Text gilt: Durch Gruppierung von Textphänomenen in hierarchischen Tagsets und Kategoriensystemen können Texteigenschaften der Modellebene zugeordnet werden (vgl. Abb. 2, in der $T_1, \#, T_n$ die Textphänomene, $F_1, \#, F_n$ die Elemente der fiktiven Welt und B_I die Beobachtungs- oder Erzählinstanz bezeichnen):

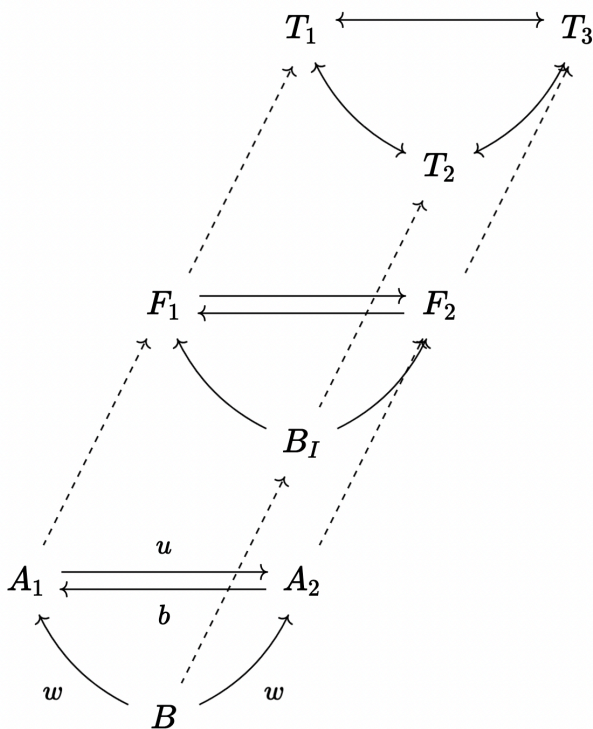


Abb. 4: Diagramm Zuordnung zwischen theoretischem Konzept, Modell und Textphänomenen

Integration der Ebenen

Das hier vorgeschlagene Framework erlaubt es, den Workflow für die Operationalisierung auf jeder der drei Ebenen zu beginnen und bei Bedarf auf eine der anderen Ebenen zu wechseln. Es gibt keine starre Beschränkung des Workflows auf top-down, middle-out oder bottom-up. Vielmehr ist es je nach Fragestellung möglich, in die Tiefe zu gehen und mehr Daten und Details einzuarbeiten oder zu abstrahieren, um größere Strukturen sichtbar zu machen. Das bedeutet, dass die genutzte Konfliktdefinition ebenso verändert werden kann wie die Umsetzung der Objekte und Relationen für literarische Texte (wie wäre es etwa, wenn man die Definition auf eine Analyse von poetologischen "Konflikten" zwischen Autor:innen ausweiten wollte?) oder deren Realisierung auf der Textebene. Dabei kann die Änderung auf einer der Ebenen jeweils Anpassungen auf den anderen Ebenen nötig machen.

Abschließende Bemerkungen

Mit dieser Formalisierung eines Konfliktbegriffs für die Literaturwissenschaft und seiner skizzierten Überführung in ein Modell, welches in der Textanalyse genutzt werden kann, haben wir gezeigt, wie man ausgehend von einer Theorie zu ihrer Anwendung auf Texte gelangen kann. Der Prozess ist – wie die angewandte Kategorientheorie selbst auch – ein allgemein anwendbarer Workflow für die Operationalisierung von Analysekonzepten für eine quantifizierende und/oder computationell-algo-

rithmische Analyse. Die angewandte Kategorientheorie ermöglicht dabei die in den Digital Humanities im Bereich von computationellen Analysen zwingend notwendige Formalisierung mit Fokus auf die Theorie. Insgesamt sehen wir vier substantielle Vorteile in diesem Workflow:

Erstens spielen in der ausschließlichen Beschäftigung mit der Theorie im ersten Schritt Fragen wie "Was ist überhaupt realistisch computationell umsetzbar?" vorerst keine Rolle. Man kann also vermeiden, dass die genutzte Theorie aus pragmatischen Gründen vorschnell unterkomplex formalisiert wird, indem aus pragmatischen Gründen auf bereits etablierte, relativ gute Analyseverfahren zurückgegriffen wird.

Zweitens erfolgt durch die angewandte Kategorientheorie und die in ihren Prinzipien festgelegte Kompositionalität immer eine feingranulare(re) und entsprechend genauere Formalisierung der zugrundeliegenden Theorie. In dieser sind die entsprechenden Unterkonzepte in Form von Objekten und Morphismen weniger komplex als das Gesamtkonzept. Die Bestimmung dieser einzelnen Elemente ist vergleichsweise einfacher, wobei das sowohl für manuelle als auch für maschinelle Zugänge gilt. Mit auf Kompositionalität beruhenden Operationalisierungen von Teilphänomenen kann man mehr und bessere Daten erzeugen, die zur Analyse genutzt werden, als wenn man versucht, die gesamte Theorie auf einmal anzuwenden. So ist es deutlich einfacher, jeweils zu bestimmen, welche Figur was fühlt und welche Handlungen Figuren vollziehen, und erst anschließend zu sehen, ob dort Konflikthaftigkeit erkennbar ist, als direkt Konflikte im Text zu annotieren.

Drittens ermöglicht die Kompositionalität es, verschiedene Theorien zusammen zu formalisieren. Durch die Identifikation von Objekten und Morphismen auf einer möglichst abstrakten Ebene wird leichter erkennbar, an welchen Stellen weitere Elemente ergänzt oder Elemente ersetzt werden können, um eine andere, ähnliche Theorie mitzuerfassen. In unserem Fall wurde dies etwa beim Wahrnehmungsmorphismus offensichtlich: Während dieser nach Glasl nur zwischen den beteiligten Akteuren besteht, ermöglichte die Einführung weiterer Objekte eine Verknüpfung mit einem literarischen Kommunikationsmodell, welches auch Erzählinstanzen enthält. Dies ist für die Analyse von Texten genauso hilfreich wie für ihre Auswertung, die nun ohne großen Mehraufwand für mehrere Theorien durchgeführt werden kann.

Viertens ist die Form, die eine nach der angewandten Kategorientheorie formalisierte Theorie hat, geeignet, bestehende Tripel-basierte Datenmodellierungen wie etwa RDF und darauf aufbauende Linked Open Data nachzunutzen.

Bibliographie

Carnap, Rudolf. 1945. „The Two Concepts of Probability: The Problem of Probability“. *Philosophy and Phenomenological Research* 5, Nr. 4 (Juni 1945): 513. <https://doi.org/10.2307/2102817>.

Dahrendorf, Ralf. 1972. *Konflikt und Freiheit: auf dem Weg zur Dienstklassengesellschaft*. Gesammelte Abhandlungen 2. München: R. Piper.

Dutilh Novaes, Catarina und Erich Reck. „Carnapian Explication, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization“. *Synthese* 194, Nr. 1 (Januar 2017): 195–215. <https://doi.org/10.1007/s11229-015-0816-z>.

Ehrig, Hartmut, Hrsg. . 2001. *Mathematisch-strukturelle Grundlagen der Informatik; 71 Tabellen*. 2. Aufl. Springer-Lehrbuch. Berlin Heidelberg: Springer.

Gius, Evelyn. 2015. *Erzählen über Konflikte: ein Beitrag zur digitalen Narratologie*. Narratologia, Band 46. Berlin; Boston: De Gruyter.

Glasl, Friedrich. 2011. *Konfliktmanagement: ein Handbuch für Führungskräfte, Beraterinnen und Berater*. Bern: Haupt.

Mac Lane, Saunders. 2010. *Categories for the Working Mathematician*. 2nd. ed., Softcover version of original hardcover edition 1998. Graduate Texts in Mathematics 5. New York, NY: Springer.

Moretti, Franco. 2013. „‘Operationalizing’: Or, the Function of Measurement in Literary Theory“. *New Left Review* 84 (Dezember): 103–19.

Pichler, Axel und Nils Reiter. 2021. „Zur Operationalisierung Literaturwissenschaftlicher Begriffe in der Algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers Hermeneutische Modellinterpretation von Kleists Das Erdbeben in Chili“. *Journal of Literary Theory* 15, Nr. 1–2 (31. Dezember 2021): 1–29. <https://doi.org/10.1515/jlt-2021-2008>.

Trilcke, Peer und Frank Fischer. 2016. „Fernlesen mit Foucault? Überlegungen zur Praxis des distant reading und zur Operationalisierung von Foucaults Diskursanalyse“. *Le foucaldien* 2 (1): 6. <https://doi.org/10.16995/le-fou.15>.

Einführung: Motivation

Text+ ist ein Konsortium der nationalen Forschungsdateninfrastruktur (NFDI). Seine Partner vereinigen die Expertise der bestehenden Infrastrukturverbünde CLARIN-D (Hinrichs und Trippel 2017) und DARIAH-DE (Neuroth u. a. 2016) und integrieren weitere Partner aus den Bereichen Datenzentren, Bibliotheken, Universitäten, Akademien und Rechenzentren. Durch diesen Zusammenschluss entsteht ein einzigartiger Schatz textueller Korpora, die Text+ durch sein Angebot nicht nur unter Beachtung der FAIR- (Wilkinson u. a. 2016) und CARE-Prinzipien (Carroll u. a. 2021) der Community zur Verfügung stellt, sondern Dienste und Beratung anbietet, um auf vielfältige Weise mit ihnen zu arbeiten. Dieser Beitrag wirft einen beispielhaften Blick auf vorhandene Korpora in Text+ (in Text+ vor allem innerhalb der Datendomäne Sammlungen), veranschaulicht die Ebenen des gemeinsamen Zugriffs und zeigt Möglichkeiten zur weiteren Integration von Ressourcen auf.

Text+ kann bei der Integration von Ressourcen auf vielfältige Vorarbeiten zurückgreifen. Die Vorarbeiten beziehen sich dabei etwa auf verschiedene Infrastrukturen, darunter CLARIN-D als nationalem Zweig der europäischen Infrastruktur CLARIN (Hinrichs und Krauwer 2014) und DARIAH-DE als nationalem Zweig der europäischen Infrastruktur DARIAH (Kálmán u. a. 2019; bzw. überblicksartig Gray 2021). Allerdings umfasst Text+ auch weitere Partner, die zuvor nicht an diesen geisteswissenschaftlichen Infrastrukturverbünden beteiligt waren, die aber unabhängig davon insbesondere große Inventare an Referenzdaten erstellt haben und weitere Erfahrungen in das Konsortium einbringen.

Korpora modular, verteilt, vernetzt in Text+

Leinen, Peter

P.Leinen@dnb.de
Deutsche Nationalbibliothek

Trippel, Thorsten

thorsten.trippel@uni-tuebingen.de
Leibniz-Institut für Deutsche Sprache

Weimer, Lukas

weimer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Witt, Andreas

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

Beispielhafte vorhandene Korpora

Auch wenn das NFDI-Konsortium Text+ erst am 1. Oktober 2020 formal begonnen hat, gibt es zahlreiche Vorarbeiten, die durch die Partner bereits zur Verfügung stehen und von vielen Forschenden (nach-)genutzt werden. Einige dieser Referenzdatensätze, die in Text+ zu den Sammlungen gehören, werden hier kurz vorgestellt.

DeReKo

Das deutsche Referenzkorpus (DeReKo, siehe Kupietz u. a. 2018; 2010; Lungen 2017; Kupietz und Keibel 2009) ist eine Sammlung elektronischer deutschsprachiger Korpora. Mit über 53 Milliarden Wörter ist diese linguistische Sammlung die weltweit größte ihrer Art. DeReKo enthält belletristische und wissenschaftliche Texte, Periodika und viele weitere Textarten der Gegenwart und früheren Vergangenheit. Über die Werkzeuge COSMAS II und KorAP ist es niedrigschwellig zugänglich und für diverse linguistische Fragestellungen nutzbar.

Zeitungskorpora der DNB

Die Deutsche Nationalbibliothek (DNB) sammelt gemäß ihres gesetzlichen Auftrags Zeitungen und Zeitschriften, die in Deutschland publiziert werden. Der umfangreiche Bestand beläuft sich aktuell auf ca. 311.000 gedruckte Titel und ca. 3,2 Millionen Ausgaben E-Paper. Insgesamt umfasst die digitale Sammlung der DNB aktuell ca. 12 Millionen Objekte.

Das Zeitungsportal der DDB.

Über das neu geschaffene Zeitungsportal der Deutschen Digitalen Bibliothek (DDB) ist eine große Sammlung historischer Zeitungen aus den Jahren 1671–1950 zugänglich.

ELTeC-Korpus

Das Korpus der European Literary Text Collection (ELTeC, Schöch u. a. 2021) im TextGrid Repository (TextGrid Repository 2020) ist eine Textsammlung von 2500 linguistisch annotierten Romanen in mindestens zehn Sprachen. Sie ermöglicht es, Methoden und Verfahren der Textanalyse über mehrere Nationalliteraturen hinweg zu vergleichen.

Baumbanken

Baumbanken haben in der Sprachwissenschaft eine lange Tradition. Im Gegensatz zu anderen Arten von Korpora und Sammlungen enthalten sie geparste Texte, die syntaktische Annotationen enthalten. Dadurch werden sie für die Forschung zur Grammatik, Typologie, Sprachtechnologie, etc. eingesetzt (siehe z. B. Kübler, McDonald und Nivre 2009; Hinrichs, Filippova und Wunsch 2005). Ein Beispiel für eine Baumbank ist die Tübinger Baumbank des Deutschen/Zeitungskorpus (TüBa-D/Z, siehe auch Telljohann, Hinrichs und Kübler 2004), die auf Artikeln der Zeitung 'die tageszeitung' (taz) basiert und in der letzten Ausgabe (Version 11) auf 3816 Artikeln und über 100.000 Sätzen mit fast 2 Millionen Token beruht (siehe <https://uni-tuebingen.de/de/134290>). Mit spezialisierten Suchprogrammen wie Tundra (Martens 2013) kann nach Wörtern und syntaktischen Strukturen gesucht werden. Tundra ist dabei nicht auf das Deutsche beschränkt. Die Spannweite reicht dabei von Werken von Thomas von Aquin (Martens und Passarotti 2014) bis zu den vielen Baumbanken der Universal Dependency Treebanks Initiative (<https://universaldependencies.org/>), die für viele Sprachen Baumbanken zur Verfügung stellt.

Beispiele weiterer Korpora in Text+

Daneben gibt es in Text+ noch zahlreiche weitere wichtige Korpora. Dazu gehört beispielsweise das Deutsche Textarchiv (DTA) an der Berlin-Brandenburgischen Akademie der Wissenschaften mit Texten ab ca. 1600 bis 1900. Gesprochensprachliche Korpora sind z. B. am Bayerischen Archiv für Sprachsignale an der LMU München, mit der Datenbank Gesprochenes Deutsch (DGD)

am IDS Mannheim oder auch in der Digitalen Bibliothek an der SUB Göttingen vorhanden. Diese Aufzählung ist bei weitem nicht vollständig, zeigt aber auf, dass die obige, beispielhafte Auswahl von Ressourcen nur einen kleinen Teil des Portfolios abdeckt und daher Lösungen innerhalb von Text+ zwar mit Hilfe einiger Beispiele implementiert werden, aber die Lösungen immer die Erweiterungsmöglichkeit auch auf weitere Daten und Partner erlauben muss. Alle Ressourcen, Einrichtungen und Angebote können als Module von Text+ gesehen werden, die durch die Infrastruktur vernetzt werden, so dass auf sie gemeinsam zugegriffen werden kann.

Ebenen des gemeinsamen Zugriffs

Ressourcen in Text+ liegen an unterschiedlichen Institutionen verteilt vor, teilweise sind sie abgeschlossen, teilweise werden sie noch aktiv weiterentwickelt. Da viele bestehende und im Aufbau befindliche Ressourcen Rechte Dritter berühren, z. B. über Verlagsrechte an Texten, bedeutet ein FAIRer Zugang nicht, dass der Zugang offen und frei sein kann. Vielmehr stehen viele Ressourcen unter expliziten Lizenzen, die die Nutzungsart einschränken. So kann beispielsweise auf viele digitale textuelle Daten der DNB nur innerhalb der Bibliothek und über deren Infrastruktur zugegriffen werden, um rechtliche Bestimmungen zu wahren. Für andere Sammlungen gibt es Zugangsbeschränkungen, die über ein Login den Zugang auf autorisierte Nutzende einschränkt, um Lizenzbestimmungen und -verträge einhalten zu können. Eine Grundlage für die konsortiumsweite und communityintegrierende Arbeit an und mit den Korpora ist daher ein gemeinsamer, niederschwelliger Zugriff. In Text+ wie in verschiedenen Vorarbeiten wird dazu eine gemeinsame AAI-Lösung (Authentication and Authorization Infrastructure) verwendet, über die Nutzende die verschiedenen Ressourcen und Werkzeuge verwenden können.

Da der Zugriff auf die Daten oft eingeschränkt ist, gibt es häufig keinen Vollzugriff auf die (digitalen) Daten. Über ein föderiertes Anmeldeverfahren, das auf Shibboleth und den Diensten des Deutschen Forschungsnetzes (DFN) und deren europäischer Zusammenarbeit im Rahmen von GÉANT aufsetzt, sind viele Dienste institutionsunabhängig zugänglich, z. B. Dienste, die über die AcademicCloud unseres Partners GWDG angebunden sind. Dienste wie die Federated Content Search oder die DARIAH Collection Registry bieten bereits jetzt übergreifende Suchmöglichkeiten nicht nur in Korpora. Daten werden über verschiedene Archive bereitgestellt, darunter das DARIAH-DE Repository an der SUB Göttingen sowie die zahlreichen zertifizierten CLARIN-Zentren.

In vielen Fällen besteht ein freier Zugriff auf Informationen in Katalognachweissystemen. Diese enthalten in der Regel zumindest Informationen über die Existenz der Korpora (bzw. anderer Forschungsdaten) und andere grundlegende, beschreibende Metadaten. Im Bereich der Metadaten besteht eine Herausforderung in den unterschiedlichen verwendeten Metadatenformaten, die nicht zuletzt bezüglich ihrer integrierten semantischen Interoperabilität divergieren. Hier reicht die Spannweite von Standardformaten (z. B. dem TEI-Header,

Dublin Core, Marc21, ISO 24622-X CMDI) bis zu relativ frei definierten Beschreibungen von Ressourcen. Mit mehreren Hunderttausend Datensätzen der Text+ Partner insbesondere im Bereich der Sammlungen stellt die Vernetzung der Daten eine erhebliche Herausforderung dar. In Arbeitsgruppen von Text+, die über die verschiedenen Daten- und Aufgabendomänen hinweg gebildet wurden, werden für den Zugang Lösungen entwickelt. Die Arbeitsgruppe zur Text+ Registry arbeitet dabei an den einzusetzenden Verfahren und Schnittstellen, um ein Inventar der verschiedenartigen Ressourcen anbieten zu können. Dabei zeichnet sich ab, dass für eine solche große Datenmenge eine Listendarstellung nicht ausreichend ist und zumindest grundlegende gemeinsame Informationseinheiten in den verschiedenen Metadaten enthalten sein müssen, um eine sinnvolle Überblicksdarstellung des Inventars zu ermöglichen. Die unterschiedlichen Anforderungen an Metadaten durch Bestandsdaten, Datentypen und neuere Entwicklungen werden dabei gewahrt werden. Eine weitere Arbeitsgruppe beschäftigt sich mit Fragen des Linked Data, wobei hier zwischen Linked Data für beschreibende Metadaten und Objektdaten unterschieden wird. Durch die Verknüpfung von Metadaten mit z. B. Normdaten und dem Angebot im Rahmen von Linked Data-Initiativen wird untersucht, welche Mehrwerte für Forschungsfragen der Text+ Community geschaffen werden können. Die Darstellung von Objektdaten wie Korpora in Linked Data-Formaten sowie die Schaffung rechtskonformer Angebotsmöglichkeiten sind hier noch am Beginn der Entwicklungen.

Ein zusätzlicher, gemeinsamer Zugang zu den Daten über die Metadaten hinaus wird in Text+ mit sogenannten abgeleiteten Textformaten (Schöch u. a. 2020) weiterentwickelt. Abgeleitete Textformate sind dabei Informationen über die Texte, die nicht die Lizenzbedingungen verletzen, aber trotzdem jene Informationen enthalten, die für konkrete Forschungsfragen erforderlich sind. Ein Ziel besteht dabei darin, z. B. die bei Partnern von Text+ befindlichen vollständigen Sammlungen rechtskonform nutzbar zu machen, etwa die Zeitungen, die an der DNB gesammelt werden oder Korpora des IDS, die nicht im Volltext außerhalb der Lizenzbestimmungen bereitgestellt werden dürfen. Eine zusätzliche Zugangsmöglichkeit zu den zugrundeliegenden Texten unter Berücksichtigung der Rechte Dritter entwickelt Text+ auf Grundlage einer föderierten Suche über den Inhalt von Ressourcen, die Federated Content Search (FCS, Vorarbeiten siehe Stehouwer u. a. 2012). Eine Herausforderung bei der Inhaltssuche sind die unterschiedlichen Annotationssebenen, welche die teils aufwändig annotierten Korpora enthalten, seien es linguistische aber auch andere Annotationen, die z. B. in anderen TEI-Repräsentationen von Texten enthalten sind. Hier muss – ebenso wie für andere Suchen – eine gemeinsame Anfragesprache definiert werden. Da diese gemeinsame Anfragesprache zwangsläufig nur eine Teilmenge der Möglichkeiten spezialisierter Werkzeuge bieten kann, können Anfragen, die z. B. durch hochspezialisierte Forschungsfragen motiviert sind, häufig besser mit den spezialisierten Werkzeugen bearbeitet werden. Ein konkretes Beispiel sind Anfragen zu bestimmten syntaktischen Strukturen, die in Baumbanken enthalten sind, die aber auf nicht syntaktisch annotierten Daten nicht zu einem Ergebnis führen

können und daher in einer föderierten Inhaltssuche nur eingeschränkt zur Verfügung stehen.

Integration weiterer Ressourcen

Ein Anspruch, den eine nationale Forschungsdateninfrastruktur adressieren muss, ist der, dass eine Vollständigkeit kaum gewährleistet werden kann und daher eine Offenheit gegenüber neuen Daten und Datengegebenen bestehen muss. Insbesondere vor dem Hintergrund zunehmender Anforderungen an das Forschungsdatenmanagement im Rahmen des Forschungsprozesses – etwa der Anforderung, Daten, die als Grundlage von Forschungsergebnissen dienen, für mindestens 10 Jahre aufzubewahren (siehe DFG-Leitlinien zur Sicherung guter wissenschaftlicher Praxis¹) – können Forschende darauf angewiesen sein, verlässliche Partner zu finden, die sie beim Datenmanagement unterstützen. Text+ adressiert diesen Bedarf auf unterschiedliche Weise: (1) durch Kooperationsprojekte zur Förderung neuer Angebote zur Integration in die Text+ Infrastruktur; (2) durch offene Schnittstellen, durch die Datenzentren ihre Angebote auch über die Infrastruktur von Text+ bereitstellen können und (3) durch Angebote von Partnern von Text+, Daten zu hosten.

Kooperationsprojekte sind Projekte, die im Rahmen von Text+ das Portfolio an Daten und Diensten für die wissenschaftliche Gemeinschaft erweitern. Jedes Jahr erfolgt in Text+ eine Ausschreibung, durch die zusätzliche Arbeiten gefördert werden können, die eine Integration im Rahmen der NFDI-Angebote von Text+ ermöglichen. Ausgestattet mit substantiellen Mitteln können dadurch Bestandsdaten und -dienste die Infrastruktur erweitern.²

Neben den Kooperationsprojekten basiert die Offenheit von Text+ auch auf den Schnittstellen. Als ortsverteilte Infrastruktur mit einer verteilten Datenhaltung, unterschiedlichen Schwerpunkten der Beteiligten, diversen Dateiformaten und zugrundeliegenden Technologien der Archivsysteme bei den verschiedenen Partnern sind Schnittstellen zur Zusammenarbeit der Partner in gemeinsamen Angeboten unverzichtbar. Ein Beispiel dafür sind die bereits beschriebenen unterschiedlichen Korpora, die durch die föderierte Inhaltssuche über eine gemeinsame Schnittstelle zugänglich sind. Diese Schnittstellen werden in Text+ offen spezifiziert und sind damit auch für Datenzentren außerhalb des Konsortiums nutzbar. Dadurch können auch Datenzentren außerhalb von Text+ ihre Angebote über die Infrastruktur verfügbar machen und können, so sie ihre Angebote nachhaltig bereitstellen möchten, verlässlich integriert werden. Über diesen Mechanismus können sich z. B. auch Anbietende von bisher nicht einbezogenen Communities, etwas aus dem Bereich der sog. „kleinen Fächer“, über die Infrastruktur vernetzen.

Daneben übernehmen die unterschiedlichen Daten- und Kompetenzzentren von Text+ aber auch selbst Daten Dritter. Wenn Daten und Dienste zur Spezialisierung und zu den Möglichkeiten eines Partners passen, können sie dort gehostet werden. Die Prozesse – angefangen von Dateiformaten bis zu den dafür notwendigen Lizen-

zen und Übereinkommen – basieren auf den jeweiligen lokalen Kontexten der Partner. Als Fallback steht aber auch eine Bitstream-Archivierung von einem Partner in Text+ zur Verfügung, hier werden Daten archiviert, aber nicht notwendigerweise in weitere Infrastrukturdienste integriert, z. B. weil sie nicht die formalen Voraussetzungen erfüllen oder Daten nicht zu den bestehenden Daten- und Kompetenzzentren passen. Auch dies erlaubt es, die Community zu erweitern.

Die Erweiterungsmöglichkeiten von Text+, gerade im Bereich der Sammlungen, sind dabei sehr vielfältig und erlauben individuelle Absprachen mit Forschenden. Dabei ist es entscheidend, dass die Forschenden durch die Infrastruktur ihre Interessen wahren und die Verantwortung für und Rechte an Ressourcen nicht verlieren.

Ausblick auf weitere Entwicklungen in Text+

Das Bestreben der Vernetzung der an Text+ beteiligten Institutionen, ihrer Daten und Dienste wird auch in weiterer Hinsicht in diversen Arbeitsgruppen von Text+ vorangetrieben. Ein Bestandteil für diese Vernetzung und Integration ist eine übergreifende Registry, über die ein Zugang zu Ressourcen aus allen Text+-Datendomänen trotz deren Unterschieden, etwa ihrer disziplinären Genese, Institution, Sprache, etc. möglich werden soll. Diese Registry, aber auch soweit möglich die Daten an sich, sollen über eine Erweiterung der Federated Content Search durchsuchbar gemacht werden. Schließlich gibt es zur Verbesserung der Vernetzung der Daten eine Arbeitsgruppe zu Linked Data (LOD) in Text+. Um den Zugriff auf all diese Dienste zu gewährleisten und zu verbessern, wird die Text+-Homepage (<https://www.text-plus.org>) sukzessive zu einem Text+-Portal als universaler Einstiegspunkt in die Text+-Infrastruktur transformiert.

Fußnoten

1. Siehe Leitlinie 17 in DFG 2019: Leitlinien zur Sicherung guter wissenschaftlicher Praxis, https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf (korrigierte Version 1.1 vom April 2022, zuletzt aufgerufen am 2022-12-13)
2. Zum Zeitpunkt des Vortrags können zu den Kooperationsprojekten aus dem Bereich der Sammlungen konkrete Beispiele benannt werden, die derzeit noch in der Bewilligung und Kooperationsvertragsabwicklung sind.

Bibliographie

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell und Shelley Stall. 2021. „Operationalizing the CARE and FAIR Principles for Indigenous data futures“. *Scientific Data* 8 (1): 108. <https://doi.org/10.1038/s41597-021-00892-0>.

Gray, Edward J. 2021. „DARIAH ERIC: Empowering Arts and Humanities Research on a National and International Level“. Zenodo. <https://doi.org/10.5281/zenodo.5596905>.

Hinrichs, Erhard, Katja Filippova, und Holger Wunsch. 2005. „What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German“. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT 2005)*, 77–88.

Hinrichs, Erhard und Steven Krauer. 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'14)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk und Stelios Piperidis, 1525–31. Reykjavik, Island: ELRA.

Hinrichs, Erhard und Thorsten Trippel. 2017. „CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften“. *Bibliothek - Forschung und Praxis* 1 (41).

Kálmán, Tibor, Matej Ďurčo, Frank Fischer, Nicolas Larrousse, Claudio Leone, Karlheinz Mörth und Carsten Thiel. 2019. „A landscape of data – working with digital resources within and beyond DARIAH“. *International Journal of Digital Humanities* 1 (1): 113–31. <https://doi.org/10.1007/s42803-019-00008-6>.

Kübler, Sandra, Ryan McDonald und Joakim Nivre. 2009. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Springer. <https://doi.org/10.1007/978-3-031-02131-2>.

Kupietz, Marc, Cyril Belica, Holger Keibel und Andreas Witt. 2010. „The German Reference Corpus DeReKo: A primordial sample for linguistic research“. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner und Daniel Tapias, 1848–54. European Language Resources Association (ELRA) 2010.

Kupietz, Marc und Holger Keibel. 2009. „The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research“. In *Workings Papers in Corpus-based Linguistics and Language Education*, herausgegeben von Makoto/Kawaguchi Minegishi, 3:53–59. Tokyo: Tokyo University of Foreign Studies 2009.

Kupietz, Marc, Harald Lungen, Paweł Kamocki und Andreas Witt. 2018. „The German Reference Corpus DeReKo: New Developments – New Opportunities“. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, u. a., 4353–60. Miyazaki: European Language Resources Association (ELRA).

Lungen, Harald. 2017. „DeReKo - Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim“. *Zeitschrift für germanistische Linguistik* 45 (1): 161–70.

Martens, Scott. 2013. „TüNDRA: A Web Application for Treebank Search and Visualization“. In *Proceedings of the 12th International Workshop on Treebanks and Linguistic Theories (TLT 2013)*. Sofia, Bulgaria.

Martens, Scott und Marco Passarotti. 2014. „Thomas Aquinas in the TüNDRA: Integrating the Index Thomisticus Treebank into CLARIN-D“. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk und Stelios Piperidis. Reykjavík, Island.

Neuroth, Heike, Stefan Schmunk, Mirjam Blümm, Andrea Rapp, Fotis Jannidis, Dirk Wintergrün, Ulrich Schwardmann und Peter Gietz. 2016. *DARIAH-DE – Digitalität in den Geistes- und Kulturwissenschaften am Beispiel der digitalen Forschungsinfrastruktur DARIAH-DE*. Bd. 40. Bibliothek Forschung und Praxis 2. De Gruyter.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen“. *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2020_006.

Schöch, Christof, Roxana Patras, Tomaž Erjavec, und Diana Santos. 2021. „Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives“. *Modern Languages Open* 1 (25): 1–19. <https://doi.org/doi.org/10.3828/mlo.v0i0.364>.

Stehouwer, Herman, Matej Ďurčo, Eric Auer und Daan Broeder. 2012. „Federated Search: Towards a Common Search Infrastructure“. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk und Stelios Piperidis. Istanbul, Türkei.

Telljohann, Heike, Erhard Hinrichs und Sandra Kübler. 2004. „The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone“. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2004/pdf/135.pdf>.

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“. *Scientific Data* 3 (März): 160018. <https://doi.org/10.1038/sdata.2016.18>.

Korpuszusammensetzung und Verlässlichkeit des deutschsprachigen Google Ngram-Viewers

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Als Google im Jahre 2009 die erste Version des Ngram-Viewer publizierte, hat die Digital Humanities-Community recht schnell die positiven und negativen Aspekte dieses Werkzeugs analysiert: Die Möglichkeiten einer wort- und begriffsgeschichtlichen Forschung sind sprunghaft erweitert worden, wie auch der begleitende Aufsatz von Michel deutlich belegte (Michel et al. 2011). Dessen teilweise zu naive Umgang mit dem Quellenmaterial machte aber auch deutlich, dass die Autorinnen und Autoren kein Bewusstsein für die möglichen Fallen von korpusbasierten Forschungen hatten. Die wechselnde Zusammensetzung der zugrundeliegenden Textsammlung, die Unmöglichkeit auf die dahinterliegenden Texte zuzugreifen, das Fehlen von Metadaten für die Texte, falsche Jahreszahlen, OCR-Fehler u.a.m. sind dem Ngram-Korpus wiederholt vorgeworfen worden (z.B. Underwood 2012). Die Arbeit von (Pechenick et al. 2015) hat gezeigt, wie die zunehmende Menge von wissenschaftlicher Literatur die Korpusanteile in der zweiten Hälfte des 20. Jahrhundert merklich verschiebt; unklar bleibt allerdings, ob dies nicht auch eine gesellschaftliche Entwicklung reflektiert, also keineswegs nur als Manko zu betrachten ist. Besonders einschlägig ist die Arbeit (Koplenig 2017), in der Veränderungen der Korpuszusammensetzung während des zweiten Weltkriegs untersucht werden: „the German GB corpus was strongly biased toward volumes published in Switzerland during WWII“ (Koplenig 2017).

Google hat zwei größere Updates vorgelegt, die die zugrundeliegende Textmenge deutlich erweitert haben. Hier die Entwicklung des Umfangs der deutschsprachigen Korpora, auf die ich mich im Folgenden beschränke:

	Tokens	Bücher
2009	37.439.210.527	406.666
2012	64.784.628.286	657.991
2019	286.463.423.242	3.843.962

Dadurch dass die Anzahl der digitalisierten Bücher noch einmal deutlich gesteigert werden konnte – und das über den gesamten Zeitraum, den der Viewer abdeckt, – scheinen die Fragen nach einem Bias der Auswahl weniger relevant zu werden. Entsprechend wurde und wird der Ngram-Viewer auch weiterhin in vielen Kontexten als

schnell zugängliches Werkzeug verwendet, das die Möglichkeiten einer auf sehr großen Datenbeständen basierenden Forschung anschaulich macht (z.B. Chen and Yan 2016, Gonçalves et al. 2018, Richey and Taylor 2020). Insgesamt gehören der *Ngram-Viewer* und die zugrundeliegenden freiverfügbaren Daten trotz aller berechtigten Kritik für viele Forschende zu den wichtigsten Daten-Publikationen der letzten Jahrzehnte.

In diesem Sinne wollte auch eine Kollegin, die zu Fragen der literarischen Kanonisierung arbeitet, den Viewer verwenden, aber bei der Analyse der Ergebnisse fiel uns schnell auf, dass die Namen einer Reihe der hochkanonischen deutschsprachigen Autoren – Thomas Mann, Goethe, Schiller, Kafka, Brecht – nach 2005 einen auffälligen Abwärtstrend aufweisen (siehe Fig. 1).

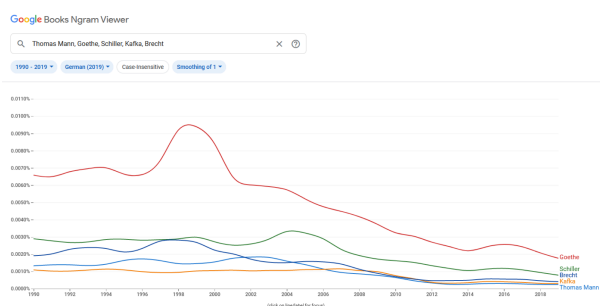


Fig. 1: Kanonisierte Autoren (Goethe, Schiller, Thomas Mann, Kafka, Brecht) 1990-2019.

Das Phänomen zeigt sich noch deutlicher, wenn man sich nur für kanonisierte Autor *innen* interessiert (siehe Fig. 2).

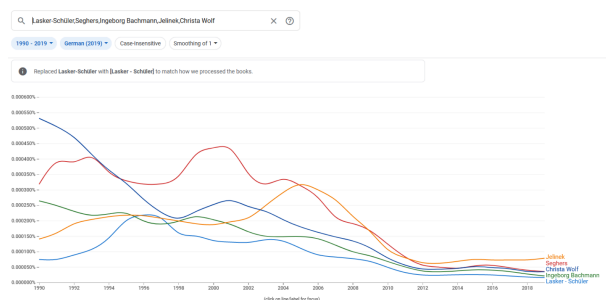


Fig. 2: Häufigkeit kanonisierter Autorinnen im Ngram-Viewer (Lasker-Schüler, Seghers, Bachmann, Jelinek, Christa Wolf) 1990-2019.

Bevor wir nun allgemeinere kulturanalytische Thesen über das Ende der Bildungskultur formulierten, wollte ich die Solidität der Daten prüfen. Doch wie kann man ein Korpus auf möglichen Bias untersuchen, wenn weder die Liste der Texte geschweige die Texte selbst vorliegen, sondern nur eine Reihe von generischen Metadaten und 1-5 Gramme?

Die Korpusanomalien

Da die NGramme, die dem *Ngram-Viewer* zugrundeliegen, ebenfalls publiziert sind, ist es naheliegend, erst ein-

mal die Rohwerte und deren Entwicklungstendenzen zu untersuchen. Das Ergebnis (Fig. 3) zeigt zumindest für drei der untersuchten Namen eine einheitliche Tendenz: aufsteigend – also genau das Gegenteil der absteigenden Entwicklung, die der *Ngram-Viewer* präsentiert. (Aufällig ist außerdem ein Abfallen in allen Verteilungen im Jahre 2010.)

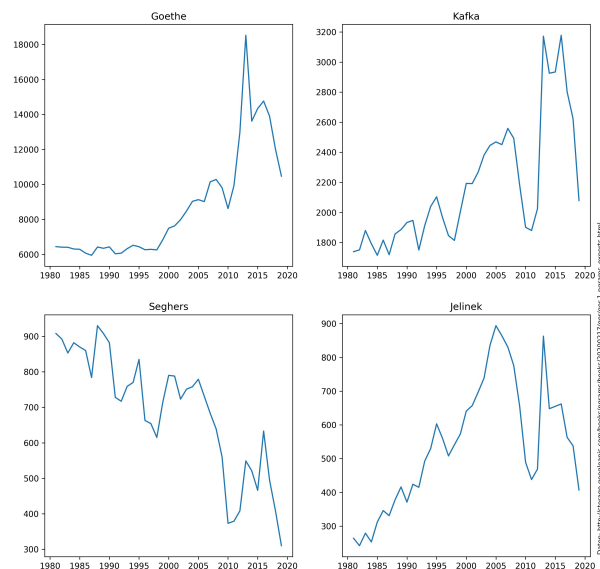


Fig. 3: Rohwerte für die Häufigkeiten im Ngram-Viewer Korpus (1-Gramme, 2019)

Wie lässt sich der Widerspruch zwischen den Ergebnissen erklären? Der *Ngram-Viewer* zeigt die *relativen* Häufigkeiten der Wörter an, d.h. den Anteil den das Wort, z.B. der Name ‚Goethe‘, an der Menge aller publizierten Wörter dieses Jahres hat. Dadurch kann man Frequenzen aus dem 18. Jahrhundert, die nur auf einigen wenigen Büchern beruhen, mit denen im 21. Jahrhundert vergleichen. Die Differenz der Ergebnisse könnte also so erklärt werden, dass zwar die Anzahl an Nennungen von Goethe weiter ansteigt, aber ab ca. 2005-2010 zugleich sehr viel mehr Bücher im Korpus sind, in denen die Namen der Autoren und Autorinnen nicht genannt werden. Das würde bedeuten, dass insgesamt im Korpus die Anzahl der Bücher pro Jahr deutlich angestiegen sein müsste. Die entsprechenden Daten sind im Ngram-Korpus vorhanden und sie bestätigen die These:

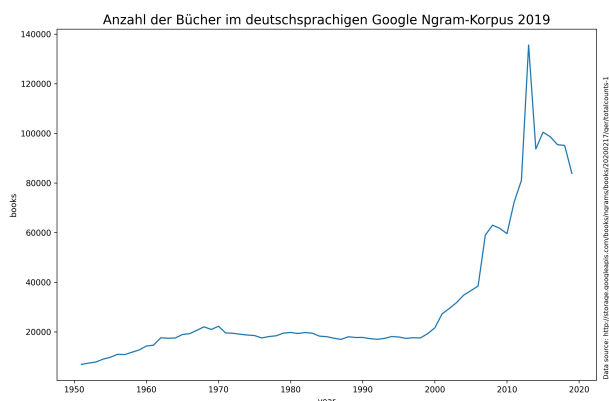


Fig. 4: Anzahl der Bücher pro Jahr im DE-Ngramm-Korpus 2019

Wenn die Rohdaten aus Fig. 3 nun mit den Daten über die Anzahl der Token pro Jahr normalisiert werden, dann ergibt sich der Trend, der im Ngram-Viewer sichtbar wurde, alle Werte stürzen nach 2005 mehr oder weniger steil ab.

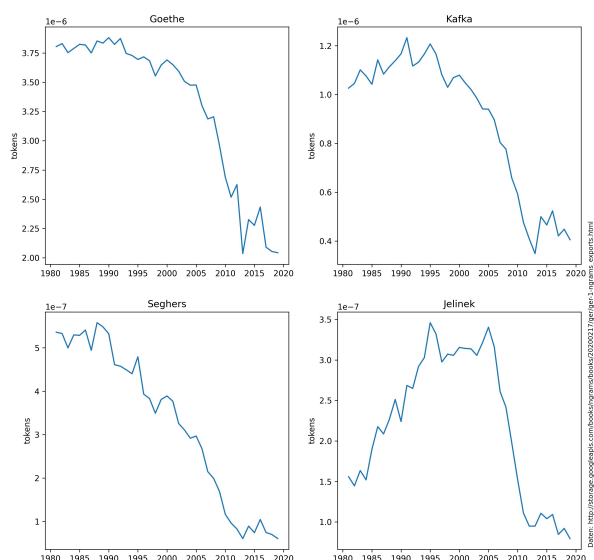


Fig. 5: Häufigkeitswerte der kanonisierten Autoren und Autorinnen geteilt durch die Anzahl der Token pro Jahr

Allerdings ist der sehr steile Anstieg der Buchzahlen in Fig. 4 ziemlich überraschend. Wir wissen, dass Google Books auf den Ergebnissen der Digitalisierungskampagne Googles in Kooperation mit einer ganzen Reihe von internationalen Forschungsbibliotheken beruht. Die deutschsprachigen Ergebnisse verdanken sich nicht zuletzt den Kooperationen mit der Österreichischen Nationalbibliothek und der Bayerischen Staatsbibliothek. Deren Bestände speisen sich in den letzten Jahrzehnten aus Pflichtabgabeexemplaren und einer umfassenden Erwerbspolitik. Woher kommt also der plötzliche Anstieg? Sind – den Klagen der Verlage zum Trotz – seit 2005 sehr viel mehr Bücher als früher gedruckt worden?

Für die Buchproduktion konnte ich zwei Quellen verwenden, die die Daten gleich in digitaler Form anbie-

ten: Statista, ein kommerzieller Datenanbieter, der Zahlen des Börsenvereins des deutschen Buchhandels zur Anzahl der Neuerscheinungen aufbereitet hat (Börsenverein 2022),¹ und die Angaben über die Anzahl der Buchpublikationen insgesamt in Thomas Rahlfs Zeitreihen zur historischen Statistik (Rahlf 2015). Sie decken allerdings nicht die gleichen Zeiträume ab. Statista hat die aktuelleren Daten, reicht aber nicht soweit zurück, während Rahlfs Datenreihe schon 2015 endet. Schon ein erster Blick auf die Fig. 6, die die Anzahl Bücher, die laut Börsenverein/Statista zwischen 2002 und 2019 gedruckt worden sind, mit der Anzahl der Bücher vergleicht, die dem Ngramm-Korpus zugrundeliegen, zeigt etwas Verblüffendes: In dem Korpus sind nach 2012 mehr Bücher enthalten, als Neuerscheinungen in dem Jahr gedruckt wurden.

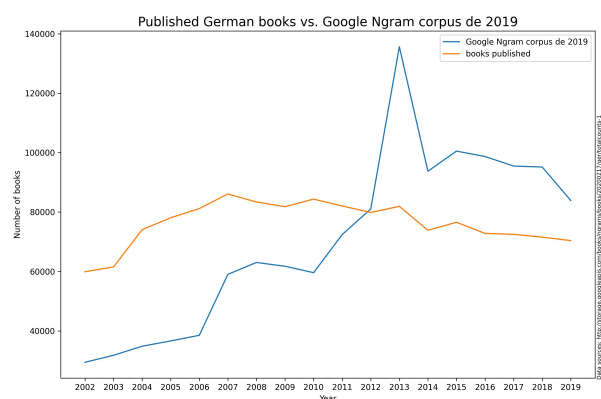


Fig. 6: Bücher im deutschsprachigen Ngramm-Korpus und die Anzahl der jährlich publizierten Bücher (Statista)

Vielleicht stimmen ja die Zahlen von Statista nicht. Vielleicht sind die Angaben in dem Ngram-Korpus aus dem Jahre 2019 falsch. In der Fig. 7 sind nun alle Informationen gemeinsam eingeflossen. Sie zeigt zum einen die Werte der Buchproduktion aufgrund der Daten des Börsenvereins – das ist der kurze violette Graph – und nach den Werten von Rahlf – die rote Kurve ab 1950. Es ist offensichtlich, dass die Werte zur Buchproduktion sich voneinander unterscheiden: Rahlf liegt immer etwas höher, da er ja nicht nur die Neuerscheinungen erfasst, aber beide beschreiben ziemlich parallel die gleiche Dynamik. Bis ca. 2009 steigt die Produktion immer weiter an und fällt danach ab. Aber welche der beiden Kurven man auch zugrundelegt, stets übersteigen die Zahlen des Ngramm-Korpus die Werte im Jahr 2013 und wohl auch danach.

Die Grafik enthält neben den Werten für das aktuelle Korpus von 2019 auch die Werte für die früheren Ngramm-Korpora von Google aus den Jahren 2009 und 2012, die deutlich kleiner waren. Beginnen wir bei den Werten vor 1995: Es ist auffällig, dass Google mit dem letzten Update den Anteil an der Buchproduktion eines Jahres, der digitalisiert vorliegt, deutlich steigern konnte, so dass bis in die Mitte der 1960er Jahre teils 50% und mehr im Korpus enthalten sind. Für den Zeitraum von den späten 1960ern bis in die späten 1990er stagnieren die Werte der Korpora, während die Buchproduktion in diesen Jahren steil angestiegen ist. Während im Jahr 1967 erstaunliche 67% der nach Rahlfs publizierten Bücher im

Korpus liegen, sind es 1997 ‚nur‘ noch 23%. Erst danach, 1998 bis 2006, wächst der Anteil wieder. In den Korpora von 2009 und 2012 geschieht dies noch relativ langsam, während die Erweiterung von 2019 hier deutlich stärker zulegt. Von 2006 auf 2007 springen die Werte allerdings steil nach oben. Das gilt für alle drei Stufen des Korpus, aber auch hier ist der Anstieg im 2019-Korpus noch einmal deutlich ausgeprägter. Danach, zwischen 2008 und 2010 verhalten sich die Werte im Korpus auf einem hohen Niveau parallel zu den Werten der Buchproduktion und fallen mit dieser sogar leicht ab. Das ändert sich wiederum 2011-2013, wo wir einen weiteren Anstieg beobachten können, der diesmal sogar das Niveau der Buchproduktion übertrifft.

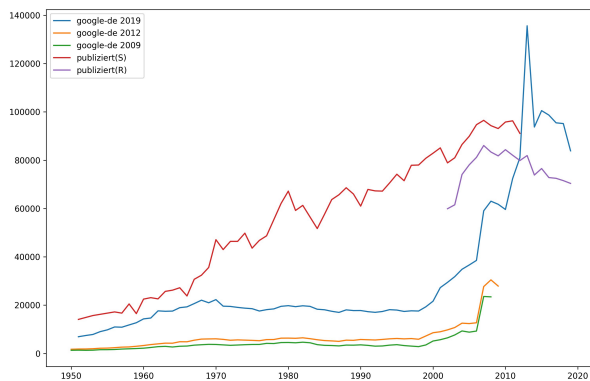


Fig. 7: Entwicklung der Buchproduktion und der Google Korpora 1950-2019

Wenn wir noch einmal auf Grafik 3 und 4 blicken, dann sehen wir dort, dass die Rohwerte für Goethe und zwei andere kanonisierte Autoren in den späten 1990er Jahren deutlich ansteigen, sie zugleich in der normalisierten Darstellung schon abfallen. Das bedeutet, dass bereits in der Phase von 1998-2006 ‚andere‘ Texte hinzugefügt wurden, deren Zusammensetzung anders war als das bisherige Korpus. Um welche Texte könnte es sich hierbei handeln?

Faktoren der Korpusverzerrung nach 1995

Die oben erwähnten verschiedenen Phasen der Veränderung legen es nahe zu vermuten, dass nicht ein einzelner Eingriff in das Korpus, sondern eine Reihe verschiedener Textinzufügungen die beobachteten Phänomene bedingen. In einem anderen Projekt, in dem es um die Analyse von Heftromanen geht, war bereits aufgefallen, dass sich die einschlägigen Verlage der Komplexität der Feststellung der richtigen Metadaten, insbesondere des Publikationsdatums, dadurch entledigt haben, dass sie einfach alle retrodigitalisierten Texte unter dem Datum veröffentlichen, an dem die digitale Kopie publiziert wird. Um zu testen, ob diese Texte auch im Ngram-Korpus sind, wurden die entsprechenden Serienautoren und -helden gesucht:

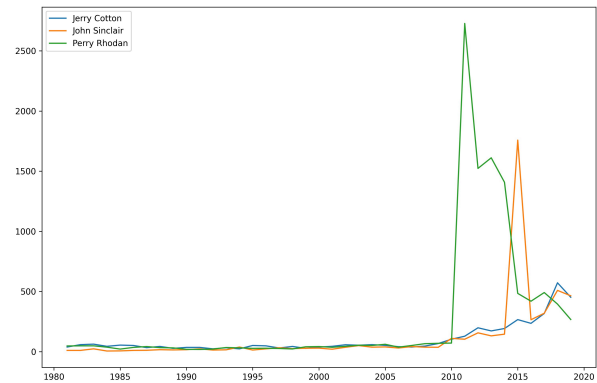


Fig. 8: Häufigkeit von Heftroman-Helden im Ngramm-Korpus (2019)

Die deutliche Steigerung ab 2010 spricht dafür, dass auch hier die Retrodigitalisate unter dem jeweiligen Jahr aufgeführt sind. Allerdings können selbst einige Tausend Heftromane nicht alleine verantwortlich sein.

Wie könnte man nun weitere Faktoren erkennen? Ein offensichtlicher Weg geht über das sprachliche Material. Die neuen Texte würden für bestimmte Worte zu einer relativen Erhöhung führen, selbst wenn viele andere Worte einen relativen Rückgang aufweisen. Da Google die Daten im 2019-Korpus mit Wortklasseninformation ausliefert, war es einfach, alle Substantive aus den 1-Grammen zu extrahieren, rd. 13 mio. Anschließend wurden die Substantive herausgefiltert, die von 1995 bis 2019 jedes Jahr auftauchen und zwar insgesamt mindestens 2 500 mal. Für diese restlichen rd. 350.000 Wörtern wurde für die Daten jedes Wortes eine lineare Regression berechnet und die Steigung der Geraden als Filterungsfaktor verwendet, um die rd. 250 Wörter zu finden, die in dieser Zeit den größten Anstieg verzeichnen.

Eine manuelle Sichtung dieser Wortliste zeigte den Einfluss mehrerer Textsorten. Vor allem aber fiel der einzige Name in der Liste auf: GRIN. Es handelt sich um eine deutsche Verlagsgruppe, zu der neben dem GRIN-Verlag selbst u.a. auch die Webseite hausarbeiten.de gehört. Der Verlag, der von Beobachtern als ‚vanity publisher‘ oder ‚predatory publisher‘ (Shrestha 2021) eingeschätzt wird, publiziert alle Texte digital oder als Book on Demand. Eine „Lektorierung findet nicht statt“ (Wikipedia). Laut Verlagswebseite wurden bis ins Jahr 2018 200.000 Texte publiziert. Seitdem sind schätzungsweise mindestens weitere 40.000 Titel publiziert worden.² Fig. 9 zeigt die Anzahl der Bücher im NGramm-Korpus, in denen das Token ‚GRIN‘ vorkommt, was wohl weitgehend identisch ist mit Titeln des Verlags.

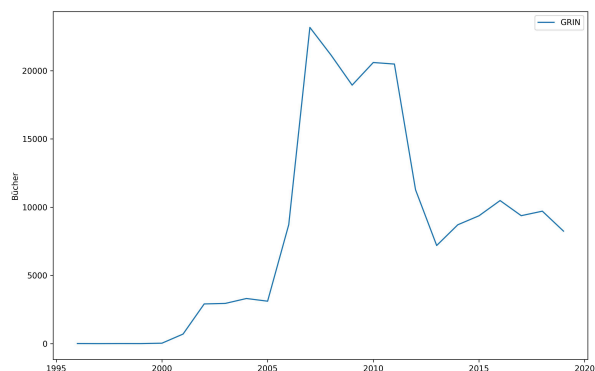


Fig. 9: Anzahl der Bücher im Ngramm-Korpus (2019), in denen das Wort 'GRIN' vorkommt.

Die Texte des GRIN Verlags haben zwar alle eine ISBN-Nummer, aber die meisten sind deutlich kürzer als Bücher im herkömmlichen Sinn, viele sind Aufsätze, die nur 20-30 Seiten lang sind. Beide Faktoren, die Unmenge an quasi-wissenschaftlichen kurzen Texten des GRIN-Verlags und die falsch datierten Heftromane führen dazu, dass die durchschnittliche Länge von Büchern im Ngramm-Korpus sich seit 2000 deutlich verringert hat (siehe Fig. 10), allerdings sind die Werte seit 2017 fast wieder auf dem alten Niveau:



Fig. 10: Durchschnittliche Anzahl der Seiten pro Buch im Ngramm-Korpus

Fassen wir zusammen: Nach 1998 ändert sich die Zusammensetzung des deutschsprachigen Ngramm-Korpus einschneidend, so dass es für die meisten Analysen zur Entwicklung von Sprache und Kultur weitgehend unbrauchbar wird. Dazu tragen eine Reihe von Faktoren bei, von denen zwei identifiziert werden konnten: Schwerwiegender ist, schon aus Umfanggründen, der Anteil der Publikationen des GRIN-Verlags. Sie geben zwar einen Einblick in eine bestimmte Form universitärer Wissenschaftskommunikation, haben aber nichts mit der sonstigen Buchproduktion zu tun. Hinzukommen die falsch datierten Retrodigitalisierungen einiger Verlage. Zugleich zeigt die Analyse der Daten, dass dies nicht die einzigen Faktoren sind, die hier ins Gewicht fallen. Wenn man sich auf die Wörter mit den steilsten Karrieren

ren in den letzten 25 Jahren konzentriert (Fig. 11), dann fällt auf, dass dies allgemeine Token sind, die sich eher in Romanen als in Fachtexten finden (besonders die Anführungszeichen, mit denen in den meisten deutschen Drucktexten direkte Rede markiert wird): „Augen Blick Du Frau Gesicht Hand Kopf Leben Mal Mann Moment Mutter Stimme Tag Tür Vater « »“ Da zugleich die Länge ansteigt und die Anzahl der Texte sehr hoch bleibt, außerdem diese Texte aber wohl nicht in den offiziellen Verlagsstatistiken auftauchen, handelt es sich vermutlich um Texte aus literarischen *Selfpublishing* Verlagen, die in der Umfrage des Börsenvereins nicht miteinbezogen waren. Darüber, ob diese nicht doch Teil eines Kulturgraphen sein sollten, lässt sich allerdings trefflich streiten.

An diese explorative Studie könnte nun eine Untersuchung anschließen, die die Veränderungen der Korpuszusammensetzung als überdurchschnittlich starke Veränderung der Token-Verteilungen formalisiert (Koplenig 2017) und so den hier etwas vernachlässigten Aspekt, wann sich genau die Veränderungen ergeben, herausarbeitet. Die ‚typische‘ Verwendung des Ngramm-Korpus und -viewers, nämlich die Untersuchung der Verwendungshäufigkeit von Termen in der schriftlichen Öffentlichkeit, ist durch die starken Schwankungen in der Zusammensetzung des Korpus sehr fragwürdig geworden. Da nach 2000 sonst eher randständige Bereiche, quasi-wissenschaftliche Texte und die Produktion der selbstverlegten Autorinnen und Autoren, das Korpus dominieren, ist es auch für rein sprachanalytische Untersuchungen, etwa zur Kollokationsanalyse, kaum verwendbar. Aber insgesamt verdient die Frage, ob und unter welchen Vorzeichen die Daten nicht doch für bestimmte Analysen herangezogen werden können, eine genauere Untersuchung.

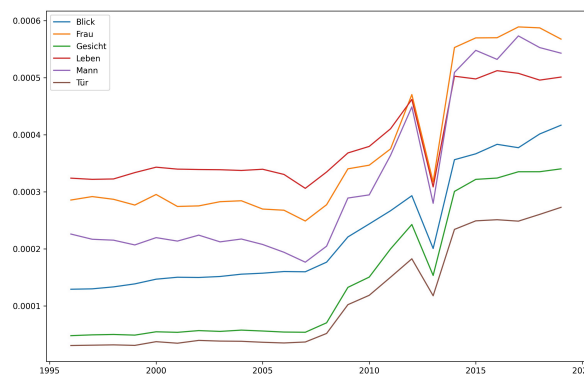


Fig. 11: Relative Häufigkeiten einiger Wörter, die 1995 bis 2019 zunehmend häufiger verwendet werden.

Fußnoten

1. Der Börsenverein dient in erster Linie der Interessenvertretung deutscher Firmen, auch wenn seit ca. 10 Jahren internationale Firmen Mitglied werden können.
2. Die Deutsche Nationalbibliothek verzeichnet rd. 394.000 Treffer für den GRIN-Verlag, allerdings wurden mindestens seit 2017 der größere Teil der Titel doppelt verzeichnet. Die DNB listet für den Zeitraum von 2019 bis heute rd. 78.000 Titel des GRIN-Verlags.

Bibliographie

Börsenverein des deutschen Buchhandels (2022) *Buchtitelproduktion: Anzahl der Neuerscheinungen in Deutschland in den Jahren 2002-2021*. Statista. Available at: <https://de.statista.com/statistik/daten/studie/39166/umfrage/verlagswesen-buchtitelproduktion-in-deutschland/> (Accessed: 8 March 2022).

Chen, Y. and Yan, F. (2016) 'Centuries of Sociology in Millions of Books'. Available at: <https://doi.org/10.1111/1467-954X.12399>.

Gonçalves, B. et al. (2018) 'Mapping the Americanization of English in Space and Time', *PLOS ONE*, 13(5), p. e0197741. Available at: <https://doi.org/10.1371/journal.pone.0197741>.

Koplenig, A. (2017) 'The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII', *Digital Scholarship in the Humanities*, 32(1), pp. 169–188. Available at: <https://doi.org/10.1093/llc/fqv037>.

Michel, J.-B. et al. (2011) 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science*, 331(6014), pp. 176–182. Available at: <https://doi.org/10.1126/science.1199644>.

Pechenick, E.A., Danforth, C.M. and Dodds, P.S. (2015) 'Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution', *PLOS ONE*, 10(10), p. e0137041. Available at: <https://doi.org/10.1371/journal.pone.0137041>.

Rahlf, T. (2015) *Deutschland in Daten: Zeitreihen zur Historischen Statistik*. Bonn: Bundeszentrale für politische Bildung.

Shrestha, J. (2021) *Vanity publishers: How to identify and avoid them*. Available at: <http://eprints.rclis.org/42635/> (Accessed: 2 August 2022).

Underwood, T. (2012) 'How not to do things with words', *The Stone and the Shell*, 25 August. Available at: <https://tedunderwood.com/category/ngrams/> (Accessed: 1 August 2022).

Korrespondenzen der Frühromantik: Ein kontrolliertes Vokabular zur Analyse von Kommunikation und Wissenstransfer für das Semantic Web

Suárez Cronauer, Elena

Elena.SuarezCronauer@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Fath, Laura

lfath@uni-mainz.de
Johannes Gutenberg-Universität Mainz

Deicke, Aline

aline.deicke@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Strobel, Jochen

jochen.strobel@uni-marburg.de
Johannes Gutenberg-Universität Mainz

Weyand, Sandra

weyands@uni-trier.de
Universität Trier

Burch, Thomas

burch@uni-trier.de
Universität Trier

Das Projekt „Korrespondenzen der Frühromantik“

Die Jenaer (und Berliner) Frühromantik gilt als die herausragende intellektuelle Revolution junger deutscher Autor*innen und Gelehrter an der Epochenschwelle um 1800. Die Gruppe agierte öffentlichkeitswirksam und nachhaltig, dispers und zugleich netzwerkbildend; sie reflektierte und praktizierte „Geselligkeit“ beispielsweise auch mittels der Kommunikationsform „Brief“. Die (auch quantitative) Auswertung dieser epistolaren Kommunikationsprozesse zwischen den Frühromantiker*innen einschließlich einer Untersuchung des dabei erfolgenden Wissenstransfers ist eines der großen Desiderate der Romantikforschung, dem das DFG-Projekt „Korrespondenzen der Frühromantik. Edition – Annotation – Netzwerkforschung“ begegnen möchte. Ein grundlegender Schritt auf dieses Ziel hin ist die Erstellung kontrollierter Vokabulare in Gestalt normierter Meta- und Registerdaten, die in einen Knowledge Graphen auf Grundlage einer domänenspezifischen Ontologie eingebunden werden sollen. Einen Aspekt dieser Modellierungsprozesse präsentiert dieser Beitrag.

Die Datengrundlage bilden die Briefe der wichtigsten Protagonist*innen der Frühromantik (wie z. B. Friedrich, Dorothea, August Wilhelm und Caroline Schlegel, Novalis u.a.) untereinander und mit ihren weiteren Korrespondenzpartner*innen zwischen 1790 und 1802, also von den diversen 'Vorgeschichten' bis zum Zerfall des Jenaer Kreises (Schanze 2018, 18). Die Daten werden systematisch und vollständig erfasst, digital in Open Access publiziert und literaturwissenschaftlich wie netzwerktheoretisch ausgewertet. Darunter fällt auch die semantische Annotation von Aussagen in diesen Brieftexten, bei der dort erfasste Registerentitäten, wie Personen,

Werke, Körperschaften oder Periodika, über eine zweiteilige Prädikatstruktur aus Illokution (dem eigentlichen Verb) und Proposition (einer Aussage, die dieses Verb näher spezifiziert) miteinander verknüpft werden. Zusammen mit den Register- und Metadaten der Briefe werden diese Annotationen in einen Knowledge Graphen überführt, in dem die Daten weiter angereichert werden. Auf dieser Basis erfolgen schließlich Auswertungen mittels quantitativer netzwerkanalytischer sowie qualitativer Ansätze. Die digitale Bereitstellung und Erschließung philologisch zuverlässiger Briefdigitalisate, die Annotation der Briefe sowie die parallele und abschließende graphen- und netzwerktheoretische Auswertung stellen demnach die drei Kernbereiche des Projekts dar.

Der Vortrag stellt die zwei Arbeitsphasen vor, in denen die kontrollierten Vokabulare der Illokutionen und Propositionen im Zusammenspiel von digitaler Edition und Annotation entwickelt werden: Zunächst wird ein festes Begriffssatz aus den Aussagen der Briefe destilliert und definiert, das danach in die Strukturen eines kontrollierten Vokabulars für die Verwendung als Linked Open Data überführt wird. Anschließend wird die Einbindung dieser Vokabulare in das Datenmodell des Knowledge Graphen präsentiert.

Entwicklung von Vokabularen zur semantischen Annotation von Aussagen

Der ursprüngliche Plan zur Erstellung des Begriffssatzes orientierte sich an Tripelstrukturen wie aus dem Semantic Web bekannt (Subjekt-Prädikat-Objekt). Insbesondere das Prädikat wurde, wie bereits angesprochen, im Laufe des Projekts jedoch differenziert und in zwei Kategorien – Illokution und Proposition – aufgesplittet. Diese können zu Analyse- und Publikationszwecken (z.B. als kontrolliertes RDF-Vokabular) wieder zusammengeführt werden, werden jedoch zur Gewährleistung größtmöglicher Flexibilität in der datenhaltenden Schicht zunächst in dieser ausführlicheren Variante vorgehalten. Ein den Semantic-Web-Prinzipien ähnlicher Ansatz findet sich auch in der *quantitative narrative analysis* (QNA) (Franzosi 2010): Nach der sozialwissenschaftlichen Methode bilden semantische Tripel die grundlegenden Einheiten einer Erzählung (Sudhahar u.a. 2015, 2). Von uns wird diese Vorgehensweise erstmals für die Annotation von Briefen genutzt. Durch die zweiteiligen Prädikate werden Meta- und Registerdaten der Briefe (Subjekt und Objekt) semantisch verknüpft. Subjekte können in diesem spezifischen Fall nur Personen sein, Objekte Personen, Körperschaften, Werke und Periodika.

Ziel dieser Auszeichnungen ist es, die Prozesse intellektuellen Austauschs und die Spuren kollaborativen Schaffens in den Brieftexten für eine formale quantitative Analyse zu öffnen. Auf Basis der kontrollierten Vokabulare der Illokutionen und Propositionen (sowie der schon in Gestalt der Meta- und Registerdaten vorgegebenen Registerinträge) werden die konkreten Akte, aus denen sich Strukturen von Kommunikation und Wissenstransfer ergeben, formal expliziert, womit sich weiterführende

datengestützte Auswertungsperspektiven im Sinne der Forschungsfrage ergeben.

Die Vokabulare entstehen teils *bottom up*, teils *top down*: mehrere Bearbeiter*innen annotieren parallel ein kleines Sample von Briefen. Aus dem im Team diskutierten Abgleich der gewählten Formulierungen ergibt sich eine Liste von zusammengesetzten Prädikaten im weiteren Sinne, die sich aus einem die jeweilige Kommunikationsfunktion bezeichnenden und eine kommunikative Handlung vollziehenden illokutionären Verb (also etwa „behaupten“, „erbitten“, „grüßen“, „positiv bewerten“) und fakultativ einer „welthaltigen“ Proposition, der Aussage, die hier auf den Punkt zu bringen ist, zusammensetzt, also etwa: „Publikation“, „Buchsendung“, „Arbeitsplan“. Die Proposition spezifiziert die abstrakte illokutionäre Aussage und beantwortet Fragen wie: „Was wird positiv bewertet?“ oder „Wozu wird jemand (das Objekt) aufgefordert?“ Bei der Erstellung dieser Liste(n) stellt der historische Bedeutungswandel der Sprache ein grundlegendes Problem dar. Wir versuchen diesem Problem zu begegnen, indem wir Anachronismen und begriffliche Überschneidungen mit romantischen Konzepten zu vermeiden versuchen. In einzelnen Fällen wird das nicht möglich sein. Hier können wir lediglich darauf verweisen, dass unser modernes Verständnis von „Kritik“ nicht identisch ist mit Friedrich Schlegels Begriff. Das Korpus der Propositionen soll 200 bis 300 nicht überschreiten, die Zahl der illokutionären Verben wird etwa 80 bis 90 erreichen. Mit diesen Termini schließen wir an sprachwissenschaftliche Standards, nämlich die Sprechakttheorie, an, die zwischen illokutionärem Akt und Proposition unterscheidet. Dabei sind nicht alle Aussagen eines Briefes zu annotieren. Die Auswahl ergibt sich aus folgenden Fragen: Sind identifizierte/identifizierbare Akteur*innen, Werke, Periodika, Körperschaften beteiligt? Ist die Aussage mit möglichst generischen Formulierungen zu erfassen? Kann die Aussage Fragen zu unseren Forschungsschwerpunkten beantworten? Wollte man auf diese Einschränkungen verzichten und eine komplette maschinenlesbare Paraphrase von 6.000 teils umfangreichen und thematisch oft äußerst heterogenen Briefen leisten, würde der dafür notwendige Zeitaufwand den eines finanzierbaren Forschungsprojekts überschreiten. Die Aussagenkette würde den Brief quantitativ um ein Mehrfaches übertreffen.

Bei der Erstellung der Begriffssatzes ist oberstes Gebot die intersubjektive Prüfbarkeit. Während Subjekt und Objekt positive Gegebenheiten sind, da Kopfdaten des Briefs oder Nennungen in den Registern, müssen Proposition und Illokution erst kontextuell abgeleitet werden. Eine Herausforderung, der man bei der Erstellung der Vokabulare begegnet, ist die Vereinheitlichung des zusammengesetzten Prädikats bei komplexen Aussagen, die auf zwei oder mehr Annotationsketten verteilt werden müssen. Dabei müssen die einzelnen Elemente – Illokution und Proposition – generisch und möglichst abstrakt gehalten werden, um das Vokabular zu begrenzen und eine quantitative Auswertung und potentielle Interoperabilität zu ermöglichen.

Auch die Entscheidung, was als Illokution gelten kann und was nicht, ist nicht immer einfach. Möglichst sollten keine Dopplungen von Begriffen im Propositions- und Illokutionsregister auftauchen wie beispielsweise *senden* als Illokution und in seiner substantivierten Form *Sen-*

„*ding*“ als Proposition. Andererseits können komplexere Aussagen wie „Buchsendung erbitten“, also das Phänomen sekundärer Prädikation, nur um den Preis gewisser Unschärfen ausgedrückt werden. So muss zuweilen auf Nuancen verzichtet werden, um die Konsistenz der Annotationspraxis zu gewährleisten. Hier dient das kontrollierte Vokabular auch der Organisation von Informationen (ANSI/NISO, 10).

Die intersubjektive Prüfbarkeit soll durch den Anschluss an sprachwissenschaftliche Standards garantiert werden. So werden Illokutionen und Propositionen jeweils Oberbegriffsklassen zugewiesen. Sie sind die Objektivation des *top down*-Elements bei der Annotationspraxis, die eben nicht allein auf einer abstrahierenden Paraphrase einzelner Briefpassagen beruht. Im Falle der Illokutionen ist das die Zuordnung zu Illokutionstypen nach Searle (Searle 1976, 1–23). Er teilt Verben in die Gruppen Assertive, Direktive, Kommissive, Expressive und Deklarative ein. Die Oberbegriffe zu den Propositionen sollen abgeglichen werden mit Erkenntnissen der Romantikforschung, der Briefforschung und dem Wissen über die Zeit um 1800 einschließlich ihrer Alltagskultur. Dies ist mehr noch als bei den Illokutionen ein *top down*-Element, das auf vorgängigen Erkenntnisinteressen und dem Stand der Forschung beruht. Mit diesen in einem Trial-and-Error-Verfahren zu erarbeitenden Begriffsklassen verfolgen wir mehrere Absichten: Wir dokumentieren den Anschluss an vorausgehende wissenschaftliche Überlegungen (wir vermeiden dabei Objektsprache), wir erhöhen den Grad an Disambiguierung in unserem Korpus (oder stellen uns den zwischen den Klassen sicherlich unvermeidlichen Ambiguitäten, wenn etwa Unterbegriffe mehrfach zugeordnet werden können), wir geben Nutzer*innen unserer Editionsplattform als zusätzliche Suchoption Registerelemente an die Hand, erhöhen also die Usability unserer Textsammlung, und fügen schließlich für maschinelle Auswertungen eine zentrale, wenngleich händisch erhobene Datenquelle hinzu. Dieses Angebot soll u.a. die literaturwissenschaftliche Forschung befruchten, aber auch ein Beitrag sein zur Erforschung der Leistung der Kommunikationsform „Brief“ in ihrer Textualität. Nicht zuletzt wird im Projekt selbst ein Wechselspiel von quantitativem und qualitativem Arbeiten erprobt, aus dem neue Erkenntnisse und Arbeitsweisen für die digitalen Geisteswissenschaften hervorgehen können.

Implementierung in den Knowledge Graphen

Für die Implementierung in den Knowledge Graphen dient folglich zunächst ein festes Set aus Begriffen als Grundlage, das die Briefaussagen, die im Zusammenhang mit Kommunikationsprozessen und dem daraus folgenden Wissenstransfer stehen, als die Verknüpfung von Registerentitäten durch Illokution und Proposition mit zugehörigen Oberklassen abbildet. Dieses Begriffset wird bei der Modellierung einer Ontologie der „Korrespondenzen der Frühromantik“ berücksichtigt und anschließend in einer RDF-Serialisierung in ein kontrolliertes Vokabular im Sinne des Semantic Webs überführt.

Eine Ontologie wird als eine formale und explizite Spezifikation einer gemeinsamen Konzeptualisierung von Wissen definiert (Gruber 1993, 199). Ein kontrolliertes Vokabular als eine „Zusammenstellung von Bezeichnern (URIs) mit klar definierter Bedeutung“ (Hitzler u.a. 2008, 48) stellt zunächst Informationen dar, identifiziert diese Informationen darüber hinaus aber auch nochmals eindeutig in einer maschinenlesbaren Form und bildet somit die Voraussetzung für die Beschreibung von semantischen Beziehungen innerhalb einer Ontologie. Hier werden diese Begriffe dann in komplexe, ggf. hierarchische Beziehungen gesetzt.

Herausfordernd ist in vorliegendem Fall die Transformation der oben beschriebenen Aussagen in Konzepte: Die Aussagen in den Briefen spiegeln deren natürliche Sprache; sie werden in den Vokabularen aus Illokutions- und Propositionsklassen formalisiert und in einen semantischen Zusammenhang gestellt. Diese Arbeit wird schließlich im Knowledge Graphen fortgesetzt, indem diese Aussagen als Konzepte innerhalb einer Ontologie abgebildet werden. Der Begriff 'Konzept' wird hier als eine aus der Wahrnehmung abstrahierte Vorstellung, also als eine mentale, wortähnliche Repräsentation von Dingen verstanden und somit eben nicht als ein spezifischer, in einem Brief beschriebener Sachverhalt (Margolis 2022). Von vornherein werden nicht lediglich die Aussagen der Briefe rekonstruiert bzw. formalisiert. Vielmehr wird das übergreifende und somit für weitere semantische Verarbeitung relevante Konzept hinter dieser Aussage bereits in den generischen Elementen des Begriffssets sichtbar. So ist beispielsweise nicht relevant, dass Schlegel in seinem Brief Schleiermacher um seine Kritik an einer bestimmten von ihm übersetzten Textpassage bittet. Relevant sind vielmehr die Begriffe „Kritik“ und „erbitten“ bzw. „Bitte“ (also: Proposition und illokutionäres Verb) im Verhältnis zu Subjekt und Objekt, da durch diese Konzepte sowohl die Selbstreferentialität der Kommunikation (Illokutionen) als auch die relevanten Themen der Kommunikationsprozesse (insbesondere der Wissenstransfer) angesprochen werden. Wie diese Aussagen in das kontrollierte Vokabular übernommen werden – z.B. ob man „Kritik erbitten“ als ein Konzept wertet oder nicht – hat Auswirkungen auf das weitere Datenmodell, also die Ontologie, und ihre Fähigkeit, generische Aussagen der frühromantischen Wissensdomäne abzubilden.

Um dieser Problematik für das Datenmodell adäquat zu begegnen, wurden zunächst verschiedene Ansätze geprüft, die sich mit der Modellierung geisteswissenschaftlicher Daten, insbesondere Briefdaten, beschäftigen.¹ Das Datenmodell der „Korrespondenzen der Frühromantik“ referenziert auf das Cidoc Conceptual Reference Model (CRM) als Upper Ontology. Auch wenn das Cidoc CRM eigentlich als eine Ontologie für die Museumsdomäne entwickelt wurde, folgt die Orientierung daran der Definition einer Ontologie als einer *gemeinsamen* Konzeptualisierung von Wissen, da sich zahlreiche geisteswissenschaftliche Projekte an dem Modell orientieren. Zudem kommt die logische Ausrichtung des Cidoc CRMs auf Ereignisse (*event driven architecture*) den Forschungsfragen des Projekts zugute, da so bspw. die Korrespondenzen als Prozesse modelliert werden und somit Flexibilität innerhalb des Modells garantiert wird. So wird jede Klasse der frühromantischen Domänenontologie als Superklasse einer Klasse des Cidoc CRMs

übergeordnet oder es werden, falls möglich, die Klassen des Cidoc CRMs direkt nachgenutzt. Ebenso wird bei den Properties vorgegangen. Darüber hinaus wurden Aspekte der Auszeichnungslogik von *correspdesc* miteinbezogen, da die bereits publizierten Daten der August-Wilhelm-Schlegel-Edition (Strobel 2014-2020) dieser folgen.

Für die Modellierung der Aussagen bedeutet dies nun Folgendes: Innerhalb des Konzeptes Brief (Klasse *Letter* mit Superklasse *E33 Linguistic Object* des Cidoc CRMs) können Personen (Klasse *E21 Person*), Zeitschriften (*Periodical* mit Superklasse *E33*), Werke (*Work* mit Superklasse *E33*), Institutionen (*Institution* mit Superklasse *E74 Group*), Schlagwörter (*Theme* mit Superklasse *E55 Type*) und Aussagen (*Statement* mit Superklasse *E13 Attribute Assignment*) miteinbezogen sein. Verbunden sind diese Klassen mit *Letter* jeweils mit der Property *P129 is about* des Cidoc CRMs. Das kontrollierte Vokabular wird in den Klassen *Illocution* und *Proposition* abgebildet, beide mit Superklasse *E55 Type*. *Illocution* ist zudem mit der Gruppe der Illokutionstypen durch die Klasse *Illocutiongroup* verbunden, die ebenfalls die Superklasse *E55 Type* hat. Innerhalb der Klasse *Statement* als Domain werden die Aussagen nun modelliert, indem das Subjekt durch *has_Subject* (Subproperty von *P140 assigned attribute to*) mit *E21 Person* als Range verbunden wird, das Objekt als Range mit den Klassen *E21 Person*, *Periodical*, *Work* und/oder *Institution* mit der Property *has_Object* (Subproperty von *P141 assigned*) sowie das Prädikat mit *has_Predicate* (Subproperty von *P177 assigned property of type*) mit den Klassen *Illocution* und *Proposition* (Abb. 1). Die jeweiligen Einheiten einer Aussage werden somit als gesonderte Klassen betrachtet, die erst innerhalb der Aussage selbst wieder zusammengeführt werden, wobei die Properties die Subjekt-Illokution-Proposition-Objekt-Struktur festlegen. Perspektivisch bedeutet dies, dass bspw. die Proposition „Kritik“ und die Illokution „erbitten“ erst für die Netzwerkanalyse aus zuvor getrennt vorgehaltenen Einheiten der Graphdatenbank zusammengesetzt werden. Hierdurch wird eine flexible Struktur garantiert: Die Komplexität der Konstruktion von Aussagen wird durch das Aufgliedern in ihre einzelnen Bestandteile reduziert, sodass eine Nachnutzung der Daten – unter der Berücksichtigung der Auswahl der Begriffe, die von den angesprochenen Forschungsfragen bestimmt war – auch in anderen Kontexten möglich ist. Es gilt dementsprechend aus den spezifischen und eine konkrete „Sache“ betreffenden Aussagen des Quellmaterials generische und allgemeine Konzepte zu entwickeln und diese als Linked Open Data darzustellen, welche dann einerseits für den im Projekt zu entwickelnden Knowledge Graphen bzw. die Ontologie genutzt, andererseits aber auch für andere Forschungsinteressen, welche die Zeit um 1800 oder das Kommunikationsmedium „Brief“ betreffen, nachgenutzt werden können. Diese Entwicklung generischer und offener Forschungsdaten aus konkretem Quellenmaterial stellt eine Aufgabe dar, denen viele Forschungsprojekte aus dem Bereich der Digital Humanities begegnen, die jedoch nicht minder signifikant ist: Nur so werden Forschungsdaten erzeugt, die auch über den Projektkontext hinaus genutzt werden und folglich die Forschungslandschaft ergänzen können. Zudem können durch den Linked Open Data-Ansatz die projektinternen Forschungsdaten an andere Datenkor-

pora angeschlossen werden, was einerseits zu einer größeren Sichtbarkeit, andererseits zu vielfältigen Anschlussperspektiven führt, beispielsweise für die Historische Netzwerkforschung, die Geschichtswissenschaften (v.a. zur Kultur-, Sozial- und Ideengeschichte, aber auch zu Alltags- und Emotionsgeschichte), die Genderforschung oder die Judaistik.

Ausblick

Durch den Aufbau solcher kontrollierter Vokabulare, die Entwicklung des Datenmodells und die Integration in einen Knowledge Graphen eröffnet sich die Möglichkeit tiefergehender Analysen mit Methoden der historisch-geisteswissenschaftlichen Netzwerkforschung, innerhalb derer die Strukturen des Netzwerks der Jenaer (und Berliner) Frühromantik sowie seine Genese und Entwicklung als relationale Phänomene rekonstruiert werden. So wird ein neues Bild der einleitend erwähnten Akteur*innen gezeichnet, werden zentrale Personen, homo- und heterogene Strukturen, Überschreitungen sozialer Barrieren und die Kreuzung sozialer Kreise ausgemacht sowie Dynamiken dieser Strukturen aufgeschlüsselt. Da sich die Netzwerkforschung in Bezug auf die umfassende, auch inhaltliche Erschließung von Korrespondenzen noch in den Anfängen befindet, ergibt sich in der Kombination mit einer umfassenden, hochstrukturierten Datenbasis die Möglichkeit, neue Forschungsansätze in der Synthese literaturwissenschaftlicher und netzwerktheoretischer Arbeitsweisen zu entwickeln und für den Anwendungsfall der Jenaer Frühromantik eingehend zu prüfen.

Abbildung

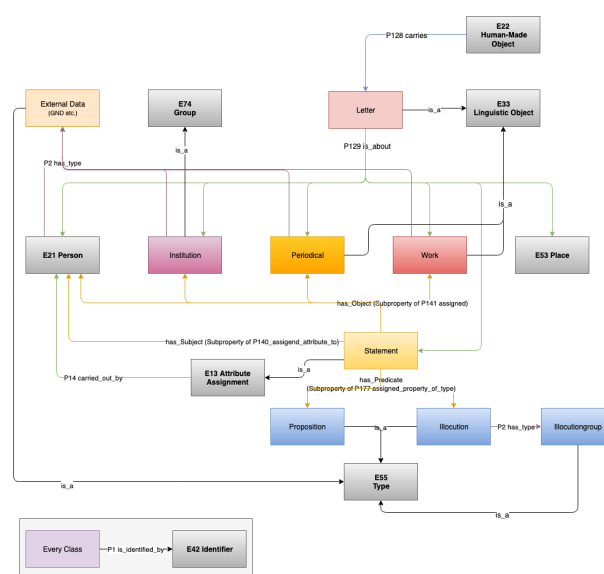


Abb. 1: Ausschnitt aus Datenmodell der „Korrespondenzen der Frühromantik“

Fußnoten

1. Hier sei auf das CIDOC CRM (Bekiari u.a. 2022), das „Letter Model“, das im Rahmen der COST Action „Re-assembling the Republic of Letters“ entwickelt wurde (Jeffries u.a. 2019), das *correspdesc*-Element der TEI Guidelines (TEI Consortium [Hg.] 2021), das European Data Model (Doerr u.a. 2010), die Ontologie „OntoAndalus“ (Almeida u.a. 2021) sowie die Ontologie „OntoBellini-Letters“ (Cristofaro u.a. 2022) verwiesen.

Bibliographie

Almeida, Bruno, and Rute Costa. „OntoAndalus: an ontology of Islamic artefacts for terminological purposes.“ *Semantic Web* 12.2 (2021): 295–311.

ANSI/NISO. 2005. *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, 10. <https://www.niso.org/publications/ansi-niso-z3919-2005-r2010> (zugegriffen: 28. Juli 2022).

Bekiari, Chrysoula, George Bruseker, Martin Doerr, Ore Christian-Emil, Stead Stephen, Athanasios Velios, Erin Canning, und Philippe Michon. 2022. *Volume A: Definition of the CIDOC Conceptual Reference Model. Version 7.2.1*. ICOM/CIDOC Documentation Standards Group/CRM Special Interest Group. http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v.7.0_%2020-6-2020.pdf (zugegriffen: 28. Juli 2022).

Cristofaro, Salvatore, Pietro Sichera und Daria Spampinato. 2022. „An ontology proposal for a corpus of letters of Vincenzo Bellini: formal properties of physical structure and the case of rotated texts“. *International Journal of Metadata, Semantics and Ontologies* 15, Nr. 4: 269–279.

Dumont, Stefan, Ingo Börner, Dominik Leipold, Jonas Müller-Laackman und Gerlinde Schneider. 2019. „Correspondence Metadata Interchange Format.“ In *Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf*, hg. von Stefan Dumont, Susanne Haaf und Sabine Seifert. Berlin. <https://encoding-correspondence.bbaw.de/v1/CMIF.html> (zugegriffen: 28. Juli 2022).

Franzosi, Roberto. 2010. *Quantitative narrative analysis*. Los Angeles u.a.: SAGE Publications (=Quantitative Applications in the Social Sciences 162).

Gruber, Thomas. 1993. „A Translation Approach to Portable Ontology Specifications.“ *Knowledge Acquisition* 5, Nr. 2: 199–220.

Hitzler, Pascal, Markus Krötzsch, Sebastian Rudolph und York Sure. 2008. *Semantic Web: Grundlagen*. Berlin: eXamen.press.

Jeffries, Neil, Howard Hotson, Christoph Kudella, Miranda Lewis, Thomas Stäcker, und Gertjan Filariski. 2019. „Letter Model“. In *Reassembling the Republic of Letters in the Digital Age. Standards, Systems, Scholarship*, hg. von Howard Hotson und Thomas Wallnig, 171–89. Göttingen: Universitätsverlag Göttingen. <http://www.univerlag.uni-goettingen.de/handle/3/ISBN-978-3-86395-403-1> (zugegriffen: 28. Juli 2022).

Margolis, Eric und Stephen Laurence, „Concepts“. In *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), hg. von Edward N. Zalta und Uri Nodel-

man, <https://plato.stanford.edu/archives/fall2022/entries/concepts/> (zugegriffen: 07. Dezember 2022).

Miles, Alistair und Dan Brickley. 2005. „SKOS Core Vocabulary“ <https://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/> (zugegriffen: 28. Juli 2022).

Schanze, Helmut. 2018. *Erfindung der Romantik*. Stuttgart: Metzler.

Searle, John R. 1976. „A classification of illocutionary acts.“ *Language in Society* 5, Nr. 1: 1–23.

Sudhakar, Saatviga, Giuseppe A Veltri und Nello Cristianini. 2015. „Automated Analysis of the US Presidential Elections Using Big Data and Network Analysis.“ *Big Data & Society* 2, Nr. 1. <https://doi.org/10.1177/2053951715572916> (zugegriffen: 28. Juli 2022).

Strobel, Jochen und Claudia Bamberg (Hg.). 2014–2020. *August Wilhelm Schlegel: Digitale Edition der Korrespondenz*. <https://august-wilhelm-schlegel.de> (zugegriffen: 07. Dezember 2022).

TEI Consortium (Hg.) 2021. „2.4.6 Correspondence Description.“ *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 4.2.1. TEI Consortium. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD44CD> (zugegriffen: 28. Juli 2022).

Lemmabasierte Publikationsformate weiter denken mit dem "Open Encyclopedia System"

Grote, Brigitte

brigitte.grote@fu-berlin.de
Freie Universität Berlin, Deutschland

Strobl, Maren

maren.strobl@fu-berlin.de
Freie Universität Berlin, Deutschland

Einleitung

Eine zentrale Anforderung an die „Offenheit“ in der Wissenschaft, für die Geisteswissenschaften unter dem Label „Open Humanities“ eingeführt, ist die Publikation in offenen Formaten (Kleineberg, Kaden 2017; CfP Dhd 2023¹). Bisher orientieren sich geisteswissenschaftliche digitale Publikationen meist an den etablierten Printformaten Monographie, Zeitschriftenartikel und Sammelwerk², mit *Open Monograph Press* und *Open Journal System*³ liegen erprobte Open-Access-Publikationsinfrastrukturen für Buchpublikationen und Zeitschriften vor. „Offenheit“ erfordert jedoch mehr als die freie Zugänglichkeit, sondern betrifft weiterhin die Offenheit der der

Publikation zugrundeliegenden Daten (Open Data), des Quellcodes der Publikationsinfrastruktur (Open Source) und eine strukturelle Offenheit der Veröffentlichungen (u.a. Kleineberg, Kaden 2017; CfP Dhd 2023). Angesichts einer sich verändernden Publikationskultur in den Geisteswissenschaften ist es daher nicht mit der Schaffung digitaler Äquivalente zu Printformaten getan, vielmehr sollten Online-Publikationslösungen der Diskussion um alternative Veröffentlichungsformen Rechnung tragen: Zu nennen sind hier kollaborative Autorschaft (Kaden 2016, 4) und diskursive Schreibprozesse mit einer Bewegung „von der Ergebnis- zur Prozesspublikation“ (AG Digitales Publizieren 2021, 13), kleinformatige Veröffentlichungsformen und eine „strukturelle Ausdifferenzierung des Publikationsobjektes“ (Kaden 2016, 19), die semantisch annotiert sind. Entsprechend fordert der Wissenschaftsrat (2022, 54) dass „Voraussetzungen dafür zu schaffen [seien], dass auch neue, innovative Publikationsorte entstehen“ können.

Mit dem *Open Encyclopedia System* (OES)⁴ existiert eine webbasierte Plattform zur Erstellung, Publikation und Pflege von lemmabasierten Online-Publikationen. Darunter verstehen wir eine Sammlung von wissenschaftlichen Beiträgen zu einem bestimmten Themen- und/oder Methodenbereich, die von unterschiedlichen Autor:innen namentlich gekennzeichnet erstellt werden und deren Publikation durch ein Redaktionsteam betreut wird. OES unterstützt für diese Publikationsform die inkrementelle Erstellung und Verwaltung von Inhalten und Metadaten bis hin zur Einbettung von multimedialen Inhalten, Normdaten und bibliographischen Informationen, die Registererstellung, die Veröffentlichung mehrsprachiger zitierfähiger Artikel sowie vielfältige Formen der Navigation und Suche in und Interaktion mit den publizierten Inhalten, möglichst unter Berücksichtigung der o.g. Facetten von „Offenheit“. OES stellt so zum einen eine Open-Source-Anwendung für eine weitere zentrale Publikationsform der Geisteswissenschaften dar, und bietet zum anderen Lösungsansätze für offene und innovative Formate. Damit grenzt sich OES von bestehenden Publikationsinfrastrukturen für wissenschaftliche Online-Referenzwerke ab, die entweder proprietäre Lösungen oder an Verlagen angesiedelt sind, nicht primär für den wissenschaftlichen Gebrauch (Wikimedia) gedacht oder noch in Entwicklung (OAPEnz) sind.⁵

OES wurde im Rahmen einer DFG-Förderung (LIS)⁶ am Center für Digitale Systeme der Freien Universität Berlin als Open Source Software entwickelt.⁷ Seit 2019 werden mit fachwissenschaftlichen Partner:innen Online-Publikationen mit OES realisiert:⁸ wissenschaftliche Online-Enzyklopädien, Online-Compendien und -Lexika, sowie aktuell „Living Handbooks“ zur (kollaborativen) Verschriftlichung, Diskussion und Präsentation der Leitbegriffe interdisziplinärer Forschungsgruppen. OES wird kontinuierlich weiterentwickelt; Betrieb und Pflege der Software werden durch die Freie Universität Berlin gesichert.

OES-Systemarchitektur und -Komponenten

Die OES-Software basiert auf dem Open-Source Content-Management-System WordPress⁹ mit übernimmt dessen Systemarchitektur mit der Unterscheidung in Publikations- und Redaktionsumgebung, und ist als WordPress-Plugins mit zugehörigem Themes umgesetzt (vgl. Abb. 1). OES erweitert WordPress um vielfältige Funktionen zur Erstellung, Publikation und Pflege lemmabasierter Publikationen.

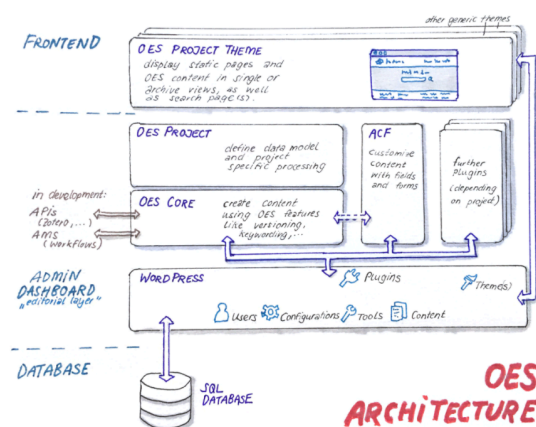


Abb.1: Aufbau der OES-Software

Über die interaktive Präsentationsschicht („frontend“) ist das Sammelwerk online zugänglich. Für die Präsentation der digitalen Inhalte entsprechend der definierten Datenstrukturen wurde ein WordPress-Theme entwickelt.¹⁰ Das OES-Theme erfüllt die gängigen Webstandards und ermöglicht eine barrierefreie und responsive Darstellung. Im Redaktionssystem („editorial layer“) werden die redaktionellen Prozesse der Inhaltserstellung und Datenpflege sowie die Administration der Präsentationsschicht standardübergreifend und kollaborativ organisiert. Der OES Core-Plugin („OES Core“) enthält die OES-spezifischen WordPress-Erweiterungen und kann mit weiteren WordPress-Plugins kombiniert werden. OES nutzt für das Arbeiten mit strukturierten Daten das Open-Source Plugin Advanced Custom Fields (ACF).¹¹ Die erstellten Inhalte werden in einer relationalen mySQL-Datenbank („database“) verwaltet und können über eine csv-Schnittstelle strukturiert importiert und exportiert werden. WordPress selbst ermöglicht den Export der Datenbank als SQL-Datensatz und über eine REST API in XML/RDF-Formate. Weitere Exportformate können projektabhängig festgelegt und implementiert werden.

OES unterscheidet standardisierte Datenstrukturen und Funktionalitäten, die im OES-Core und OES-Theme zusammengefasst werden, von projektspezifischen Erweiterungen, die die zusätzlichen Bedarfe einzelner OES-Anwendungen abdecken. Jede OES-Instanz implementiert über einen OES Projekt-Plugin („OES Project“) ein projektabhängiges Datenmodell, über welches projekt-

spezifische Datenstrukturen, Präsentationen und Interaktionen der digitalen Inhalte definiert werden. Über Konfigurationsoptionen in der Redaktionsumgebung („OES-Settings“) lassen sich zahlreiche Aspekte der Präsentation administrieren; für weitere projektspezifische Anpassungen kann in Ergänzung zum OES-Theme ein Child-Theme („OES Project Theme“) erstellt werden.

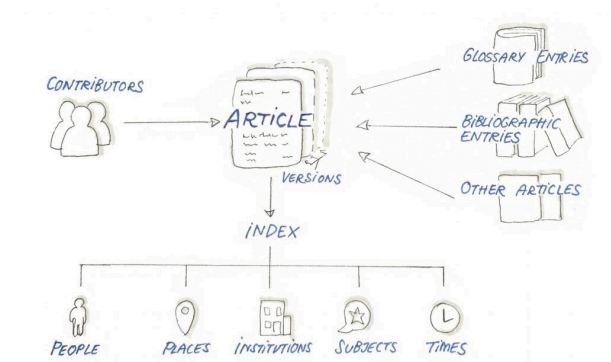


Abb. 2: Exemplarisches Datenmodell für OES

Abb. 2 illustriert anhand der für lemmabasierte Publikationen charakteristischen Inhaltstypen die Struktur eines OES-Datenmodells mit Datenobjekten wie Lemma, Verfasser:in und bibliographischer Eintrag, die mit Attributen angereichert werden können, und die über Relationen in Beziehung zu anderen Datenobjekten stehen. Schnittstellen zu externen Datenbeständen wie Normdatenbanken und Zotero¹² erlauben eine dynamische Verknüpfung mit diesen strukturierten Informationssammlungen und die Übernahme von Daten. Attribute und Relationen werden bei der Darstellung in der Publikationsumgebung ausgewertet und unterstützen eine dynamische Anzeige der Daten.

Mit OES im Open Access publizieren

OES unterstützt die Open-Access-Publikation von lemmabasierten Werken und reagiert dabei auf die Anforderungen, die an zeitgemäße digitale Publikationen formuliert werden:¹³ Über die Redaktionsumgebung kann jedes Lemma separat lizenziert werden; OES unterstützt dabei die Auswahl unterschiedlicher CC-Lizenzen. Ebenso können für multimediale Elemente die spezifischen Lizenzierungen und Rechte einzeln ausgewiesen werden. Referenzierbarkeit und Zitierfähigkeit sind durch Angabe einer DOI gegeben; dieses gilt jeweils für die Fassung (Version oder Sprachfassung) eines Artikels, sodass die Persistenz der Referenz garantiert ist. Autor:innen und andere Beitragende können einzeln ausgewiesen und mit ORCID/GND-IDs aufgeführt (vgl. AG Digitales Publizieren 2021, 41) und „verschiedene Autorschafts- und Beiträge*innenrollen“ (ebd., 18) sichtbar gemacht werden. Eine standardisierte Zitierweise als obligatorische Angabe wird durch eine automatische Generierung nach einem durch die Redaktion definierten

Muster¹⁴ garantiert (vgl. Abb. 4). Neben offener Zugänglichkeit und stabilem Nachweis ist die menschliche und maschinelle Weiternutzung der Inhalte eine zentrale Anforderung an offene Publikationen. Die Bereitstellung der XML/TEI-Kodierung der generischen Textstruktur und der typischen Entitäten im Frontend ist in Vorbereitung. Ein PDF-Download kann über das Frontend bereitgestellt werden (vgl. Abb. 3).



Abb. 3: Ansicht eines Artikelheaders am Beispiel des Compendium Heroicum des SFB 948¹⁵

Offene Inhalte und Daten erstellen und pflegen

Die OES-Redaktionsumgebung umfasst Funktionen zur Inhaltserstellung und Datenpflege, die kollaborativ von einem Redaktionsteam wahrgenommen werden. Zentrales Datenobjekt ist das Lemma („article“, Abb. 2), welches mit Inhaltsverzeichnis, Verfasser:in, Querverweisen und externen Verlinkungen, Einzelnachweisen in Form von Endnoten, bibliographischen Angaben, Zitervorschlag und inhaltlichen und strukturellen Metadaten erstellt und publiziert werden kann (vgl. Abb. 3 und 4). OES unterstützt den Schreibprozess durch den um OES-spezifische Funktionalitäten erweiterten Gutenberg-Editor¹⁶ und durch an die OES-Inhaltstypen angepassten WordPress-Formulare für die Erstellung und Pflege der Vokabulare. Neben Textblöcken können multimediale Elemente einfach eingebunden und Querverweise innerhalb des Sammelwerks und zu externen Datenbeständen integriert werden. Für bibliographische Daten ist ein Import aus dem Literaturverwaltungsprogramm Zotero oder eine dynamische Anbindung über das Zotpress-Plugin¹⁷ möglich. Die Qualitätssicherung der Inhalte erfolgt in den aktuellen Anwendungen durch die Redaktion; die Review-Prozesse werden i.d.R. außerhalb von OES organisiert.¹⁸

OES geht über das an Printformaten orientierte Konstrukt einer „Ergebnispublikation“ hinaus und ermöglicht die Umsetzung dynamischer und kollaborativer Inhaltserstellung im Sinne einer Prozesspublikation (u.a. AG Digitales Publizieren 2021): Lemmata sind in OES veränderbar und können nach Erstveröffentlichung weitergeschrieben werden. Sie existieren in verschiedenen Text- und Sprachfassungen, den sog. Versionen, die über ein „Stammlemma“ miteinander in Beziehung stehen, und die Inhalte und strukturierte Daten teilen können.

Dieser Abkehr einer reinen Ergebnispublikation geht mit der Herausforderung einher, Zitierfähigkeit und Referenzierbarkeit zu erhalten. OES sieht deshalb vor, dass alle Versionen mit einer eigenen DOI versehen werden können. Die aktuellste Version kann wie in Abb. 3 als Volltext, weitere als Liste angezeigt werden.

Inkrementelles und dynamisches Publizieren kann noch weitergedacht werden als eine „strukturelle Ausdifferenzierung des Publikationsobjektes in unterschiedlich verarbeitbare und aktualisierbare Teile“ (Kaden 2016, 19). OES bietet erste Lösungen, indem es den modularen Aufbau von Lemmata ermöglicht,¹⁹ und so eine dynamische Bündelung und Ausgabe von Inhalten im Frontend unterstützt. Im Rahmen verschiedener Anwendungen²⁰ werden aktuell strukturierte Formate und modulare Ansätze ausgearbeitet, um das „Dokument selbst als Einheit zu öffnen und zu verflüssigen“ (Kaden 2016, 19). Zu betrachten ist hierbei die Ausweitung des Autor:innenbegriffs, da diese Form der Publikation i.d.R. mit geteilter Autorschaft einhergeht, die neue Zitationspraktiken erfordern.

The screenshot shows the 'compendium heroicum' website. At the top, there are navigation links: 'ÜBER UNS', 'ARTIKEL A-Z', 'RUBRIKEN', 'INDEX', and a search icon. Below the navigation bar, the 'Zitierweise' (Citation) section is displayed, followed by a list of authors and their affiliations. The 'Metadaten' (Metadata) section is highlighted, showing various fields such as DOI, Lizenz, Rubrik, Schlagworte, Karlsruher Virtueller Katalog, and Index. The 'Index' field lists a long list of authors and their works, including Achilles (Figur), Homer, Agamemnon (Figur), Johann Wolfgang Goethe, Hector (Figur), Gilgamesch (Figur), Aristoteles, Wilhelm Tell (Figur), Jossia (Figur), Eliza Harris (Figur), Harriet Tubman, Harriet Beecher Stowe, Mweido (Figur), Shemwindo (Figur), Werther (Figur), George Cruikshank, Joseph Campbell, Wladimir Propp, Hans Robert Jauss, Räume: Russland, USA, Europa, Afrika, Jordan, Ohio River, Epochen: Antike, Frühe Neuzeit, Mittelalter, Amerikanischer Bürgerkrieg / Sezessionskrieg (1861–1865), Moderne, Postmoderne.

Abb. 4: Anzeige von Zitierweise und Metadaten am Beispiel des Compendium Heroicum des SFB 948²¹

Auf der Präsentationsebene werden den Nutzenden vielfältige Optionen zur Erschließung der erstellten Inhalte angeboten: Die Inhalte können neben der für Referenzwerke typischen alphabetischen Listenansicht über thematische Rubriken, Register sowie über eine Volltextsuche mit Filteroptionen (facettierte Suche) aufgerufen werden. Für einzelne OES-Anwendungen wurden Visualisierungen wie Karten und Zeitstrahl zur Exploration der Inhalte entwickelt.²² Die Unterschiede zwischen den Anwendungen manifestieren sich aufgrund spezifischer Anforderungen der Zielgruppe und der Daten vor allem im Frontend, und erfordern projektspezifische Anpassungen (über OES-Konfigurationsoptionen) bzw. Erweiterungen (projektspezifische OES-Plugins und Themes).

Datenbestände verknüpfen und verfügbar machen

Wie oben beschrieben werden die in OES gemäß dem Datenmodell erstellten Forschungsdaten über normierte Schnittstellen bereitgestellt. Gleichzeitig greift OES aber auch zum Zwecke der strukturierten Beschreibung, der eindeutigen Referenzierung, der Vernetzung der Inhalte sowie einer verbesserten Auffindbarkeit (u.a. Wissenschaftsrat 2022, 9; AG Digitales Publizieren 2021, 14) der OES-Datenbestände auf Normdateien und Verbundkataloge zu. OES unterstützt die Suche in und Datenübernahme aus ausgewählten Normdateien: Über die von lobid bereitgestellte API²³ werden GND-Daten abgefragt, ortsgebundene Normdaten werden direkt vom Ortsnamensdienst Geonames²⁴ abgerufen und die Subject Headings von der Library of Congress²⁵. Eine Anbindung von Wikidata und der Schlagwortnormdatei Rameau befindet sich in Umsetzung. Eine in der Redaktionsumgebung integrierte Auswahlmaske ermöglicht die Auswahl von GND- bzw. LCSH-Schlagworten und deren Zuweisung zu Lemmata sowie die Generierung von „shortlinks“ zu GND- und Geonames-Kennungen, mit denen Named Entities in den Volltexten verknüpft werden können. Diese Verweise ermöglichen im Frontend die dynamische Anzeige von ausgewählten Informationen aus dem verknüpften Normdatensatz. Weiterhin können über die LOD-Auswahlmaske für strukturierte Datenobjekte semi-automatisch Angaben aus den Normdatenbanken in den OES-Datenbestand übernommen werden. Weitere Verknüpfungsmöglichkeiten ergeben sich durch die bereits erwähnte persistente Identifizierung der Autor:innen mit ORCID und GND-ID sowie die dynamische Verknüpfung mit bibliographischen Einträgen in Zotero-Bibliotheken.

Diskussion und Ausblick

Das *Open Encyclopedia System* fügt den Angeboten für wissenschaftliches digitales Publizieren eine weitere Open-Access-Lösung hinzu. Dabei bildet OES nicht primär bestehende Printformate nach, sondern strebt innovative und offene Publikationsformen an. Während hinsichtlich Open-Access-Veröffentlichungen und offene Daten tragfähige Lösungen entstanden sind, sind im Kontext offener Formate, sowohl inhaltlich als auch strukturell, zwar erste Lösungsansätze benannt, zu deren weiterer Ausgestaltung bedarf es jedoch noch konzeptioneller Überlegungen: Wie definiert sich Autorschaft bei modularen Publikationen? Wie kann in modularen und „fluiden“ Formaten punktgenau zitiert werden? Bis auf welche Strukturebene soll verschlagwortet werden? Welche Rollen und Beitragsmodi müssen ausgewiesen werden? Über welche Verfahren kann die Qualitätssicherung erfolgen? Diese Diskussion wird im OES-Kontext verstärkt im Rahmen der in interdisziplinären Forschungsverbünden verorteten OES-Anwendungen wie dem LHTC und dem Organon-Lexikon fortgeführt, die mit „lebhaften“ und modularen Publikationsformen experimentieren. Durch solche Diskussionen und die prak-

tische Ausgestaltung weiterer OES-Anwendungen wird die Entwicklung des OES-Frameworks bzw. -Plugins auch in generischer Sicht vorangetrieben.

Fußnoten

1. <https://tcdh.uni-trier.de/de/newsbeitrag/call-papers-dhd2023>
2. Alternativ werden „Sammelbandbeiträge“ (Kleineberg, Kaden 2017; Flüh 2019), „Sammelwerke“ (Projekt AuROA 2022) oder „Enzyklopädie“ (Kohle 2017) genannt.
3. <https://pkp.sfu.ca/omp/>, <https://pkp.sfu.ca/ojs/>
4. <https://www.open-encyclopedia-system.org/>; Apostolopoulos et al. (2017)
5. OAPEnz wird als Erweiterung des Open-Access-Publikationsportal PUBLISSO entwickelt <https://www.publissso.de/>
6. Im Rahmen des DFG- Projekts "Von 1914-1918-online zum Open Encyclopedia System" (2016-2020). Die Arbeiten bauen auf das Vorgängerprojekt „1914-1918 online: An International Encyclopedia of the First World War“ (DFG, 2012-2015) und dessen konzeptionellen Überlegungen zu Erstellung und Publikation wissenschaftlicher Online-Enzyklopädien auf.
7. Der Quellcode ist unter einer GPLv2-Lizenz auf GitHub publiziert: <https://github.com/open-encyclopedia-system/>
8. Für Zusammenstellung aktueller OES-Anwendungen: <https://www.open-encyclopedia-system.org/use-cases/>. Weitere sind in Vorbereitung bzw. Antragsstellung.
9. <https://wordpress.com/de/> WordPress wird als Open-Source-Software von einer weltweiten Community permanent weiterentwickelt:
10. Mit GPLv2-Lizenz auf GitHub publiziert <https://github.com/open-encyclopedia-system/oes-theme>
11. <https://www.advancedcustomfields.com/>
12. <https://www.zotero.org/>
13. Die Ausführungen nehmen Bezug auf die in AG Digitales Publizieren (2021), Projekt AuROA (2022), Wissenschaftsrat (2022) genannten Anforderungen.
14. Neben den bibliographischen Angaben sind dieses i.d.R. Publikationsdatum, Versionsnummer und der Persistent Identifier (AG Digitales Publizieren 2021, 71).
15. <https://www.compendium-heroicum.de/lemma/erinnerung-gedaechtnis-1-3/>
16. Einen in WordPress integrierten blockbasierten WYSIWYG-Editor, der die Umsetzung modular aufgebauter Lemmata ermöglicht: <https://de.wordpress.org/gutenberg/>
17. <https://de.wordpress.org/plugins/zotpress/>
18. Z.B. unter Verwendung von Trello oder JIRA. Mit dem Artikel-Management-System ist eine OES-spezifische Lösung in Arbeit.
19. Technisch wird dieses über spezifische Custom post types und deren Bündelung in sog. Containern bzw. der Verwendung von (OES-spezifischen) Blöcken im Gutenberg-Editor realisiert.
20. CeMoG-Wissensbasis im Online-Compendium der deutsch-griechischen Verflechtungen (<https://comdeg.eu/wissensbasis/>), Living Handbook of Temporal Communities (<https://www.temporal-communities.de/research/digital-communities/projects/lhct/>), Organon-Lexikon (<https://organon-lexicon.org>)

21. <https://www.compendium-heroicum.de/lemma/helldennarrative/>
22. Eiserner Vorhang: <https://todesopfer.eiserner-vorhang.de/>, B.forscht! <https://projektbrowser.berliner-antike-kolleg.org/>
23. Lobid ist ein Dienst des Hochschulbibliothekszen-trums NRW und bietet u.a. Zugriff auf die GND: <https://lobid.org/gnd>
24. <https://www.geonames.org/>
25. <https://id.loc.gov/authorities/subjects.html>

Bibliographie

- AG Digitales Publizieren (Hg). 2021. „Digitales Publizieren in den Geisteswissenschaften: Begriffe, Standards, Empfehlungen.“ In *Zeitschrift für digitale Geisteswissenschaften / Working Papers*, 1. Wolfenbüttel. text/html Format. DOI: 10.17175/wp_2021_001 (zugegriffen: 25. Juli 2022).
- Apostolopoulos, N., I. Egilmez und C. Schimmel. 2017. „Open Encyclopedia System. Open Source Software for Open Access Encyclopedias“. In *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium for Information Science (ISI 2017)*, hg. von M. Gäde, V. Trkulja, und V. Petras., Berlin, Glückstadt: 380-385. http://isi2017.ib.hu-berlin.de/ISI_17_ONLINE_FINAL.pdf (zugegriffen: 25. Juli 2022).
- Flüh, M. 2019. „Digitales Präsentieren und Publizieren“. In *forTEXT. Literatur digital erforschen*. <https://fortext.net/routinen/methoden/digitales-praesentieren-und-publizieren> (zugegriffen: 25. Juli 2022).
- Kaden, B. 2016. Zur Epistemologie digitaler Methoden in den Geisteswissenschaften. Zenodo. <https://doi.org/10.5281/zenodo.50623> (zugegriffen: 25. Juli 2022).
- Kleineberg, M. und B. Kaden. 2017. „Open Humanities? Expertenmeinungen über Open Access in den Geisteswissenschaften“. In *LIBREAS: Library Ideas*, 32.
- Kohle, H. 2017. „Digitales Publizieren“. In *Digital Humanities. Eine Einführung*, hg. Von F-Jannidis, H. Kohle und M. Rehbein, 199–205. Stuttgart: J.B. Metzler.
- Offen-Definition, Version 2.1. 2022. „Open Knowledge Foundation: Open Definition. Defining Open in Open Data, Open Content and Open Knowledge“. <https://opendefinition.org/od/2.1/de/> (zugegriffen: 1. August 2022).
- Projekt AuROA. 2022. *Publizieren und Open Access in den Geisteswissenschaften: Erkenntnisse aus dem Projekt AuROA zu den Stakeholdern im Publikationsprozess*. Essen. <https://projekt-auroa.de/wp-content/uploads/2022/03/AuROA-Publizieren-und-Open-Access-in-den-Geisteswissenschaften.pdf> (zugegriffen: 25. Juli 2022).
- Wissenschaftsrat (Hg.) 2022. *Empfehlungen zur Transformation des wissenschaftlichen Publizierens zu Open Access*. <https://doi.org/10.57674/fyrc-vb61> (zugegriffen: 25. Juli 2022).

"Mind the Gap": Von Lücken in der Provenienzforschung und ihrer Präsenz im digitalen Raum

Lang, Sabine

sab.lang@fau.de

Friedrich-Alexander-Universität Erlangen-Nürnberg,
Deutschland

Ein Plädoyer für mehr Offenheit

Das Tagungsmotto *Offenheit* hat für die Provenienzforschung¹ eine große Relevanz und bezieht sich unter anderem auf den offenen Umgang mit Lücken in Provenienzanangaben, die z.B. auf eine:n unbekannte:n Besitzer:in hinweisen. Damit Angaben den bekannten Informationsstand widerspiegeln und korrekt bewertet werden, müssen Lücken deutlich gekennzeichnet sein – auch im digitalen Raum. Aufgrund dieser gesteigerten Bedeutung widmet sich der Vortrag der Lücke: Wie werden Provenienzlücken² im digitalen Raum abgebildet und welche weiteren provenienzbezogenen Lücken lassen sich identifizieren? Warum müssen Lücken vor allem im Digitalen thematisiert und gekennzeichnet werden? Dazu werden verschiedene museale Online-Kataloge untersucht, wobei die Provenienzanangaben auf Basis des *Leitfadens zur Standardisierung von Provenienzanangaben* (fortan: *Leitfaden Standardisierung*), herausgegeben vom *Arbeitskreis Provenienzforschung e.V.*, bewertet werden.

Die Auswahl der Kataloge erfolgte anhand folgender Kriterien: eine umfassende und gut aufbereitete Online-Sammlung, (sichtbare) Provenienzforschung am Museum, Beispiele aus verschiedenen deutschen Bundesländern.³ Die ausgewählten Werke zeigen eine Bandbreite, wie Lücken in Provenienzketten abgebildet werden, und stammen von Künstler:innen, die häufig Gegenstand der Provenienzforschung sind. Die Bewertung der Objekteinträge erfolgte anhand folgender Fragestellungen: Sind Provenienzinformationen vorhanden? Folgen sie dem im *Leitfaden Standardisierung* veröffentlichten Standard? Wurden Lücken explizit gekennzeichnet? Sind sie verständlich? Spiegelt der Objekteintrag den Informationsstand wider und verlinkt auf interne und externe Quellen?⁴

Lücken sind ein Bestandteil vieler Wissenschaften: Im Kontext von Archiven und Sammlungen beschäftigen sich Forschende intensiv mit Lücken (Farrenkopf et al. 2021); in der Kunstgeschichte wird das fragmentarische Objekt thematisiert (Schädler-Saub/Weyer 2021) und gefragt, wie man über Objekte schreibt, die abwesend sind

(Fricke/Kumler 2022). In der Provenienzforschung wird auch der Umgang mit und die Bewertung von (Informations-)Lücken im Rahmen von Provenienzforschungen diskutiert (Geldmacher/Kulbe 2022). Die Aufgabe der Vermittlung von Forschungsergebnissen und Provenienzen – auch im digitalen Raum – wird zudem einschlägig besprochen (Türnich 2019). Hierbei findet auch eine Beschäftigung mit musealen Webseiten, im Besonderen mit Online-Sammlungen, und der Darstellung von Provenienzanangaben statt (Haffner 2020; Haffner 2019).⁵ Grundsätzlich ist die digitale Erfassung und online Veröffentlichung von musealen oder universitären Beständen ein wichtiger Forschungsgegenstand, der zudem Fragen nach Zugang und Verhältnis von Original und Digitalisat miteinbezieht (Andraschke/Wagner 2020). Vorhandene Forschungsbeiträge liefern eine wichtige Grundlage für den vorliegenden Beitrag. Eine dezidierte Auseinandersetzung mit Lücken in Online-Katalogen und ihrer Problematik im Digitalen erfolgte bisher aber nicht in ausreichendem Maß. Derzeit widmen sich zudem mehrere Initiativen historischen Forschungsdatenstandards (z.B. *NFDI4Memory*). Damit Lücken in diese Überlegungen mit einbezogen werden, muss eine Auseinandersetzung zu diesem Zeitpunkt stattfinden. Indem der Beitrag Defizite aufzeigt, sollen Bestrebungen hinsichtlich der Entwicklung und Etablierung notwendiger Standards für Provenienzanangaben unterstützt werden. Schließlich fordert der Beitrag zu einer Zusammenarbeit der mit Daten arbeitenden Akteur:innen auf. Der im Oktober 2020 gegründete Verein *Nationale Forschungsdateninfrastruktur (NFDI e.V.)* und die darin agierenden themenübergreifenden Konsortien könnten die dafür notwendigen Plattformen bieten (NFDI).

Die Visualisierung von Lücken

Der *Leitfaden Standardisierung* bietet Richtlinien für die Erstellung von Provenienzanangaben, einschließlich Lücken (vgl. Abb. 1).⁶ Demnach sollen unbekannte Besitzstationen entweder mit [...] oder dem Hinweis *Verbleib unbekannt* gekennzeichnet werden. Hingegen sind einzelne unbekannte Informationen innerhalb einer Besitzstation wie Zeitraum oder Besitzer:in folgendermaßen zu kennzeichnen: Die Abkürzungen *o.D.* (= ohne Datum) oder *o.J.* (= ohne Jahr) markieren eine unbekannte Zeitangabe, *unbekannter Besitzer/Käufer* entsprechend einen unbekannte:n Besitzer:in (Arbeitskreis Provenienzforschung 2018, 11-12, 15-18).⁷ Wurden diese Empfehlungen von Museen umgesetzt?

1919-o.D.	Lyonel Feininger (1871–1956), Weimar [1]
	[...] [2]
spätestens März 1928- vermutlich 1931	Helene (1895–1945) und Hermann Mayer-Freudenberg (1894–1945), Berlin [3]
spätestens 1931- mindestens 06.02.1932/ spätestens Dezember 1932	Maria Miriam Daus, geb. Freudenberg (1907–vermutlich 1995), Berlin [4]
spätestens Dezember 1932–17.01.1933	Galerie Ferdinand Möller, Berlin, wohl in Kommission erhalten von Maria Daus [5]
17.01.1933	Unbekannter Besitzer, angekauft von der Galerie Ferdinand Möller [6]
	[...] [7]
o.D.–08.07.1949	Galerie Franz, Berlin [8]
08.07.1949–1968	Magistrat von Groß-Berlin, Galerie des 20. Jahrhunderts, Berlin, angekauft von der Galerie Franz [9]
seit 1968	Nationalgalerie, Staatliche Museen zu Berlin – Preußischer Kulturbesitz, erhalten als Dauerleihgabe des Landes Berlin [10].

Abb. 1: Screenshot der Provenienzzangabe für Lyonel Feiningers *Kirche von Niedergrunstedt* (1919) (Arbeitskreis Provenienzforschung 2018, 29).

Der Online-Objektkatalog des *Germanischen Nationalmuseums (GNM)* in Nürnberg umfasst derzeit über 168,000 Einträge (GNM a). Eine erste Durchsicht zeigt, dass der Katalog keine Provenienzzangaben für die Objekte bereithält – eine erste signifikante Informationslücke. Ein:e interessierte:r Nutzer:in mag bald auf die Seite des Provenienzforschungsprojekts stoßen, das rund 1,300 Objekte umfasst (GNM b). Die Rechercheergebnisse werden allerdings nicht im Objektkatalog, sondern in einer dazu separaten Datenbank präsentiert, wo für jedes Objekt ein Eintrag mit umfassenden Provenienzzinformationen einschließlich Abbildungen angelegt wurde (GNM c). Abb. 2 zeigt beispielhaft die Provenienzzangabe für *Die Muttergottes mit der Meerkatze* (spätes 16. Jhr.) eines Augsburger Dürernachahmers (GNM d). Erkennbar ist, dass das Museum weitestgehend den Empfehlungen des *Leitfadens Standardisierung* folgt und unbekannte Elemente mit *Unbekannte(r) Vorbesitzer* (anstatt *Unbekannter Besitzer*) und fehlende Besitzstationen mit *Verbleib unbekannt* kennzeichnet (Arbeitskreis Provenienzforschung 2018, 15,18). Das *Städel Museum* in Frankfurt visualisiert letzteres mit ... und weicht dadurch etwas von den im *Leitfaden Standardisierung* alternativ vorgeschlagenen Auslassungspunkten in eckigen Klammern ab (Arbeitskreis Provenienzforschung 2018, 15): Der Eintrag für Max Liebermanns (1847–1935) *Freistunde im Amsterdamer Waisenhaus* (1881–1882) zeigt dies exemplarisch (vgl. Abb. 3) (Städel Museum). Der Eintrag für *Die Muttergottes* enthält zudem einen Link zur entsprechenden Seite im Objektkatalog (GNM d). Dort findet man keine Provenienzzangaben und auch ein Link zur Seite in der Datenbank des Provenienzforschungsprojekts fehlt. Obwohl umfassende Informationen zur Herkunft bekannt sind, zeigt der Objektkatalog eine erhebliche Informationslücke (GNM f).

Datum	Provenienz
spätestens 1905	Johann Nepomuk Sepp, München, erworben von Unbekannte(r) Vorbesitzer [1]
zwischen spätestens 1905 und spätestens 1919	Verbleib unbekannt
spätestens 1919	Joseph Helldörfer, München, erworben von Unbekannte(r) Vorbesitzer [2]
spätestens 17.03.1919	Aloys Laichner, München, wohl erhalten in Kommission von Joseph Helldörfer, München [3]
17.03.1919	Guido von Volckamer, München, erworben durch Kauf von Aloys Laichner, München [4]
Mai 1941	Germanisches Nationalmuseum, erworben im Erbgang von Guido von Volckamer[5]

Die Provenienz für den Zeitraum 1933 bis 1945 ist rekonstruierbar und unbedenklich.

[5]

Abb. 2: Screenshot der Provenienzzangabe für *Die Muttergottes mit der Meerkatze* (spätes 16. Jhr.), Augsburger Dürernachahmer (GNM b; GNM d).

<ul style="list-style-type: none"> Erworben aus der Ausstellung des Pariser Salon von Jean-Baptiste Faure (1830-1914), Paris, Mai 1882 (Nr. 1679) in Kommission an Paul Durand-Ruel, Paris, ca. 1897 verkauft an Paul Cassirer, Berlin, 1899 verkauft an den Städelischen Museums-Verein e.V., Frankfurt am Main, 9. März 1900 Verlust am Auslagerungsort Amorbach, April 1945 ... Rückwerbung durch den Städelischen Museums-Verein e.V., Frankfurt am Main, 1964.
<p>Informationen</p>

Abb. 3: Screenshot der Provenienzzangabe für Max Liebermanns *Freistunde im Amsterdamer Waisenhaus* (1881–1882) (Städel Museum).

Die Sammlung Online des *Sprengel Museums* in Hannover hält mehr als 19,000 Einträge bereit, darunter auch Max Beckmanns (1884–1950) *Stilleben [sic.] mit schiefer Schnapsflasche und Buddha* (1939) (Sprengel Museum). Im Vergleich zu anderen Beispielen ist die Provenienzzangabe für Beckmanns Stilleben sehr unübersichtlich und schwer verständlich (vgl. Abb. 4). Auffällig ist die Abwesenheit von Lücken; ob für das Werk tatsächlich eine vollständige Biografie vorliegt oder ob fehlende Informationen nur nicht abgebildet werden, ist schwer zu sagen. Sowohl Format als auch bereitgestellte Informationen entsprechen allerdings nicht dem Standard, der im *Leitfaden Standardisierung* vorgeschlagen wird (vgl. Abb. 1).

Weitaus transparenter ist das *Museum Folkwang* in Essen, das neben genauen Angaben zur Provenienz einzelner Objekte auch Lücken deutlich kennzeichnet (Haffner 2019, 93). Die Sammlung Online hält aktuell über 93,000 Werke bereit, darunter auch Lyonel Feiningers (1871–1956) *Leuchtbarke I* (um 1913) (Museum Folkwang a; Museum Folkwang b). Abb. 5 zeigt die Provenienzzangabe für das Ölgemälde, deutlich erkennbar sind Lücken vor 1930 und zwischen 1930 und 1951, visualisiert durch einen Platzhalter in Form von eckigen Klammern und Auslassungspunkten. Die Darstellung weicht damit auch von den Empfehlungen des *Leitfadens Standardisierung* ab. Hierin werden die beschriebenen Platzhalter für unbekannte Besitzstationen und die Abkürzungen

o.D. oder o.J. (anstatt [...]–[...] und [...]–04.1951) für unbekannte Zeitangaben vorgeschlagen (Arbeitskreis Provenienzforschung 2018, 15–17). Es sei zudem auf die zusätzlichen Vermerke und das Ampelsymbol hingewiesen.⁸

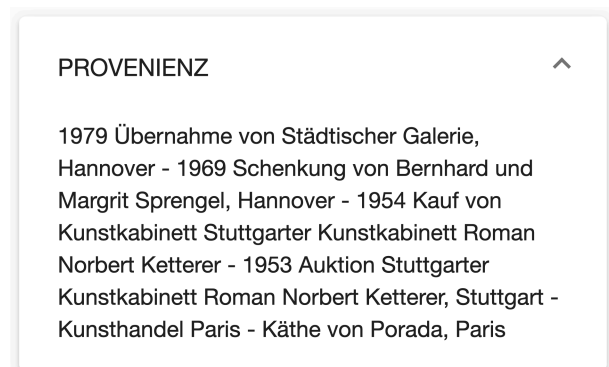


Abb. 4: Screenshot der Provenienzanzeige für Max Beckmanns *Stilben* [sic.] mit schiefer Schnapsflasche und Buddha (1939) (Sprengel Museum).

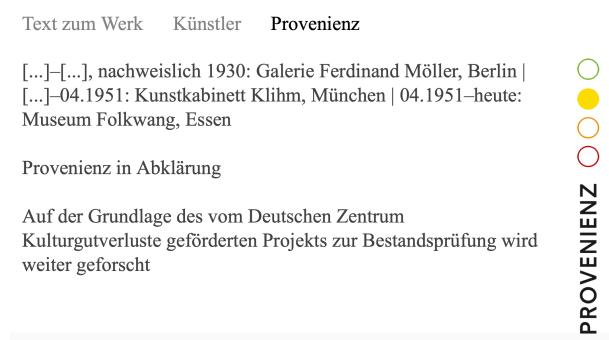


Abb. 5: Screenshot der Provenienzanzeige für Lyonel Feingolds *Leuchtbarke I* (um 1913) (Museum Folkwang).

Seit September 2022 enthält die Online-Sammlung der *Bayerischen Staatsgemäldesammlungen* (BStGS) ausführliche Provenienzinformationen für über 1,200 Werke, weitere Einstellungen sollen folgen (BStGS a; BStGS b). Davor waren die Angaben zur Herkunft der Werke nur sehr rudimentär. Am Beispiel des Gemäldes *Die Falknerin* (um 1880) von Hans Makart (1840–1884) wird dies deutlich. Abb. 6 zeigt den Eintrag in der Online-Sammlung vor der Aktualisierung: Er enthält eine Beschreibung des Werkes und weitere Objektinformationen. Ein Provenienzfeld fehlt, einzig ein Herkunftsvermerk informiert, dass das Werk „1962 als Überweisung aus Staatsbesitz erworben [wurde].“ Dass tatsächlich sehr viel mehr über die Provenienz des Werkes bekannt ist, war nicht ersichtlich. Darüber erfuhr man zum Beispiel in der *Lost Art*⁹ Datenbank, in welcher das Gemälde seit 2007 als Fundmeldung gelistet ist (vgl. Abb. 7) (Lost Art b). Diese Information sowie ein Link zu *Lost Art* fehlte in der Online-Sammlung der BStGS. Der neue Eintrag zeigt diese Lücken nicht mehr: Nutzer:innen finden nun umfangreiche Provenienzinformationen und einen Link zu *Lost Art* (vgl. Abb. 8 und 9) (BStGS c). Entsprechend der im Leitfa-

gaben mit o.D. gekennzeichnet (Arbeitskreis Provenienzforschung 2018, 16–17). Problematisch ist, dass obwohl einige Übergänge als unsicher einzustufen sind, mögliche Lücken zwischen den Besitzstationen nicht explizit sichtbar sind.

Die Falknerin

Künstler Hans Makart	Datierung um 1880	Ausgestellt Nicht ausgestellt	Inventarnummer 13291
Mehr über das Werk Die Falknerin ist in gründerzeitlicher „Nahbildlichkeit“ wiedergegeben. Tief liegender Horizont und Himmel dienen als bedeutungsleitender Hintergrund und verstärken den Eindruck von Größe und Einmaligkeit der Erscheinung. Die Genauigkeit der Wiedergabe des Stoffs und der Haut bewirkt darüber hinaus die Vorstellung erhöhter Präsenz. – Die Falkenjagd auf Feder- und kleineres Haarwild war im Mittelalter sehr beliebt. Gegen Ende des 19. und im 20. Jahrhundert wurde diese Jagd wieder aufgenommen und übte zunehmende Anziehungskraft aus.	Geburtsjahr des Künstlers 1840	Maße des Objekts 106,3 x 79,8 cm	Sterbejahr des Künstlers 1884
	Gattung Malerei		Material / Technik / Bildträger Öl auf Leinwand
	Herkunft 1962 als Überweisung aus Staatsbesitz erworben		Referat 19. Jahrhundert
	Bestand Bayerische Staatsgemäldesammlungen - Neue Pinakothek München		
	Zitervorschlag Hans Makart, Die Falknerin, um 1880, Bayerische Staatsgemäldesammlungen - Neue Pinakothek München, URL: http://www.sammlung.pinakothek.de/de/artwork/psaepVGJ7 (Zuletzt aktualisiert am 22.04.2022)		

Abb. 6: Screenshot des Eintrags für Hans Makarts *Die Falknerin* (um 1880). Alte Version. Online nicht mehr verfügbar. Vgl. (BStGS c).

Provenienz:

Bis 1887 Sammlung Mihail Kogălnicanu, Rumänien. - Am 9./10. Dezember 1887 Versteigerung bei Heberle, Köln. - Privatbesitz Wien und Berlin. - 1908 H. O. Miethe in Wien. - 1909 Privatbesitz, Wien. - Zu unbestimmtem Zeitpunkt vor 1934 möglicherweise in der Sammlung des Großindustriellen Moritz von Guttman, Berlin. - 1937 Bad Vöslau, Privatbesitz. - Am 3.12.1937 von der Firma Neumann und Salzer, Wien, an die Galerie Haberkamp, Berlin. - Am 9.2.1938 von dort für RM 13.400, an die Reichskanzlei, Berlin. - Am 12.1.1938 von Adolf Hitler als Geschenk an Hermann Göring. - Inventar Reichsmarschall, RM-Nr. 58. - CCP München, Münchner Nr. 5436. - Am 18.5.1961 von der Treuhandverwaltung an die Direktion der Bayerischen Staatsgemäldesammlungen für den Freistaat Bayern überwiesen, Nr. 47 der Übertragungsliste.

Abb. 7: Screenshot der Provenienz für Hans Makarts *Die Falknerin* in *Lost Art* (Lost Art b).

Die Beispiele zeigen, dass deutsche Museen teilweise Provenienzanzeigen zu den Objekten bereitstellen (Haffner 2019, 93), Umfang und Darstellungsform aber stark variieren, letzteres spiegelt sich auch in der Kennzeichnung von Lücken wider. Mögliche Gründe sind die Heterogenität der Sammlungen oder technische Limitationen der Museumsmanagement-Systeme. Diese sind für eine allgemeine Bestandsinventarisierung ausgelegt und nicht für die umfassende und strukturierte Erfassung von Provenienzanzeigen (Haffner 2019, 95–96). Auch rechtliche Restriktionen und eine allgemeine Besorgnis der Museen sind weitere mögliche Gründe (Haffner 2020, 38). Viele Museen verwenden noch Freitextfelder für Provenienzanzeigen, was zwar die Eingabe von unsicheren oder fehlenden Infos vereinfacht, aber eine Maschinenlesbarkeit erschwert. Letzteres wird im Moment z.B. durch das *Provenance Lab* an der Leuphana Universität durch die computergestützte Produktion von *Provenance Linked Open Data* adressiert, wobei u.a. die Dokumentation von Unvollständigkeit berücksichtigt wird (Libeskind 2022, 23–25). Auch die Entwicklung von Normdaten für die Provenienzforschung ist in diesem Zusammenhang zu nennen: Ein Projekt innerhalb des Konsortiums *NFDI4Objects* soll sich der Erstellung von provenienzbegleitenden Personen-Normdaten (NFDI4Objects) widmen. Dies lässt hoffen, dass die Etablierung eines Standards für Provenienzen in Online-Sammlungen in nicht allzu ferner Zukunft liegt. Standards würden nicht nur zur Sichtbarkeit von Lücken beitragen, sondern auch eine maschinelle Verarbeitung und Verknüpfung der Da-

tensätze ermöglichen und damit auch neue Forschungsfragen generieren, z.B. ob es Muster hinsichtlich der Provenienzlücken gibt.

Die angeführten Beispiele haben weitere provenienzbezogene Lücken aufgezeigt, die über die Provenienzanangaben selbst hinausgehen: Zum einen gibt es eine Diskrepanz zwischen den Angaben in den Objekteinträgen und den tatsächlich bekannten Provenienzinformatoren, die oft auf separaten Seiten abgelegt sind. Auch einfache Verknüpfungen der Seiten fehlen in den digitalen Sammlungen der Museen. Bei der Analyse verschiedener musealer Online-Kataloge wurde zudem sichtbar, dass in der englischen Version die Provenienzanangaben oder Teile der Objektinformationen weiterhin auf Deutsch erscheinen.¹⁰ Die Datenbank des Provenienzforschungsprojekts am GNM fehlt in der englischen Version sogar völlig (GNM e). Dies ist für internationale Provenienzforschende und die ursprünglichen Eigentümer:innen der Werke und deren Nachkommen äußerst problematisch, da sie oftmals kein Deutsch sprechen.

HANS MAKART (1840-1884) Die Falknerin, um 1880

Material / Technik / Bildträger Öl auf Leinwand	Maße des Objekts 106,3 x 79,8 cm	Ausgestellt Nicht ausgestellt
Referat 19. Jahrhundert	Gattung Malerei	Inventarnummer 13291
Erwerb 1962 als Überweisung aus Staatsbesitz erworben	Bestand Bayerische Staatsgemäldesammlungen - Neue Pinakothek München	

Zitiervorschlag
Hans Makart, Die Falknerin, um 1880, Bayerische Staatsgemäldesammlungen - Neue Pinakothek München, URL:
<http://www.sammlung.pinakothek.de/de/artwork/jpxegrVGJ7> (Zuletzt aktualisiert am 22.04.2022)

ÜBER DAS WERK

PROVENIENZFORSCHUNG

Abb. 8: Screenshot des Eintrags für Hans Makarts *Die Falknerin* (um 1880). Aktualisierte Version (BStGS c).

PROVENIENZFORSCHUNG

Seit 1999 forscht das Referat Provenienzforschung zur Herkunft aller Kunstwerke der Bayerischen Staatsgemäldesammlungen, die vor 1945 entstanden sind und die seit 1933 erworben wurden. Grundlage für ...
[Mehr erfahren](#)

Projekt: Kunstwerke aus ehem. NS-Besitz

Mihail Kogălniceanu (1817 - 1891), Bukarest, eingeliefert in Auktion bei J. M. Heberle (H. Lempertz' Söhne), Köln
o.D. - 09./10.12.1887

wohl Kunsthandlung Neumann & Salzer, Wien, wohl erworben auf der Auktion bei J. M. Heberle (H. Lempertz' Söhne), Köln
1887 - evtl. 1891

wohl Privatbesitz, Berlin u. Wien
evtl. 1891 - o.D.

Galerie H. O. Miethke, Wien
o.D. - 1908/09

wohl Privatbesitz, Wien, evtl. erworben von Galerie H. O. Miethke, Wien
1908/09 - o.D.

Moritz von Gutmann (1872 - 1934) und Erben, Schloss Vöslau bei Wien
o.D. (evtl. ab 1908/09) - längstens 1937

Galerie L. T. Neumann, Wien
spätestens 1937 - 03.12.1937

Kunsthandlung Karl Haberstock, Berlin, erworben von Galerie L. T. Neumann, Wien
03.12.1937 - 12.01./09.02.1938

Abb. 9: Screenshot der Provenienz für Hans Makarts *Die Falknerin* (um 1880), Auszug. Aktualisierte Version (BStGS c).

Schlussbemerkung: "Mind the Gap"

Warum müssen Lücken vor allem im Kontext des Digitalen besprochen und gekennzeichnet werden? Die Auseinandersetzung mit Lücken ist auch im Analogen unerlässlich, da es auch hier zu Fehlinterpretationen kommen kann. Im digitalen Raum scheint dieser Anspruch und die damit verbundene Sichtbarmachung von Lücken aber noch gesteigert: Aufgrund der Informationsmenge kann der Eindruck von Vollständigkeit geweckt werden. Etwaige Lücken können einfach durch Verknüpfungen mit externen Datensätzen geschlossen werden. Die konstante Datenproduktion erzeugt zudem eine hohe Dynamik und betont den Aspekt der Flüchtigkeit im digitalen Raum: Vorhandene Bild- oder Textdaten werden ständig durch neue oder ergänzende Daten überlagert; bestehende Lücken werden damit verwischt oder treten erst gar nicht in Erscheinung. Nutzende können zudem schnell zwischen einzelnen Seiten wechseln, was die Gefahr birgt, dass Inhalte nur selektiv wahrgenommen und Lücken deshalb schlichtweg übersehen werden. Aufgrund der Menge an Daten und der Flüchtigkeit des digitalen Raumes gilt "Mind the Gap"¹¹.

Fußnoten

1. Provenienzforschung will die Herkunft und Besitzverhältnisse von Objekten klären und prüft, unter welchen Bedingungen ein Besitzwechsel stattgefunden hat. Sie überprüft unter anderem NS-verfolgungsbedingt entzogenes Kulturgut, Kulturgutentziehungen in Sowjetischer Besatzungszone (SBZ) und DDR oder Sammlungsgut

aus kolonialen Kontexten (Haase/Hopp 2022, 6; Deutsches Zentrum Kulturgutverluste et al. 2019, 6).

2. Der Begriff *Provenienz* bezieht sich hier auf die Herkunft des realen Objekts. Im Kontext des digitalen Raumes kann er sich auch auf die Provenienz digitaler Daten beziehen, die z.B. Informationen über den Ersteller:in der Datei und weitere Nutzer:innen bereithält.

3. Aus Platzgründen konnten einige wichtige Museen wie die Staatlichen Kunstsammlungen Dresden nicht berücksichtigt werden.

4. Obwohl eine stichprobenartige Analyse keine Evidenz generiert, kann sie Tendenzen sichtbar machen und aufzeigen, dass fehlende Standards für die Darstellung von Provenienzlücken und andere digitale Lücken problematisch für die "richtige" Bewertung der Objekte sind.

5. Die Provenienzforschung beschäftigt sich zudem mit den Möglichkeiten des Digitalen und insbesondere mit digitalen Methoden (Fuhrmeister/Hopp 2019) und benennt konkrete Tendenzen, Desiderate und Bedürfnisse (Hopp 2018). Eine gesteigerte Auseinandersetzung mit digitalen Möglichkeiten zeigt sich unter anderem in der Zeitschrift *Archivar* mit dem Themenschwerpunkt Provenienzforschung (Landesarchiv Nordrhein-Westfalen 2022) und dem der Digitalen Provenienzforschung gewidmeten Heft *Provenienz & Forschung* (Deutsches Zentrum Kulturgutverluste 2020).

6. Für die Erfassung und Publikation von Provenienzen siehe auch das *LIDO-Handbuch* (Knaus et al. 2022).

7. Für einen vollständigen Überblick über die Kennzeichnung von Lücken in Provenienzangaben vgl. (Arbeitskreis Provenienzforschung 2018, 15-27).

8. Für eine ausführliche Erklärung des Ampelsymbols vgl. (Deutsches Zentrum Kulturgutverluste et al. 2019, 35, 89-90).

9. Die *Lost Art* Datenbank beinhaltet Such- und Fundmeldungen zu Kulturgütern, die im Zuge des Nationalsozialismus oder Zweiten Weltkriegs verlagert oder ihren Eigentümer:innen entzogen wurden (Lost Art a).

10. Beispielhaft sind die *BStGS* oder das *Städel Museum* (BStGS c; Städel Museum).

11. Der Titel wurde in Anlehnung an die Online-Ausstellung *Mind the gap. Von geraubten Büchern, fairen Lösungen ... und Lücken der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB)* gewählt, vgl. (SLUB).

Bibliographie

Andraschke, Udo und Sarah Wagner, Hg. 2020. *Objekte im Netz. Wissenschaftliche Sammlungen im digitalen Wandel*. Bielefeld: transcript Verlag. DOI: 10.14361/9783839455715.

Arbeitskreis Provenienzforschung e.V., Hg. 2018. *Leitfaden zur Standardisierung von Provenienzangaben*. Erarbeitet von Claudia Andrasschke, Jasmin Hartmann, Johanna Poltermann, Birgitte Reuter, Iris Schmeisser, Wolfgang Schödert. Hamburg. https://wissenschaftliche-sammlungen.de/files/4515/2585/6130/Leitfaden_APF_eV_online.pdf (zugegriffen: 26. Juli 2022).

BStGS a, Bayerische Staatsgemäldesammlungen, München, Provenienzforschung, [\[de/forschung/provenienzforschung\]\(https://www.pinakothek.de/forschung/provenienzforschung\) \(zugegriffen: 20. November 2022\).](https://www.pinakothek-</p>
</div>
<div data-bbox=)

BStGS b, Bayerische Staatsgemäldesammlungen, München, Provenienzforschung, Objektkatalog, <https://www.sammlung.pinakothek.de/de/provenienz-online> (zugegriffen: 20. November 2022).

BStGS c, Bayerische Staatsgemäldesammlungen, München, Online-Sammlung, Objekt, <http://www.sammlung.pinakothek.de/de/artwork/jpxegrVGJ7> (zugegriffen: 24. November 2022).

Deutsches Zentrum Kulturgutverluste, Hg. 2020. *Provenienz & Forschung. Digitale Provenienzforschung*. Dresden: Sandstein Verlag. https://www.sandstein.de/verlag/provenienz_1-2020.php (zugegriffen: 28. Juli 2022).

Deutsches Zentrum Kulturgutverluste et al., Hg. 2019. *Leitfaden Provenienzforschung*. Berlin: Königsdruck Printmedien und digitale Dienste GmbH. https://www.kulturgutverluste.de/Content/03_Recherche/DE/Leitfaden-Download.pdf (zugegriffen: 23. Juli 2022).

Farrenkopf, Michael, Andreas Ludwig und Achim Saupe, Hg. 2021. *Logik und Lücke: Die Konstruktion des Authentischen in Archiven und Sammlungen (Wert der Vergangenheit)*. Göttingen: Wallstein Verlag.

Fricke, Beate und Aden Kumler, Hg. 2022. *Destroyed - Disappeared - Lost - Never Were*. University Park: Pennsylvania: Penn State University Press.

Fuhrmeister, Christian und Meike Hopp. 2019. "Rethinking Provenance Research." *Getty Research Journal* 11: 213-231. <https://doi.org/10.1086/702755> (zugegriffen: 13. Juli 2022).

Geldmacher, Elisabeth und Nadine Kulbe. 2022. "Unvermeidbar! Über Lücken in der NS-Raubgut-Forschung und Möglichkeiten, mit ihnen umzugehen." *Archivar* 1: 31-34. https://www.archive.nrw.de/sites/default/files/media/files/Archivar_2022-1_Internet-NEU-28032022_Mod.pdf (zugegriffen: 23. Juli 2022).

GNM a, Germanisches Nationalmuseum, Nürnberg, Objektkatalog, <https://objektkatalog.gnm.de> (zugegriffen: 28. November 2022).

GNM b, Germanisches Nationalmuseum, Nürnberg, Provenienzforschungsprojekt, <https://provenienz.gnm.de> (zugegriffen: 26. November 2022).

GNM c, Germanisches Nationalmuseum, Nürnberg, Provenienzforschungsprojekt, Objektkatalog, https://provenienz.gnm.de/wisski_views/b661085ac1552c83ff3c2b8c56b693fc (zugegriffen: 26. November 2022).

GNM d, Germanisches Nationalmuseum, Nürnberg, Provenienzforschungsprojekt, Objektkatalog, Objekt: Gm1625, <https://provenienz.gnm.de/objekt/Gm1625> (zugegriffen: 26. November 2022).

GNM e, Germanisches Nationalmuseum, Nürnberg, englische Version, <https://www.gnm.de/your-museum-in-nuremberg/museum/> (zugegriffen: 26. November 2022).

GNM f, Germanisches Nationalmuseum, Nürnberg, Objektkatalog, Objekt: Gm1625, <http://objektkatalog.gnm.de/objekt/Gm1625> (zugegriffen: 28. November 2022).

Haase, Sven und Maike Hopp. 2022. "Einführung in den Themenschwerpunkt." *Archivar* 1: 6-10. https://www.archive.nrw.de/sites/default/files/media/files/Archivar_2022-1_Internet-NEU-28032022_Mod.pdf (zugegriffen: 23. Juli 2022).

Haffner, Dorothee. 2020. "Provenienzen in Sammlungsdatenbanken. Digitale und virtuelle Chancen für die Vermittlung." In *Provenienz & Forschung. Digitale Provenienzforschung*, hg. vom Deutschen Zentrum Kulturgutverluste, 36-42. Dresden: Sandstein Verlag.

Haffner, Dorothee. 2019. "Provenienzforschung digital vernetzt. Ergebnisse sichtbar machen." *Museumskunde* 84: 90-97. <https://www.museumsbund.de/wp-content/uploads/2022/07/museumskunde-2019-1-online.pdf> (zugegriffen: 12. Juli 2022).

Hopp, Meike. 2018. "Provenienzforschung und digitale Forschungsinfrastrukturen in Deutschland: Tendenzen, Desiderate, Bedürfnisse." In *...kein Ende in Sicht. 20 Jahre Kunstrückgabegesetz in Österreich. Schriftenreihe der Kommission für Provenienzforschung Band 8*, hg. von Eva Blimlinger und Heinz Schödl, 35-59. Wien, Köln: Böhlau Verlag. <https://doi.org/10.7767/9783205201274.37> (zugegriffen: 12. Juli 2022).

Knaus, Gudrun, Regine Stein und Angela Kailus. 2022. *LIDO-Handbuch für die Erfassung und Publikation von Metadaten zu kulturellen Objekten: Band 2: Malerei und Skulptur*, hg. vom Deutschen Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg und Christian Bracht. Heidelberg: arthistoricum.net. <https://doi.org/10.11588/arthistoricum.1026> (zugegriffen: 26. Juli 2022).

Landesarchiv Nordrhein-Westfalen und Verband deutscher Archivarinnen und Archivare e.V., Hg. 2022. *Archivar, Provenienzforschung*. Siegburg: Verlag Franz Schmitt. https://www.archive.nrw.de/sites/default/files/media/files/Archivar_2022-1_Internet-NEU-28032022_Mod.pdf (zugegriffen: 28. Juli 2022).

Libeskind, Daniel. 2022. "A New Approach to Provenance: The 'Provenance Studies' Program at Leuphana University Lüneburg." *Newsletter of the Network European Restitution Committees on Nazi-looted Art* 13: 23-25. <https://www.restitutiecommissie.nl/wp-content/uploads/2022/05/Network-Newsletter-no.13-May2022.pdf> (zugegriffen: 19. November 2022).

Lost Art a, Startseite, <https://www.lostart.de/de/start> (zugegriffen: 26. November 2022).

Lost Art b, Fundmeldung, Einzelobjekt, <https://www.lostart.de/de/Fund/391081> (zugegriffen: 26. November 2022).

Museum Folkwang a, Essen, Sammlung Online, <https://www.museum-folkwang.de/de/sammlung-online> (zugegriffen: 28. November 2022).

Museum Folkwang b, Sammlung Online, Objekt, <http://sammlung-online.museum-folkwang.de/eMP/eMuseumPlus?service=ExternalInterface&module=collection&objectId=3278&viewType=detailView> (zugegriffen: 26. November 2022).

NFDI, Verein, <https://www.nfdi.de/verein/#historie> (zugegriffen: 28. November 2022).

NFDI4Objects, TRAILS, <https://www.nfdi4objects.net/index.php/arbeitsprogramm/trails> (zugegriffen: 24. November 2022).

Schädler-Saub, Ursula und Angela Weyer. 2021. *Das Fragment im digitalen Zeitalter: Möglichkeiten und Grenzen neuer Techniken in der Restaurierung*. Berlin: Hendrik Bäßler Verlag.

SLUB, Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden, Online-Ausstellung, *Mind the gap. Von geraubten Büchern, fairen Lösungen ... und Lücken*, <https://ausstellungen.deutsche-digitale-bibliothek.de/mind-the-gap/> (zugegriffen: 26. November 2022).

Sprengel Museum, Hannover, Sammlung online, Objekt ID: 2328, <https://sprengel.hannover-stadt.de/home> (zugegriffen: 26. November 2022).

Städel Museum, Frankfurt, Digitale Sammlung, Objekt <https://www.staedelmuseum.de/go/ds/1351> (zugegriffen: 28. November 2022).

Türnich, Ruth. 2019. "Provenienzforschung weiterdenken. Vermittlung von Provenienzforschungen und Forschungsergebnissen." In *rhein-form. Information für die rheinischen Museen 2*, hg. vom LVR-Dezernat Kultur und Landschaftliche Kulturpflege, 19-25. Köln: LVR-Druckerei. https://rheinform.lvr.de/media/medienrhein-form/archiv/rheinform_2-2019.pdf (zugegriffen: 14. Juli 2022).

Washingtoner Prinzipien. 1998. "Grundsätze der Washingtoner Konferenz in Bezug auf Kunstwerke, die von den Nationalsozialisten beschlagnahmt wurden (Washingtoner Prinzipien)." <https://www.kulturgutverluste.de/Webs/DE/Stiftung/Grundlagen/Washingtoner-Prinzipien/Index.html> (zugegriffen: 10. Juli 2022).

Minimal Editing: Die Hyperdiplomatische Transkriptionsplattform

Galka, Selina

selina.galka@uni-graz.at
Karl-Franzens-Universität Graz, Österreich

Klug, Helmut W.

helmutwklug@gmail.com
Karl-Franzens-Universität Graz, Österreich

Minimal Editing

Digitale Editionen sind in der Regel kostspielige Großprojekte, die von Drittmittelförderung abhängig sind. Projektmitarbeiterinnen und -mitarbeiter müssen umfangreiche technische und philologische Fähigkeiten mitbringen und die Projekte werden meist in sehr spezifischen Editionsumgebungen entwickelt – Elena Pierazzo nennt derartige digitale Editionen "Haute Couture"-Editionen (Pierazzo 2019, 213).

Kleinere Editionsprojekte werden oft nur von Einzelpersonen gestemmt oder vielleicht im Kontext einer Lehrveranstaltung oder Qualifikationsarbeit umgesetzt; eine Vermittlung der Ergebnisse in Form eines Webauftritts oder die Archivierung in einem zertifizierten Repository ist in derartigen Fällen nur selten möglich; oft

scheitert es am entsprechenden Know-How oder an finanzieller Unterstützung. Diese Arbeiten verschwinden eher früher als später ungenutzt in dunklen Schubladen. Um dieser Problematik Herr zu werden, müssen nachhaltige Lösungen her! Eine solche Lösung für derartige Fälle wäre sicher *minimal editing*.

Minimal Editions stellen reduzierte digitale Editionen im Gegenzug zu kostspieligen, aufwendigen Editionsprojekten dar. Wie eine derartige Minimaledition zu definieren wäre, ist Gegenstand des aktuellen Forschungsdiskurses: Basiert sie auf dem effizientesten Workflow, um den Arbeitsaufwand der Datenverarbeitung zu reduzieren, der kleinstmöglichen technischen Infrastruktur (i.e. *minimal computing*, cf. Gil 2015), um die Hemmschwelle vor der Technik möglichst niedrig zu halten, der sparsamsten Modellierung der Objekte, um Bearbeitungszeit zu sparen und die weitere Verarbeitung zu vereinfachen oder zeichnet sie sich durch einen reduzierten Funktionsumfang bei der Onlinepublikation aus?

Minimal editing kann es möglich machen, digitale Editionen mit vorgefertigten Workflows, bereits etablierten, einfachen technischen Lösungen und leicht zu integrierenden Basis-Funktionalitäten im Bereich der Datenmodellierung wie auch der Datenpräsentation schnell, einfach und mit einem geringen Kostenaufwand umzusetzen. Das bringt allerdings die unterschiedlichsten Einschränkungen mit sich! Demgegenüber stehen aber die Erwartungen und Ideen der Editorinnen und Editoren und die Vorgaben der Förderorganisationen. Allerdings würde der Aufbau einer zeitgemäßen Infrastruktur, die Langzeitverfügbarkeit garantiert, mit Komponenten, die *ad hoc* für unterschiedliche Editionsprojekte benutzbar sind, den Aufwand für die Umsetzung einer digitalen Edition reduzieren und somit vermutlich zu einer größeren Anzahl von digitalen Editionsprodukten führen. Pierazzo bezeichnet diese Editionen als "Prêt-à-Porter"-Editionen (Pierazzo 2019, 213).

Weitere Stichworte im Zusammenhang mit *minimal editing* sind *minimal functionalities* und *minimal computing*. Federico Caria und Brigitte Mathiak untersuchten, welche minimalen Funktionalitäten der Webaufruf einer digitalen Edition unbedingt aufweisen sollte, damit diese von der Community auch zufriedenstellend genutzt werden kann (2018); Greta Franzini, Melissa Terras und Simon Mahony führten ebenfalls eine Umfrage zu Erwartungen und dem Gebrauch von digitalen Editionen durch (2019). Die dabei gesammelten Ergebnisse können helfen, die wichtigsten Komponenten einer *Minimal Edition* zu definieren, wie z. B. eine Such- und Exportfunktion über bzw. in den Daten oder eine grundsätzlich benutzerfreundliche Navigationsstruktur (Caria/Mathiak 2018, 360), oder die Einbindung von Faksimiles, leicht erkennbare Angabe von Lizenzen, die Möglichkeit der Nutzbarkeit der Daten und eine umfassende Dokumentation (Franzini/Terras/Mahony 2018, 1).

Minimal computing hingegen ist ein Ansatz, bei dem versucht wird, die technischen Ansprüche zu reduzieren und für die Arbeit nur jene Technologien zu verwenden, welche tatsächlich notwendig für die Umsetzung eines bestimmten Vorhabens sind. Erst 2022 erschien eine ganze Ausgabe des *Digital Humanities Quarterly*, welche sich den unterschiedlichsten Aspekten des *minimal computing* widmet (Risam/Gil 2022a). Um zu bestimmen, welche Technologien tatsächlich notwendig und

ausreichend sind, schlagen Risam und Gil vor, folgende vier Fragen zu stellen (Risam/Gil 2022b): Was brauchen wir, was haben wir, was müssen wir priorisieren und was sind wir bereit aufzugeben?

Alle diese Publikationen zeigen einen Trend in Richtung maßgeschneidertes Editionsmanagement auf. Es gibt bereits Lösungen, Tools und Plattformen, die unter dem Aspekt Minimal Editing zusammenzufassen sind – darunter fällt z.B. EVT (Edition Visualization Technique), ein Open-Source-Tool, welches es ermöglicht, XML-TEI Dokumente als digitale Editionen im Browser anzuzeigen. Dasselbe gilt für CETIclean und TEI-Publisher, welche ebenfalls Lösungen anbieten, um XML- und TEI-Dokumente ohne Erfahrung in der Webprogrammierung im Browser zugänglich zu machen. Transcriptions erlaubt es, Transkriptionen zu teilen und kollaborativ zu überarbeiten, wobei hier jedoch kein Fokus auf die Modellierung mit XML/TEI oder eine nachhaltige Archivierung gelegt wird. Den Tools, die eine Kodierung mit TEI unterstützen, ist gemeinsam, dass zwar niederschwellig eine Webansicht generiert werden kann, allerdings wird hier der Aspekt der Langzeitarchivierung, d.h. dass die Daten permanent zitierbar verfügbar bleiben, nicht berücksichtigt.

Ein Zugang zur "Minimal Edition", der sich am Paradigma eines effizient gestalteten Workflows orientiert und den Aspekt der Langzeitarchivierung miteinschließt, ist in Österreich im Rahmen des bundesgeförderten Projektes DiTAH - *Digitale Transformation der österreichischen Geisteswissenschaften* umgesetzt worden.

Hyper: Hyperdiplomatische Transkriptionsplattform

Im geplanten Vortrag soll eine Variante des *minimal editing* anhand der hyperdiplomatischen Publikationsplattform vorgestellt werden: Die Hyperdiplomatische Transkriptionsplattform (Hyper, <http://gams.unigraz.at/context:hyper>) wurde entwickelt, um das Archivieren und Publizieren von Editionsdaten, die auf dem standardisierten Datenmodell der hyperdiplomatischen Transkription (Böhm/Klug 2021) beruhen, möglichst ressourcenschonend und im Sinne von *minimal editing* zu gestalten – mit den Ressourcen, die im vorliegenden Fall möglichst sparsam verwendet werden sollen, sind vor allem die nötigen Arbeiten und der finanzielle Aufwand im Zusammenhang mit einer Publikation und Langzeitarchivierung der Editionsdaten gemeint.

Eine hyperdiplomatische Transkription versucht, die historische Quelle möglichst detailreich bis hin zur Teilzeichenebene (z. B. Superskripte oder sogar Teilstiche eines Zeichens) bzw. unter Berücksichtigung der Quellentopographie (Verortung der Informationseinheiten in einem digitalen Abbild der Quelle) in ein modernes Zeichensystem zu übertragen. In Grazer Projekten (*Mittelalterlabor*, *Cooking recipes of the Middle Ages*) wird nach einer hyperdiplomatischen Transkriptionsmethode gearbeitet, die ursprünglich auf die Philosophie der "Grazer dynamischen Editions Methode" (Hofmeister-Winter 2003) zurückgeht: Es wird dabei eine "Basis-transliteration" (Hofmeister-Winter 2003, S. 101) erstellt, deren Ziel es ist, den Text der Handschrift möglichst getreu mit typographischen Mitteln abzubilden, wobei

die Quellentopographie dabei keine bzw. nur bedingt eine Rolle spielt. In dieser Phase soll noch keine inhaltliche Interpretation stattfinden, sondern vorerst nur der paläografische Informationsgehalt der Handschrift, also die Schriftsymbole wiedergegeben werden. Es werden in einem Graphinventar nicht nur alphabetische Schriftsymbole, sondern alle graphischen Phänomene wie Abkürzungen, Superskripte oder Verzierungen mit bestimmten von den Transkribierenden individuell festgelegbaren Kodierungen festgehalten. Im Graphinventar werden die von der Schreiberin oder dem Schreiber verwendeten Zeichen beschrieben und mit einer proprietären Kodierung versehen, sodass es gleichzeitig als Transliterationsschlüssel dient, der im Rahmen der digitalen Workflows eine zentrale Stellung einnimmt. Die Transkription soll damit sowohl sprach- und literaturwissenschaftlichen als auch geschichtswissenschaftlichen Ansprüchen genügen bzw. sie in manchen Fällen übertreffen – das Ziel dieser Herangehensweise ist es aber prinzipiell, eine Transkription zu erstellen, die über den konkreten Anwendungsfall hinaus auch von anderen Forscherinnen und Forschern verwendet werden kann (Böhm/Klug 2020).

Die Hyperdiplomatische Transkriptionsplattform bietet daher ein vorgefertigtes Datenmodell für die hyperdiplomatische Transkription (Klug/Böhm 2021), vorgefertigte Routinen und Transformationen, die darauf aufsetzen, und unterschiedliche Beschreibungszugänge, um den Workflow zu verdeutlichen und einfach nachnutzbar zu machen. Die Plattform (Klug/Galka 2022) wurde mit Testusern entwickelt, die sich willig auf das Experiment eingelassen haben. Die dafür produzierten Daten stehen mittlerweile offen in einem funktionell umfangreichen Webauftritt zur Verfügung: 33 Texte zum Thema „Jagd in den deutschsprachigen Texten des Mittelalters und der Frühen Neuzeit“ (<http://jagd-im-mittelalter.de/>), transkribiert und modelliert von Timo Bülters und Simone Schultz-Balluff, und die hyperdiplomatische Transkription der Handschrift Wien, ÖNB, Cod. 3085, fol. 1r-39v von Astrid Böhm (2022).

Workflow

Die hyperdiplomatische Transkription erfolgt mittels Transkribus, einer Transkriptionssoftware, die eine automatische Textsegmentierung und Text-Bild-Verknüpfung erlaubt und darüber hinaus eine äußerst benutzerinnenfreundliche Arbeitsoberfläche bietet. In Transkribus wird mittels proprietärer Codierung transkribiert, wobei die proprietäre Kodierung, die im Transkriptionsprozess verwendet wird, im TEI-XML in der Zeichenbeschreibung (<charDecl>) festgehalten wird. Mit Daten aus Projekten, die diese Transkriptionsmethode anwenden (*Mittelalterlabor* , *Cooking recipes of the Middle Ages*), wurde in Transkribus außerdem ein HTR-Modell für spätmittelalterliche Bastarda trainiert (German Bastarda, hyperdiplomatic), das kurz vor der Publikation steht. Auch Textauszeichnungen (Initialen, Überschriften usw.) und die Annotationen wie z.B. Revisionen der Schreiberin oder des Schreibers oder Anmerkungen/Notizen der Editorinnen und Editoren erfolgen in Transkribus. Generell wird aber versucht, die Annotationen in Transkribus auf ein notwendiges Minimum zu

beschränken. Die Software ermöglicht einen TEI/XML-Export der Transkription – die exportierten Daten werden mittels X-Technologien bzw. mit auf der Plattform bereitgestellten XSLT-Stylesheets weiterverarbeitet und ins finale Datenmodell transformiert. Die Kollationierung der Transkription mit der Quelle erfolgt außerhalb von Transkribus und setzt auf den selben Transformationen auf. Das publikationsfähige TEI-Dokument enthält einen umfangreichen <teiHeader> mit <editorialDecl>, <msDesc> und <charDecl> und den Editionstext, der die Struktur des Faksimiles abbildet. Die XSLT-Stylesheets, Schema-Dateien zur Validierung, Templates für Handschriftenbeschreibung (<msDesc>) und Editionsrichtlinien (<editorialDecl>) werden auf GitHub zum Download bereitgestellt (<https://github.com/ditah-at/hyper>).

Die Archivierung und Publikation erfolgt mittels des Geisteswissenschaftlichen Asset Management Systems (GAMS). GAMS ist ein Asset Management System zur Verwaltung, Publikation und Langzeitarchivierung digitaler Ressourcen aus allen geisteswissenschaftlichen Fächern, welches auf der Open-Source-Lösung Fedora (*Flexible Extensible Digital Object Repository Architecture*) basiert und am Zentrum für Informationsmodellierung an der Universität Graz ständig weiterentwickelt wird. Mit Hilfe des Systems können unterschiedliche Ressourcen verwaltet, mit Metadaten angereichert und persistent zitierbar publiziert werden (Stigler/Steiner 2018). Das Asset-Management-System entspricht den Empfehlungen des OAIS-Referenzmodells (*Open Archival Information System*), welche eine verlässliche Langzeitarchivierung von digitalen Daten garantieren sollen. Die Auffindbarkeit der Daten wird durch die Vergabe von persistenten Identifikatoren und der Anreicherung mit umfassenden Metadaten ermöglicht. GAMS ist seit 2019 mit dem Core Trust Seal als vertrauenswürdige digitales Repositorium und seit Frühjahr 2020 außerdem als CLARIN Datencenter zertifiziert.

Das finale TEI/XML-Dokument wird gemeinsam mit den Faksimiles in das Repositorium ingestiert und unter einem systeminternen Permalink langzeitarchiviert. Die Daten werden außerdem mit einem PID aus dem Handle.net-System versehen und stehen stabil referenzierbar zur Verfügung. Im Zuge des Ingest wird automatisch die HTML-Anzeige *on the fly* mittels gängigen Webtechnologien basierend auf vorgefertigten XSLT-Transformationen generiert. Der Webauftritt der Hyperdiplomatischen Transkriptionsplattform umfasst eine vertikale wie auch horizontale Text-Bild-Synopse (wahlweise mit diplomatischer Transkription oder Lesetext), die Editionsrichtlinien, die Handschriften- und Zeichenbeschreibung, Metadaten und Lizenzinformationen sowie Zitiervorschläge.

Die Hyperdiplomatische Transkriptionsplattform stellt zur Nachvollziehbarkeit und Nachnutzbarkeit des Workflows sämtliche XSLT-Stylesheets etc. zur Verfügung. Außerdem werden in einer umfangreichen Dokumentation (<http://gams.uni-graz.at/o:hyper.documentation/sdef:TEI/get?mode=overview>), die sich der Transkription, dem Datenmodell, der Kollationierung und Validierung sowie der Archivierung und Publikation der Daten widmet, die einzelnen Schritte mit Verlinkungen zu weiteren Ressourcen, wie z.B. dem *KONDE-Weißbuch* (Klug/Galka/Steiner 2021) mit Einträgen zur digitalen Edition, genau erläutert. Eine Kurzbeschreibung (<http://gams.uni-graz.at/o:hyper.documen->

tation/sdef:TEI/get?mode=shortdescription) der einzelnen Schritte erspart das ständige zur Rate ziehen der ausführlichen Dokumentation und ein Aktivitätsdiagramm veranschaulicht die einzelnen Arbeitsschritte und bietet eine Verlinkung zu den jeweiligen Anleitungen (<http://gams.uni-graz.at/archive/objects/context:hyper/methods/sdef:Context/get?mode=diagram>).

Fazit

Die Plattform und das Editions-konzept richtet sich grundsätzlich an all jene Editorinnen und Editoren, die bereits Vorkenntnisse in Bezug auf digitale Editionen mitbringen, wie z.B. Wissen über XML/TEI, XSLT, oder den Umgang mit dem Oxygen-XML-Editor. Das Paradigma der "Minimal Edition" bezieht sich hier auf einen nachhaltig erarbeiteten und von weiteren Editionsprojekten nachnutzbaren Workflow. In der Regel evaluiert jedes Editionsprojekt in unterschiedlichen Stadien mögliche Tools und Workflows, mit denen im Team gearbeitet werden soll, sei es für die Transkription, Annotation oder für die Publikation. Im Projekt Hyper wurde jedoch ein standardisierter Workflow entwickelt, der auf wenigen, zur sofortigen Nutzung bereitstehenden Tools und Routinen aufbaut und in einer umfangreichen Dokumentation beschrieben wird, wodurch Editorinnen und Editoren, welche diesen Text auch einem größeren Publikum im Zuge einer digitalen Edition möglichst einfach zugänglich machen wollen, auf diesen zugreifen können. Die "Einschränkung" ist, dass die hyperdiplomatische Transkriptionsmethode übernommen werden muss. Risam/Gill 2022b schlagen für ihre Variante des *minimal editing* vor, nur die absolut notwendigen Technologien für die Umsetzung eines bestimmten Vorhabens zu verwenden. Da einer der Grundpfeiler von digitalen Edition aber die Modellierung der Daten in einem standardisierten Datenformat ist, wurde bewusst entschieden, dass bei der hyperdiplomatischen Modellierung mit TEI/XML keine Abstriche gemacht werden sollen. Da sowohl Editierende wie auch Benutzerinnen und Benutzer von digitalen Editionen gewisse Standards in der Datenpräsentation erwarten (Caria/Mathiak 2018), werden auch dafür komplexe Technologien verwendet, die allerdings an die verwendete Repositoriums-lösung angeknüpft sind. Sowohl die Modellierung der Quelle als auch die Präsentation sind in einem projektspezifischen Datenmodell verwirklicht. *Hyper* unterscheidet sich von anderen Tools und Lösungen auch dadurch, dass die Daten langzeitar-chiviert werden und unter einem PID erreichbar sind. "Minimal Editing" bedeutet im Rahmen dieser Plattform also nicht nur, dass es ein fertiges Datenmodell und einen erprobten Workflow gibt, sondern auch, dass der Aufwand seitens der Repositoriums-betreiber für die Archivierung und Präsentation der Daten sehr gering ist. Alle veröffentlichten Quellen stehen *open access* zur Verfügung, außerdem werden sämtliche Dateien, die im Workflow benötigt werden, wie XSLT-Stylesheets oder Schema-Dateien offen auf GitHub bereitgestellt.

Ausblick

Das Ziel der *Minimal Edition* ist es nicht, möglichst viele Editionsprojekte möglichst schnell umzusetzen, sondern einen niederschweligen Zugang anzubieten, damit ohne großen Aufwand "reduzierte" digitale Editionen geschaffen werden können, die trotzdem qualitativen Standards entsprechen. Die *Hyperdiplomatische Transkriptionsplattform* ist zwar momentan hauptsächlich für all jene interessant, die hyperdiplomatisch transkribieren – allerdings ist es ein erster Schritt in eine Zukunft, in der auch andere Editionen, die dem *minimal-editing*-Konzept folgen, ohne großen organisatorischen Aufwand publiziert werden können.

Bibliographie

- Böhm, Astrid. 2022. "Hyperdiplomatische Transkription der Handschrift Wien, ÖNB, Cod. 3085, fol. 1r-39v." In *Hyperdiplomatische Transkriptionsplattform*, hg. von Helmut W. Klug unter Mitarbeit von Selina Galka. GAMS. PID: o:hyper.OENB3085 (zugegriffen: 25. Juli 2022).
- Böhm, Astrid, und Helmut W. Klug. 2020. "Quellenorientierte Aufbereitung historischer Texte im Rahmen digitaler Editionen: Das Problem der Transkription in mediävistischen Editionsprojekten." In *Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen. Akten der Tagung Bamberg, 08.-10. November 2018*, hg. von Ingrid Bannwitz und Martin Fischer, S. 51-72. Bamberg: University of Bamberg Press.
- Caria, Federico, and Brigitte Mathiak. 2018. "Minimal Functionality for Digital Scholarly Editions." In *Digital Cultural Heritage: Final Conference of the Marie Skłodowska-Curie Initial Training Network for Digital Cultural Heritage, ITN-DCH 2017, Olimje, Slovenia, May 23-25, 2017, Revised Selected Papers*, hg. von Marinos Ioannides, 350-63. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-75826-8_28 (zugegriffen: 25. Juli 2022).
- CETELcean. TEI in HTML5 Custom Elements. Raffaele Vigilanti. <https://github.com/TEIC/CETELcean> (zugegriffen: 22. Dezember 2022).
- CoreTrustSeal. <https://www.coretrustseal.org/> (zugegriffen: 25. Juli 2022).
- Data Seal of Approval Synopsis. <https://www.coretrustseal.org/about/history/data-seal-of-approval-synopsis-2008-2018/> . (zugegriffen: 25. Juli 2022).
- EVT (Edition Visualization Technique). Roberto Rosselli Del Turco. <http://evt.labcd.unipi.it> (zugegriffen: 22. Dezember 2022).
- FAIR Principles. <https://www.go-fair.org/fair-principles/> (zugegriffen: 25. Juli 2022).
- FEDORA Commons. <http://www.fedora-commons.org/> (zugegriffen: 25. Juli 2022).
- Franzini, Greta, Melissa Terras und Simon Mahony. 2019. "Digital editions of text: Surveying user requirements in the Digital Humanities." In *ACM Journal on Computing and Cultural Heritage (JOCCH)* 12 (1): 1-23. <https://doi.org/10.1145/3230671> (zugegriffen: 25. Juli 2022).

GAMS. <https://gams.uni-graz.at> (zugegriffen: 25. Juli 2022).

Gil, Alex: "The User, the Learner and the Machines We Make", 21. Mai 2015. <http://go-dh.github.io/min-comp/thoughts/2015/05/21/user-vs-learner/> (zugegriffen: 25. Juli 2022).

Handle.net Registry. <https://www.handle.net/> (zugegriffen: 25. Juli 2022).

Hofmeister-Winter, Andrea. 2003. "Das Konzept einer „Dynamischen Edition“ dargestellt an der Erstaussgabe des „Brixner Dommessnerbuches“ von Veit Feichter (Mitte 16. Jh.).", Göppingen: Kümmerle.

GitHub: Hyper. <https://github.com/ditah-at/hyper> (zugegriffen: 25. Juli 2022).

Klug, Helmut W., Astrid Böhm und Elisabeth Raunig (Hrsg.). 2019. Mittelalterlabor. Transkription der Handschrift Graz, UB, Ms. 1609. hdl.handle.net/11471/521.60 (GAMS. 521.60.) (zugegriffen: 25. Juli 2022).

Klug, Helmut W., Astrid Böhm und Christian Steiner (Hrsg.). 2020-02. CoReMA - Cooking Recipes of the Middle Ages. Corpus - Analysis - Visualisation. hdl.handle.net/11471/562.10 (GAMS. 562.10) (zugegriffen: 25. Juli 2022).

Klug, Helmut W., Selina Galka und Elisabeth Steiner (Hrsg.). 2021. KONDE Weißbuch. hdl.handle.net/11471/562.50. (GAMS. 562.50.) www.digitale-edition.at (zugegriffen: 25. Juli 2022).

Klug, Helmut W. und Astrid Böhm 2021. "Datenmodell "Hyperdiplomatische Transkription"." In *KONDE Weißbuch*, hg. von Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". Handle: hdl.handle.net/11471/562.50.50. PID: o:konde.50 (zugegriffen: 25. Juli 2022).

Klug, Helmut W. und Selina Galka (Hrsg.). 2022. Hyperdiplomatische Publikationsplattform. gams.uni-graz.at/hyper. (zugegriffen: 25. Juli 2022).

OAIS Reference Model (ISO 14721). <http://www.oais.info/> (zugegriffen: 25. Juli 2022).

Pierazzo, Elena. 2019. "What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter." In *International Journal of Digital Humanities* 1 (2): 209-20. doi:10.1007/s42803-019-00019-3.

Risam, Roopika und Alex Gil. 2022a. Minimal Computing. *Digital Humanities Quarterly*, Vol 16 (2). <http://www.digitalhumanities.org/dhq/vol/16/2/index.html> (zugegriffen: 25. Juli 2022).

Risam, Roopika und Alex Gil. 2022b. "Introduction: The Questions of Minimal Computing." In *Digital Humanities Quarterly* 16 (2). <http://www.digitalhumanities.org/dhq/vol/16/2/000646/000646.html> (zugegriffen: 25. Juli 2022).

Stigler, Johannes, und Elisabeth Steiner. 2018. "GAMS – Eine Infrastruktur zur Langzeitarchivierung und Publikation geisteswissenschaftlicher Forschungsdaten." In *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 71: 207–16.

TEI-Publisher. The Instant Publishing Toolbox. <https://teipublisher.com/index.html> (zugegriffen: 22. Dezember 2022).

transcriptiones. Transkriptionen historischer Manuskripte erstellen, teilen, nutzen. <https://transcriptiones.ch> (zugegriffen: 22. Dezember 2022).

Musikgeschichte im distant reading: Präsentation der Musikverlagsdatenbank mvdb

Rosenthal, Maximilian

maximilian.rosenthal@hmt-leipzig.de

Hochschule für Musik und Theater Leipzig, Deutschland

Richter, Matthias

matthias.richter@slub-dresden.de

Sächsische Landesbibliothek, Staats- und Universitätsbibliothek Dresden

Das DFG-geförderte Projekt *Geschmacksbildung und Verlagspolitik*, eine Kooperation der Hochschule für Musik und Theater (HMT) Leipzig und der Sächsischen Landesbibliothek (SLUB) Dresden, erfasst die Geschäftsdaten von Leipziger Musikverlagen des 19. Jahrhunderts (Hofmeister, C. F. Peters, Rieter-Biedermann) in der eigens dafür gebauten Musikverlagsdatenbank (*mvdb*, <https://musikverlage.slub-dresden.de>). Zielsetzung des Projekts ist es, den Einfluss von Musikverlagen auf Musikgeschichte im weiteren bzw. auf Geschmacksbildungs- und Kanonisierungsprozesse im engeren Sinne zu verstehen. In vorliegendem Papier sollen zunächst Methoden und Zielsetzung des Projekts kurz erläutert werden. Abschnitte 2 und 3 werden dann die *mvdb* genau vorstellen, die das digitale Herzstück des Vorhabens bildet. Es werden sowohl ihre wesentlichen Funktionen erläutert, als auch die dafür nötigen Lösungen einiger typischer Probleme von Geisteswissenschaften im digitalen Medium vorgestellt. Als Ausblick werden abschließend einige erste Ergebnisse aus der Auswertung der Forschungsdaten präsentiert.

Einleitung: Projektvorhaben, Methode, Projektziele

Der Transfer der Musikwissenschaft ins Digitale ist unbestreitbar auf dem Vormarsch.¹ Damit öffnen sich zwar neue wissenschaftliche Handlungsspielräume, diese müssen aber nun auch mit geeigneten Fragestellungen ausgefüllt werden. In der ‚verspäteten Disziplin‘ (Gerhard 2000), der „eine größere Widerständigkeit eigen“ ist (Unsel 2019, 172) gegenüber interdisziplinären Transfers, setzt sich in jüngerer Zeit vermehrt ein Bewusstsein dafür durch, dass Musikgeschichte losgelöst von den sozial-, wirtschafts- und kulturgeschichtlichen Kontexten ihrer Entstehung kaum noch hinreichend ist. Für das 19. Jahrhundert hat sich dies z.B. in Publikationen zu Verlags-, Salon-, und Repertoiregeschichte niederge-

schlagen (Ballstaedt/Widmaier 1989, Beer 2000, Keym/Schmitz (Hg.) 2016, Gerber 2016, Beer 2020). Größere Repertoiregeschichtliche Forschungsvorhaben gibt es erst wenige (etwa Widmaier 1998, Hartmann-Enke 2022, auch <https://performance.musicconn.de/>). Das liegt unter anderem daran, dass hier einerseits geeignete Quellen vorliegen müssen, andererseits Datenmengen bearbeitet werden, die sich nur durch digitale Methoden hinreichend bewältigen lassen. Neben digitalen Musik- und Korpusanalysen (Neuwirth/Rohrmeier 2016) und digitalen Editionen besteht darin eines der großen Potentiale der Digital Musicology. Schon Helmut Loos hat darauf hingewiesen, dass es „im vordringlichen Interesse der Musikforschung [liegt], das bereitstehende innovative technische Potential der Gegenwart gewinnbringend zu nutzen und in die Erforschung eines bislang beinahe undurchdringlichen Gebiets [= Repertoireforschung, die Autoren] einzubringen“ (Loos 2010, 158). Oder in Worten von Laurent Pugin: „There are at least two key areas in which digital technology is transforming research: access and scale.“ (Pugin 2015, 1)

Hier knüpft das Projekt *Geschmacksbildung und Verlagspolitik* an, indem es aus Geschäftsunterlagen die Verlagsprogramme dreier Musikverlage inklusive ihrer Wirtschaftsdaten erfasst. Aus ca. 20.000 Verlagsnummern lassen sich dann im Zeitraum von 1807 bis ca. 1945 Bewegungen des Musikalienmarkts als ‚distant reading‘ (Moretti 2016) der Geschmacksentwicklung rekonstruieren und sowohl zu den Agenden von Musikverlegern und Musikjournalisten als auch dem musikalischen ‚Kanon‘ in Beziehung setzen. ‚Distant reading‘ ist hier adaptiert zu verstehen und meint nicht die ursprüngliche korpusanalytische Konzeption nach Moretti, denn das Projekt arbeitet nicht mit Werk-Texten. Stattdessen wird das Repertoire anhand bibliographischer und verlegerischer Metadaten analysiert. Die Prämissen sind jedoch dieselben, weshalb die Begriffswahl adäquat erscheint: Im Kontrast zu bisherigen Perspektiven auf die Musikgeschichtsschreibung, die sich entlang einer vergleichsweise überschaubaren Zahl von (kanonischen) ‚close reading‘ Beispielen bewegt, wird hier ein weiterer Blickwinkel eingenommen, der aufgrund des Verfahrens mehr Kompositionen erfasst, als eine Einzelperson je lesen könnte. Methodisch stellt sich das als ‚mixed methods‘-Zirkel dar: Deskriptive Statistiken erzeugen eine neue Lesart von musikgeschichtlichem Verlauf, erzeugen aber auch Erklärungsbedarfe, für die traditionelle Textquellen herangezogen werden. Andererseits lassen sich Annahmen über Musikgeschichte als Hypothesen statistisch am Material prüfen. Multivariate Statistiken sind ebenfalls möglich. Beispiele in Form erster Ergebnisse dafür werden am Schluss dieses Papiers vorgestellt.

Die Musikverlagsdatenbank *mvdb*



Abbildung 1: Startseite mvdb

Die *mvdb* ist „Herzstück“ des Projekts, weil sie einerseits der Datenerfassung und -organisation dient, andererseits aber auch die Plattform ist, auf der die Verlagsdaten der Forschungscommunity zur Verfügung gestellt werden.

Technisch gesehen ist die *mvdb* eine Erweiterung des PHP-basierten Content Management Frameworks TYPO3 (Version 9.5). Es bietet eine umfangreiche Programmierschnittstelle für Extensions, die solche Funktionen realisieren. Die Entwicklung von TYPO3-Extensions wird darüber hinaus durch Tools der TYPO3-Community wie den Extensionbuilder (Version 9.10) unterstützt. So konnten wir in verhältnismäßig kurzer Zeit einen Prototyp für unsere Forschungsumgebung realisieren. Sie besteht aus vier Komponenten (Abb. 2): (1) zur Normdatenverwaltung, (2) zur Terminologieverwaltung, (3) zur Errechnung von Infografiken und (4) einer Core-Komponente. Die Komponenten für Normdaten und Terminologie versorgen die Core-Komponente mit wesentlichen Metadaten. Letztere selbst dient der Erfassung und Präsentation der durch das Projekt erfassten Wirtschaftsdaten. Sie enthält ein Backendmodul mit grundlegenden Funktionalitäten zur Dateneingabe und Qualitätssicherung und mehrere Plugins, die verschiedene Teile des Frontends der *mvdb* steuern. Dazu zählen u. a. ein Plugin, das die API bereitstellt, und ein Recherche-Plugin. Während die Core-Komponente projektspezifisch gestaltet ist, lassen sich die übrigen auch für ähnliche Projekte fortnutzen. Deshalb wird die komplette Code-Basis des Projekts frei zur Verfügung gestellt.

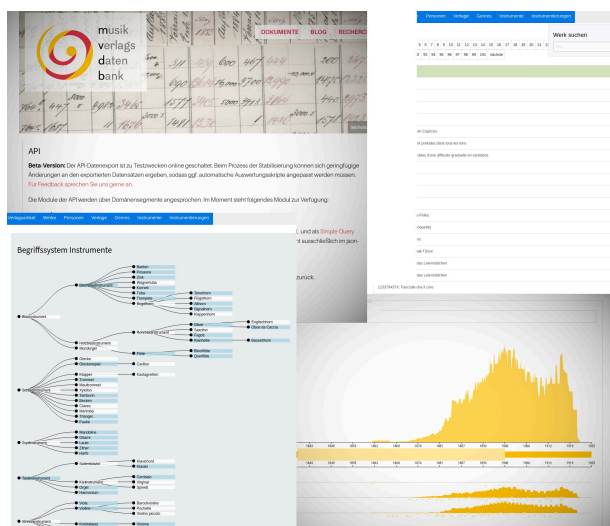


Abbildung 2: Die vier aktuellen Komponenten der mvdb (i. Uhrzeigersinn von links oben: API, Core, Terminologieverwaltung, Infografiken)

Im Projekt werden in der *mvdb* aus den Geschäftsbüchern der Verlage Hofmeister (1807 bis ca. 1939), Rietter-Biedermann (1856 bis 1917) und Peters (1867 bis ca. 1940) vier Hauptinformationen händisch erfasst: 1) Verlagsnummer und 2) Titel der Ausgabe sowie 3) das Auflagedatum und 4) die Höhe der Auflage. Stand August 2022 sind etwa 30% der rund 20.000 Verlagsnummern erreicht. Händische Erfassung ist nötig, da OCR an den komplizierten handschriftlichen Quellen scheitert (vgl. unten Abb. 4). Außerdem bedürfen manche Unregelmäßigkeiten noch einer Interpretationsleistung durch die Bearbeitenden. Um die Daten für statistische und datenbasierte Verfahren anzureichern, wird jeder Verlagsartikel durch bibliographische Daten der Gemeinsamen Normdatei (GND, s. Behrens 2011) der Deutschen Nationalbibliothek ergänzt. Das bietet gegenüber eigener Auszeichnungen drei Vorteile: 1) Effizienz: Obwohl viele Datensätze von Bibliothekar*innen der SLUB für das Projekt neu angelegt werden müssen, sind etliche tausend Werkdatensätze bereits vorhanden. 2) Qualität: Für die GND gelten fundierte und erprobte Erfassungsstandards, zudem sind sie untereinander bereits reichhaltig vernetzt. 3) Gute wissenschaftliche Praxis: Aufgrund ihrer Standards gewährleisten die GND-Daten Interoperabilität und Nachnutzbarkeit. Dass dabei die Übertragung nicht immer reibungslos möglich ist, wird weiter unten nochmals thematisiert.

Für die musikwissenschaftliche Forschung leistet die *mvdb* insbesondere drei Dinge:

Aufbereitung und Archivierung

Bei der Erstellung der verlinkten GND-Werk- und Personennormdatensätze werden die Informationen bibliothekarisch qualifiziert anhand von biographischen und bibliographischen Quellen, sodass die aus den sporadischen Geschäftsbucheinträgen hervorgehenden Verlagsartikel eindeutig identifiziert und kontextualisiert sind. Sofern es die Projektressourcen noch zulassen, ist auch geplant, die entsprechenden Quellendigitalisate und relevante Kontextquellen abzulegen. Nur am Rande

sei erwähnt, dass die durch die *mvdb* erfassten Daten von der SLUB unbefristet vorgehalten und archiviert werden.

Präsentation

Über unser Frontend werden die Daten für die Community nutz- und nachschlagbar, sodass die Geschäftsdaten der Musikverlage einfach sichtbar und navigierbar sind. Die Recherche ist über das Recherche-Plugin auf der Datenbank katalogisch oder über eine Suchfunktion möglich. Im einfachsten denkbaren Fall kann die *mvdb* über einfache philologische Fragen zur Identität einer Ausgabe wie z.B. Ersterscheinungsdatum oder Zahl der Titelaufgaben beantworten. Zu den bibliographischen Mehrinformationen zu Werk und Personen lässt sich von jedem Verlagsartikeldatensatz einfach navigieren. Auch einige weitergehende ‚on-board‘-Visualisierungen zur Datenauswertung sind eingebaut, so werden beispielsweise in der Verlagsartikel- und der Personenansicht summarische Grafiken angeboten, die einen schnellen Überblick über die Konjunktur der Ausgabe bzw. aller Werke einer*r Komponist*in zulassen (Abb. 3).

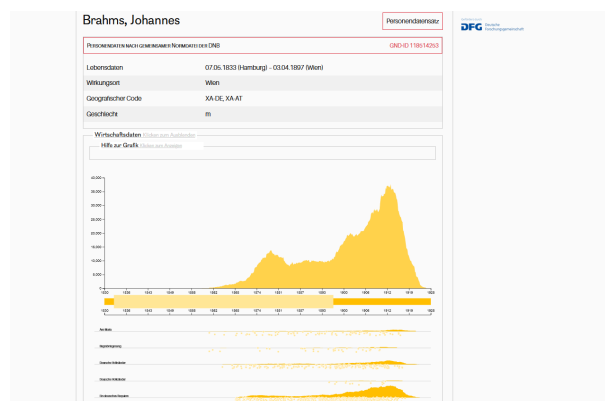


Abbildung 3: Personengrafik Brahms mit einzelnen Werken und aufaddierter Gesamtkurve (Fünf-Jahres-Mittel, Screenshot mvdb)

Nachnutzung

Die Daten lassen sich über das API-Plugin mit Lizenz CC-BY ungefiltert für eigene Forschung exportieren. Sie können im JSON-Format mithilfe der Elasticsearch-Query-DSL abgefragt werden, wozu Clientmodule in allen gängigen Programmier- und Statistiksprachen existieren, von R und Python für eigene statistische Auswertungen bis zu PHP und Javascript zum dynamischen Einbinden in projektfremde Kontexte. Außerdem wird es der Community möglich sein, Daten aus ähnlichen Quellen zu ergänzen, sodass die Bildung einer weiteren Digitalen Inselressource verhindert wird. So kann die *mvdb* perspektivisch ein digitaler Hub für musikwissenschaftliche Wirtschaftsdaten und bibliographische Informationen zu Verlagsausgaben des 19. Jahrhunderts werden.

Problemlösungen

Bei einer vergleichsweise jungen Disziplin wie den digitalen Geisteswissenschaften gehören auch Lösungsansätze für methodische Probleme, wie sie nachfolgend vorgestellt werden, zu den wichtigen Ergebnissen.

Was ist eine Ausgabe?

Dass geisteswissenschaftliche Gegenstände gelegentlich Unschärfen aufweisen, die sich in den rigiden Strukturen des Digitalen nicht recht abbilden lassen, ist keine Neuheit (z.B. Kuczera u.a. 2019) und gilt auch für das musikalische Werk (Pugin 2015, 2). Bei der Einrichtung der *mvdb* standen wir insbesondere vor dem Problem, dass ‚die Ausgabe‘ in der Verlagspraxis des 19. Jahrhunderts keine homogene Einheit darstellt.² Aus dem losen Zusammenhang der Druckplatten ließen sich beliebig ganze Bände oder separat publizierte Einzelnummern drucken, die oft simultan erschienen. Häufig lässt sich beobachten, dass erfolgreiche Einzelnummern eines Bandes nach einigen Jahren als Ableger separat gedruckt wurden, weil diese von geringerem Umfang und damit günstiger und besser verkäuflich waren. Außerdem wurden etliche Werke in Einzelstimmen gedruckt, von denen die Stimmen einzeln und unterschiedlich oft verkauft wurden. Damit zerfällt ‚das Werk‘ in ein Konglomerat von losen Publikationsmodulen, die aber trotzdem kognitiv eine eindeutige Sinn- und Werkeinheit darstellen.

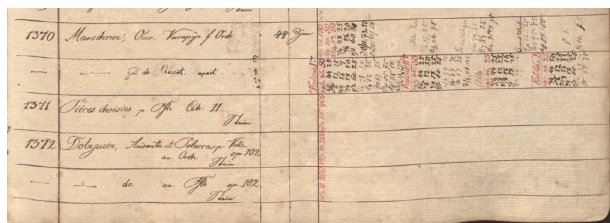


Abbildung 4: Nr. 1370 des Verlags Hofmeister, H. Marschner, Overtüre zum „Vampyr“ in Ausgaben für Orchester sowie Quartett einzeln, mit jeweils einzelnen Nachauflagen verschiedener Stimmen, D-LEsta 21072 F. Hofmeister, Nr. 43.

Dem haben wir bei der Einrichtung des Datenmodells Rechnung getragen, indem jede Ausgabe als flexibler Zusammenhang von übergeordnetem Verlagsartikel (= „Makro“) und untergeordneten Teilartikeln (= „Mikros“) verstanden wird, die nur zusammen eine Ausgabe abbilden können. Was diese Aufteilung für die Auflagestatistiken bedeutet, loten wir durch verschiedene Bereinigungsverfahren noch aus, aber klar ist, dass sich Auflagezahlen einzelner Orchesterstimmen oder Nummern nur unter Vorbehalt zu einem ‚Gesamterfolg‘ einer Ausgabe aufaddieren lassen. Auch solcherlei Modellierungen erachten wir bereits als wichtige Ergebnisse, denn daraus lassen sich gleich mehrere Konsequenzen für die Digital Humanities und die Musikwissenschaft ziehen. Inhaltlich nämlich, dass die musikwissenschaftliche Auffassung des Werkbegriffs anhand dieser Erkenntnisse zu revidieren wäre. Damit wird die modulare Modellierung womöglich auch für andere Forschungs-

projekte und insbesondere digitale Editionsprojekte relevant. Formal bedeutet eine solche, von bisherigen Konzepten abweichende Modellierung wahrscheinlich auch, dass statistische und datenbezogene Kompetenz vermehrt Teil des fachlichen Methodenkanons werden müssen, um die richtige Interpretation der so abgelegten Daten zu gewährleisten.

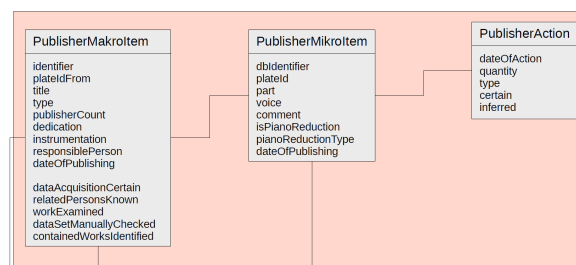


Abbildung 5: Auszug Datenmodell mvdb, Abbildung Makro-Mikro-Struktur

„Norm“-Daten?

Das Potential von Normdaten als musikwissenschaftlichen Forschungsdaten wurde bereits elaboriert (Wiermann 2018, Bicher/Wiermann 2018), und auch deren Probleme benannt: Dateninkonsistenzen, Dubletten, Fehlende Datensätze, variable Tiefenerschließung und heterogene Differenzierungsgrade der Informationen (Wiermann 2018, 347–8). Selbst wenn sich die Fehlermagnitude durch die projektinterne Normdatenerfassung verringern lässt, sind GND-Daten auch in Bezug auf das Begriffssystem selbst problematisch. Eigentlich sollen dessen vertikale Ober- und Unterbegriffsrelationen in die *mvdb* übernommen werden. Unter anderem stellte sich aber heraus, dass aufgrund der nicht eindeutig festgelegten Relationstypen grundsätzlich mehrere Oberbegriffe angegeben werden können – was eine begriffstheoretisch wünschenswerte Informationsablage in einer Baumstruktur verhindert. Daneben finden sich je nach Begriffsbereich sehr unterschiedliche Granularitäten, die bei unreflektierter Übernahme und ohne spezielle Maßnahmen quantitative Untersuchungen verzerren können. Ähnliche Probleme entstehen im Bereich der Instrumente, wo Besetzungen und Einzelinstrumente gleichberechtigt in Werkdatensätzen abgelegt werden dürfen. Schließlich werden Alternativbesetzungen in der Form ‚Klavier alternativ für Orchester‘ angegeben. Die tatsächliche Alternativbesetzung kann so nur intellektuell im Kontext der Hauptbesetzung erschlossen werden. Um diese Probleme zu bearbeiten, hat die *mvdb* ein eigenes Sachbegriffsmodul, das baumartig strukturierte Systeme von GND-IDs erfasst und Normdaten zur Benennung der Einzelknoten aus der Normdatei abfragt (s. Abb. 2). Diese Begriffssysteme können einheitlich granular strukturiert werden. Besetzungsknoten können einheitlich ausgezeichnet werden. Alternativbesetzungen werden automatisch erschlossen und sauber mit vollständiger Instrumentierung abgelegt. Auch dass es keine Versionierung der GND-Daten gibt, ist nicht unproblematisch bei der Vernetzung der Daten. Hier hoffen wir ebenfalls, dass unsere Lösungen einen gewissen

Modellcharakter erhalten können: das Sachbegriffsmodul zur Anpassung von GND-Normdaten werden wir der Community als open source Repository zur Verfügung stellen.

Ausblick auf Ergebnisse

Da die Datenerfassung noch in vollem Gang ist, müssen erste Ergebnisse als Einblick genügen. Diese Ergebnisse sind zwar endgültig, aber in Bezug auf die Tiefe der Datenanalyse und Interpretation noch nicht repräsentativ für die Ambitionen des Projekts. Bereits abgeschlossen ist die Erfassung des vergleichsweise kleinen Programms von Rieter-Biedermann (ca. 2900 Nummern) von 1856 bis 1917. Eindeutig ist an der beinahe linear wachsenden jährlichen Gesamtproduktion (Abb. 6) erkennbar, dass der Verlag weder vom Auftreten der günstigen ‚Klassikerausgaben‘ nach 1867 (*Collection Litolf*, *Edition Peters*) beeinträchtigt wurde, noch von Konkurrenztechnologien mechanischer Musikinstrumente, die um 1900 aufkamen (vgl. Heise 2022). Deutlich dagegen wird sichtbar, wie Kriege und Krisen den Absatz von Kulturgütern hemmten (s. etwa die Kriege 1870/71, 1914-1918).

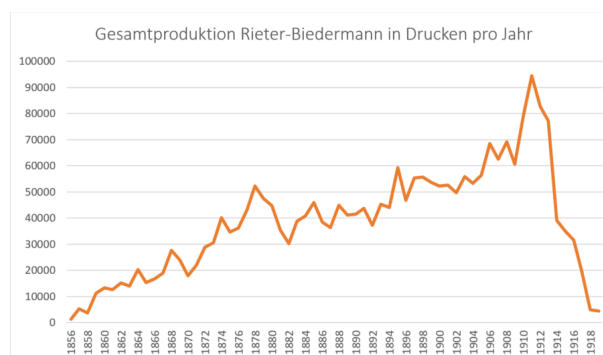


Abbildung 6: Gesamtproduktion Rieter-Biedermann jährlich

Vergleicht man dies mit der Zahl der produzierten (Neu-)Ausgaben pro Jahr (Abb. 7) wird wiederum deutlich, dass Rieter-Biedermann etwa ab den 1890ern immer weniger einzelne Verlagsartikel pro Jahr, diese dafür aber in höheren Auflagen auf den Markt bringt. Ähnlich verhält es sich mit den Komponisten, die der Verlag jährlich neu ins Programm aufnimmt (Abb. 8). Anfangs nimmt der Verlag bis zu 37 neue Komponisten in einem Jahr auf, die Zahl reduziert sich aber bald drastisch. Neuaufnahmen erfolgen dann in immer kleiner werdenden Wellen, der Gesamtrend ist abnehmend. Es eröffnet sich also eine Schere zwischen Produktpalette und Produktionsmenge, die man – noch ohne die Identität des Verlegten genauer zu hinterfragen – als Selektionsprozesse erfolgreicher Werke deuten kann. Womöglich wird dadurch hier bereits die ‚invisible hand‘ (Winko 2002) der Kanonisierung ein Stück weit sichtbar.

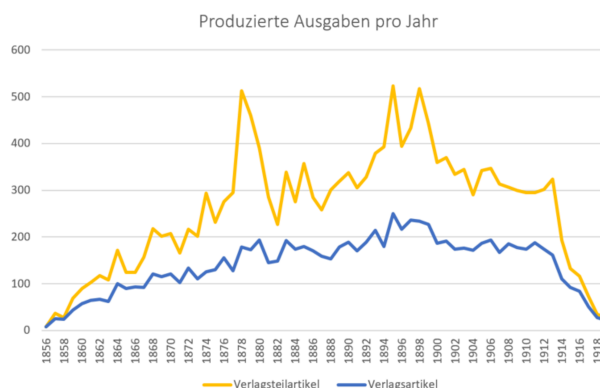


Abbildung 7: Ausgaben Rieter-Biedermann jährlich

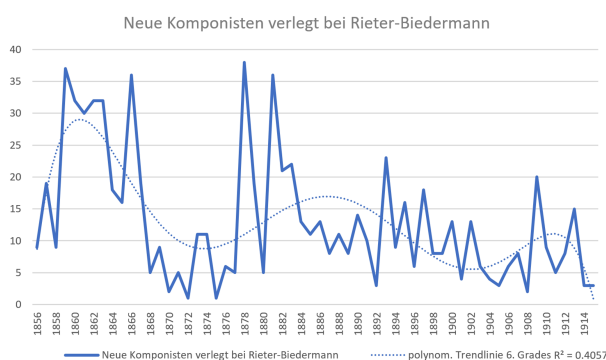


Abbildung 8: Neu ins Verlagsprogramm von Rieter-Biedermann aufgenommene Komponisten pro Jahr

Diese Tendenz steht wahrscheinlich auch in Zusammenhang mit dem Höhenflug von Johannes Brahms. Brahms ist im Gesamtvergleich der mit Abstand meistgedruckte Komponist im Verlagsprogramm von Rieter-Biedermann (Abb. 9), und das obwohl sein Hauptverleger eigentlich Simrock war. Brahms' *Deutsches Requiem* op. 45 ist wiederum mit einem guten Viertel aller Brahms-Drucke der ‚erfolgreichste‘ Verlagsartikel aus Rieter-Biedermanns Programm gewesen (Abb. 10), was das Werk wahrscheinlich nicht zu geringem Anteil den kriegerischen und nationalistischen Tendenzen des Kaiserreichs zu verdanken hat (s. Zunahme ab den 1890ern in Abb. 3). Das wiederum ist freilich bemerkenswert, da es die Erwartung unterläuft, die die Rangplätze 2 bis 10 bestätigen: dass die erfolgreichsten Verlagsartikel allgemein eigentlich Klavierstücke und Lieder für den häuslichen Gebrauch waren. Die Beispiele veranschaulichen, auf welche Weise die *mvd*b und das Projekt ermöglichen, das Repertoire des 19. Jahrhunderts ‚aus der Distanz‘ zu entschlüsseln und im Lichte des Verlagshandels neu zu beleuchten. Weitere solche Ergebnisse wären zu zeigen, würden aber den Rahmen sprengen. Der Fortschritt des Projekts kann stattdessen auf dem Portal der *mvd*b transparent nachverfolgt werden.

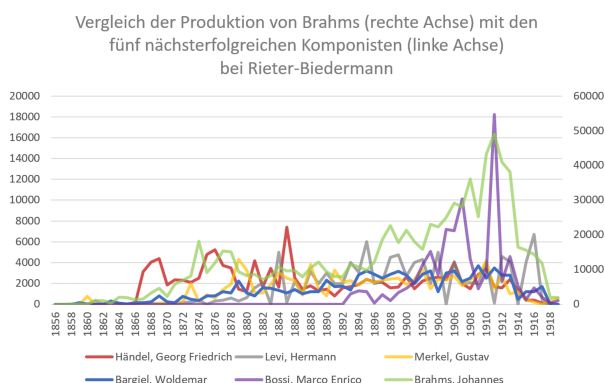


Abbildung 9: Jährliche Produktion der sechs meistgedruckten Komponisten im Verlag Rieter-Biedermann im Vergleich

Rang	Werktitel	Opusnr.	Komponist	Totale Drucke	Erstauflage
1	Ein deutsches Requiem	45	Brahms, Johannes	195316	1868-10-17
2	Gesänge	43	Brahms, Johannes	86595	1868-12-23
3	Frühlingslieder	35	Bargiel, Woldemar	69965	1867-10-17
4	Deutsche Volkslieder	NA	Brahms, Johannes	56455	1864-12-10
5	Lieder	2	Levi, Hermann	53827	1861-06-20
6	Romanzen	33	Brahms, Johannes	53406	1865-04-28
7	Lieder	2, Nr. 6	Levi, Hermann	51710	1872-05-15
8	Lieder und Romanzen	44	Brahms, Johannes	49565	1866-11-08
9	Lieder und Gesänge	32	Brahms, Johannes	44740	1864-02-23
10	Volkskinderlieder, Nr. 4, G-Dur	WoO 31	Brahms, Johannes	34175	1872-11-21

Abbildung 10: Rangliste der 10 meistgedruckten Artikel im Verlag Rieter-Biedermann

Fußnoten

- Die Fortschritte der letzten Jahre lassen sich gut nachzeichnen anhand von Cook 2004, Crawford/Gibson 2009, Schmale 2016 und dem Themenheft *Digitalität der Musikforschung* 71/4 (2018).
- Die beste, aber noch immer unzureichende Terminologie hat Rasch 2005 vorgelegt.

Bibliographie

Ballstaedt, Andreas und Widmaier, Tobias. 1989. *Salonmusik. Zur Geschichte und Funktion einer bürgerlichen Musikpraxis*. Stuttgart: Steiner.

Beer, Axel. 2000. *Musik zwischen Komponist, Verlag und Publikum. Die Rahmenbedingungen des Musikschaffens in Deutschland im ersten Drittel des 19. Jahrhunderts*, Tutzing: Hans Schneider.

Beer, Axel. 2020. *Das Leipziger Bureau de Musique (Hoffmeister & Kühnel, A. Kühnel). Geschichte und Verlagsproduktion (1800–1814)*, München/Salzburg: Katz-bichler.

Behrens, Renate. 2011. „Die Gemeinsame Normdatei – ein Kooperationsprojekt.“ *Dialog mit Bibliotheken* 1: 37–40. <https://d-nb.info/101606361X/34>

Bicher, Katrin und Wiermann, Barbara. 2018. „Normdaten zu ‚Werken der Musik‘ und ihre Potenzial für die digitale Musikwissenschaft.“ *Bibliothek – Forschung und Praxis* 42/2: 222–235. 10.1515/bfp-2018-0043

Cook, Nicholas. 2004. „Comparative and Computational Musicology.“ In *Empirical Musicology: Aims, Methods, Prospects*, hg. von Eric Clarke und Nicholas Cook, 103–126. Oxford: Oxford University Press.

Crawford, Tim und Gibson, Lorna (Hg.). 2009. *Modern Methods for Musicology. Prospects, Proposals, and Realities*. Farnham: Ashgate.

Gerber, Mirjam. 2016. *Zwischen Salon und Geselligkeit. Henriette Vogt, Livia Frege und Leipzigs bürgerliches Musikleben*, Hildesheim: Olms.

Gerhard, Anselm (Hg.). 2000. *Musikwissenschaft – eine verspätete Disziplin? Die akademische Musikforschung zwischen Fortschrittsglauben und Modernitätsverweigerung*. Stuttgart/Weimar: Metzler.

Hartmann-Enke, Linus. 2022. *Gewandhaus repertoire levels for composers and works 1800–1895* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5862025>

Heise, Birgit (Hg.). 2022. *Paul Ehrlich und die Anfänge der Leipziger Musikautomaten-Industrie*, Altenburg: Kamprad.

Keym, Stefan und Schmitz, Peter (Hg.). 2016. *Das Leipziger Musikverlagswesen. Innerstädtische Netzwerke und internationale Ausstrahlung*. Hildesheim: Olms.

Kuczera, Andreas u. a. (Hg.). 2019. *Die Modellierung des Zweifels – Schlüsselideen und –konzepte zur graphbasierten Modellierung von Unsicherheiten*. Sonderband 4 *Zeitschrift für digitale Geisteswissenschaften*. 10.17175/sb004

Loos, Helmut. 2010. „Vom Idealen zum Realen. Ein Paradigmenwechsel.“ *Lietuvos muzikologija* 11: 152–160.

Moretti Franco. 2016. *Distant Reading. Aus dem Englischen übersetzt von Christine Pries*, Konstanz: Konstanz University Press.

Neuwirth, Markus und Rohrmeier, Martin. 2016. „Wie wissenschaftlich muss Musiktheorie sein? Chancen und Herausforderungen musikalischer Korpusforschung.“ *Zeitschrift der Gesellschaft für Musiktheorie* 13/2: 171–193. 10.31751/915

Pugin, Laurent. 2015. „The Challenge of Data in Digital Musicology.“ *Frontiers in Digital Humanities* 2: 1–3. 10.3389/fdigh.2015.00004

Rasch, Rudolf. 2005. „Basic Concepts.“ In *The Circulation of Music, Bd. 1: Music Publishing in Europe 1600–1900. Concepts and Issues, Bibliography*, hg. von dems., 15–46. Berlin: BWV.

Stadler, Peter. 2019. „Musikwissenschaft und Digital Humanities.“ In *Historische Musikwissenschaft- Gegenstand – Geschichte – Methodik*, hg. von Frank Hentschel, 330–339.

Unsold, Melanie. 2019. „Historische Musikwissenschaft als Kulturwissenschaft.“ In *Historische Musikwissenschaft- Gegenstand – Geschichte – Methodik*, hg. von Frank Hentschel, 170–182. Laaber: Laaber.

Schmale, Wolfgang. 2016. „Digital Musicology im Kontext der Digital Humanities.“ In *Wissenskulturen der Musikwissenschaft. Generationen – Netzwerke – Denkstrukturen*, hg. von Sebastian Bolz u. a., 299–310. Bielefeld: transcript.

Widmaier, Tobias. 1998. *Der deutsche Musikalienleihhandel. Funktion, Bedeutung und Topographie einer Form*

gewerblicher Musikaliendistribution vom späten 18. bis zum frühen 20. Jahrhundert, Saarbrücken: Pfau.

Wiermann, Barbara. 2018. „Bibliothekarische Normdaten und digitale Musikwissenschaft.“ *Die Musikforschung* 71/4: 338–357.

Winko, Simone. 2002. „Literatur-Kanon als *invisible hand*-Phänomen.“ *TEXT+KRITIK. Zeitschrift für Literatur, Sonderband Literarische Kanonbildung 2002*: 9–24.

Narrativität und Handlung: Zum Verhältnis von Handlungs- zusammenfassungen und relevanten Ereignissen

Hatzel, Hans Ole

hans.ole.hatzel@uni-hamburg.de
Universität Hamburg

Gius, Evelyn

evelyn.gius@tu-darmstadt.de
Technische Universität Darmstadt

Stierner, Haimo

stierner@linglit.tu-darmstadt.de
Technische Universität Darmstadt

Biemann, Chris

christian.biemann@uni-hamburg.de
Universität Hamburg

Relevante Ereignisse in Erzähltexten

Welche Ereignisse in Erzähltexten sind besonders relevant? Diese Frage wird in der Literaturwissenschaft im Kontext von verschiedenen Konzepten verhandelt. So können relevante Ereignisse identifiziert werden, indem man die für die Textinterpretation als besonders wichtig erachteten Stellen (so genannte „Schlüsselstellen“) betrachtet (Arnold & Fiechter, 2022). Auf die Rezeption orientiert sind ebenfalls die empirische Leser:innenforschung (Groeben 1977; Miall & Kuiken 2001) oder die Rezeptionsästhetik (Iser 1976). Steht hingegen der Text im Fokus, kann die Frage nach der Wichtigkeit von Ereignissen in Bezug auf Ereignishaftigkeit oder die so genannte

Erzählwürdigkeit untersucht werden (z. B. Hühn 2014; Baroni 2012). Allen Ansätzen gemeinsam ist, dass sie bestimmte Qualitäten von Texten bzw. Textbestandteilen betrachten.

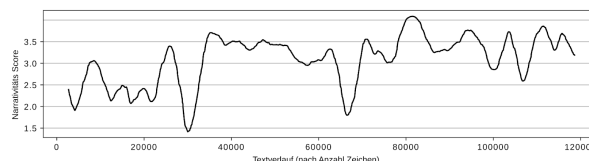


Abb. 1: Narrativitätskurve für Kafkas Die Verwandlung (Vauth et al., 2021)

In unserem EvENT-Projekt wurde bereits die Identifikation und Klassifikation von Ereignissen anhand von textuellen Merkmalen vorgenommen. Den hierbei entwickelten vier Ereignistypen haben wir Werte entsprechend ihrer Ereignishaftigkeit zugewiesen und darauf basierend Narrativitätskurven erzeugt, die den Verlauf der Ereignishaftigkeit über einen Text abbilden (siehe Abb. 1).¹

Im Fortgang des Projektes wollen wir überprüfen, inwiefern zwischen diesen grundlegenden Ereignistypen, die auch unter dem Konzept „Event I“ subsumiert werden können, und besonders erzählwürdigen Ereignissen bzw. sogenannten Events II, eine Verbindung besteht.² Nimmt man an, dass ein Event II ein Event I mit weiteren Qualitäten ist, so ist die Bestimmung von Events II insofern komplexer, als sie zusätzliches Wissen erfordert – etwa um Regelmäßigkeiten und Auffälligkeiten in der fiktionalen Welt, aber auch um den Kontext. Mit anderen Worten, wir wollen die von uns im Vorlauf vorgenommene textoberflächenbasierte Detektion von Ereignissen dahingehend prüfen inwiefern mit ihr auch jene Textstellen erfasst werden, die in Handlungszusammenfassungen als besonders handlungsrelevant bzw. erzählwürdig gelten.

Was bereits mit unseren bestehenden Daten und ohne die weitere Operationalisierung des Event II-Konzepts möglich ist, ist der Abgleich unserer Annotationen mit als besonders handlungsrelevant markierten Textstellen. Diesen Vergleich stellen wir im vorliegenden Beitrag an, indem wir unsere Annotationen von Ereignissen in literarischen Texten mit Zusammenfassungen der entsprechenden Texte abgleichen.

Semiprofessionelle, professionelle und nutzer:innengenerierte Handlungszusammenfassungen

Wir gehen davon aus, dass Textstellen durch ihre Erwähnung in Zusammenfassungen als für die Handlung wichtig markiert werden. Für den Abgleich dieser Textstellen mit unseren Narrativitätsverläufen nutzen wir drei Typen von Zusammenfassungen: (1) semiprofessionelle Zusammenfassungen, die Studierende der Literaturwissenschaft verfasst haben, (2) professionelle Zusammenfassungen aus Kindlers Literatur Lexikon und (3)

nutzer:innengenerierte Zusammenfassungen der Online-Enzyklopädie Wikipedia. Die Verwendung dieser verschiedenen Zusammenfassungstypen zielt darauf ab, zu analysieren, welche Aspekte bzw. Textstellen für die jeweiligen Zusammenfassungstypen relevant sind und dadurch Rückschlüsse auf ihre Qualität zu ziehen.

Die (1) Zusammenfassungen der Studierenden waren Teil der Studienleistungen in einem Seminar. Sie wurden als explizit auf die Handlung bezogene Zusammenfassungen verfasst, die eine maximale Länge von 20 Sätzen haben durften. Außerdem wurden die Studierenden aufgefordert, keine Hilfsmittel (wie Zusammenfassungen auf Wikipedia oder aus Literaturlexika) zu nutzen. Für die vier genutzten Primärtexte *Das Erdbeben in Chili* (Heinrich von Kleist, 1807), *Die Judenbuche* (Annette von Droste-Hülshoff, 1842), *Krambambuli* (Marie von Ebner-Eschenbach, 1896) und *Die Verwandlung* (Franz Kafka, 1915) haben wir jeweils 11, 9, 11 und 10 Zusammenfassungen.

Aus den Beiträgen des (2) Kindler-Literaturlexikons und von (3) Wikipedia wurden nur jene Passagen verwendet, die sich auf die Handlung in den Primärtexten beziehen. Passagen, die sich der Autor:in, Rezeption oder Interpretation widmen, wurden nicht berücksichtigt. Durch eine kollaborative Annotation der einzelnen Sätze wurde deren Bezug auf die Handlung des Textes annotiert und ein entsprechender Goldstandard erstellt, der den weiteren Analysen zugrunde liegt. So wurde sichergestellt, dass alle drei Zusammenfassungstypen handlungsorientiert sind. Die Zusammenfassungen wurden dahingehend annotiert, dass jeder Satz der Zusammenfassung mit einer Referenz auf alle Spannen des Primärtextes versehen wurde, auf die er Bezug nimmt.³ Als logische Einheiten wurden entsprechend die bereits vorliegenden Annotationen aller Verbalphrasen in Bezug auf die vier Ereignistypen genutzt.⁴ Dabei wurden konkret jeweils eine oder mehrer Spannen im Originaltext als zugehörig annotiert, die Sequenzen relevanter Ereignisse enthalten; ein Satz der Zusammenfassung kann also mehrere Passagen im Originaltext referenzieren.

Evaluation der Zusammenfassungen

Um die Qualität der Zusammenfassungen und mögliche Unterschiede zwischen den drei Zusammenfassungstypen zu analysieren, evaluieren wir deren Ähnlichkeit. Dazu nutzen wir drei Metriken, mit denen implizit drei unterschiedliche Auffassungen von Ähnlichkeit verbunden sind: Eine weitgehend lexikalische (N-Gramme), eine Metrik auf Basis distributioneller Semantik (Word Embeddings) und eine, die sich weitgehend von der sprachlichen Struktur löst und inhaltsbezogene Vergleiche vornimmt (adaptierte Pyramiden-Methode).

Wir nehmen zunächst an, dass die semiprofessionellen Zusammenfassungen durchweg handlungsbezogen sind. Deshalb vergleichen wir jeweils eine semiprofessionelle Zusammenfassung mit allen anderen semiprofessionellen und jede Zusammenfassung aller anderen Typen mit allen semiprofessionellen.

N-Gramm-basierte Ähnlichkeit

Als erstes berechnen wir BLEU- (Papineni et al., 2002) und ROUGE-Scores (Lin et al., 2004), die Ähnlichkeiten unter Zusammenfassungen als N-Gramm-Ähnlichkeit abbilden.

Wir gewichten BLEU-{1,2,3} und ROUGE-{1,2,3} jeweils gleich und quantifizieren so die Überlappung von 1-, 2- und 3-Grammen zwischen den unterschiedlichen Texten und geben für BLEU die Precision und ROUGE den F1-Score an.

Anhand der Scores in Tab. 1 und Tab. 2 wird ersichtlich, dass die semiprofessionellen Zusammenfassungen nahezu durchgehend die höchsten Ähnlichkeitswerte aufweisen.

Tab. 1: BLEU-{1,2,3} Scores für Ähnlichkeiten. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,27	0,23	0,24	0,22
professionell	0,15	0,2	0,19	0,09
nutzer:innengeneriert	0,14	0,22	0,24	0,1

Tab. 2: ROUGE-{1,2,3} F-Scores. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,51	0,56	0,6	0,47
professionell	0,58	0,53	0,63	0,43
nutzer:innengeneriert	0,34	0,49	0,55	0,31

Word-embedding-basierte Ähnlichkeit

N-Gramm-basierte Metriken haben den Nachteil, dass kleine Unterschiede in der Wortwahl zu einer deutlich geringeren Ähnlichkeit führen können. Um Vergleiche stärker auf die Semantik zu fokussieren, wurde mit BERTScore (Zhang et al., 2020) eine embedding-basierte Methode etabliert. Wenden wir diese auf unsere Texte an, zeigt sich ein deutlich geringerer Unterschied der Zusammenfassungstypen (siehe Tab. 3). Dies weist darauf hin, dass Unterschiede in der N-Gramm-basierten Bewertung zu einem großen Teil auf Unterschiede in der Wortwahl zurückzuführen sind.

Tab. 3: BERTScore F-Werte. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,74	0,71	0,74	0,72
professionell	0,69	0,71	0,74	0,69
nutzer:innengeneriert	0,72	0,68	0,74	0,72

Inhaltsbasierte Ähnlichkeit

Für den letzten Vergleich der Zusammenfassungen adaptieren wir die Pyramiden-Methode, die für die automatische Evaluation maschinell generierter Zusammenfassungen entwickelt wurde (Nenkova et al., 2004). Die Zusammenfassungen werden auf Basis von sogenannten Summary Content Units (SCU) mit Referenzzusammenfassungen verglichen. Eine SCU repräsentiert dabei

eine semantische, inhaltliche Aussage aus dem Zusammengefassten. Die namensgebende Pyramide repräsentiert dabei das Vorkommen der unterschiedlichen SCUs in der Menge der Referenzzusammenfassungen, wobei die Höhe der Pyramide n der Anzahl der Referenzzusammenfassungen entspricht. Dabei ist in der Regel eine Verteilung zu beobachten, die tatsächlich eine Pyramide aufbaut: eine SCU taucht in allen n Texten auf und bildet die Spitze, einige wenige SCUs tauchen in $n-1$ Texten auf usw. bis in der letzten Stufe SCUs auftauchen, die nur in einer Zusammenfassung vorkommen. Eine zu evaluierende Zusammenfassung sollte nun, um ihren Pyramiden-Score zu maximieren, SCUs aus hohen Schichten der Pyramide enthalten. Dabei erhält die oberste Schicht das Gewicht n , sodass jede SCU aus dieser Schicht n Punkte gibt. Der Score einer Zusammenfassung wird als Anteil der tatsächlichen Punkte an denen der optimalen Zusammenfassung der gleichen Länge (in SCUs) angegeben. Entsprechend sind die Pyramiden-Scores reelle Zahlen im Intervall 0 bis 1, wobei 1 eine perfekte Zusammenfassung beschreibt.

Wir passen die Pyramiden-Methode in zwei Punkten an unsere Fragestellung an. Zum einen haben wir keine Referenztexte, sondern benutzen das Verfahren zum Vergleich verschiedener Zusammenfassungen. Zum anderen enthalten unsere Daten keine SCUs, diese werden deshalb über Textspannen approximiert. Dafür nehmen wir zunächst an, dass jede Spanne des Textes eine Menge von SCUs enthält, die insofern eindeutig ist, als sie keine Schnittmenge mit den SCUs anderer, disjunkter Textabschnitte hat. Insofern kann jede Textspanne auf eine oder mehrere SCUs abgebildet werden. Textspannen werden derart in Unterspannen zerlegt, dass Spannen sich nur überlagern, wenn sie identisch sind. Somit erhalten wir Spannen, die gemäß unserer Annahme semantisch eindeutig sind (siehe Abb. 2). Eine Textspanne kann nach unseren Annahmen mehrere SCUs enthalten die wir als eine behandeln, dies entspricht einer Ereignis Modellierung in größerer Granularität.

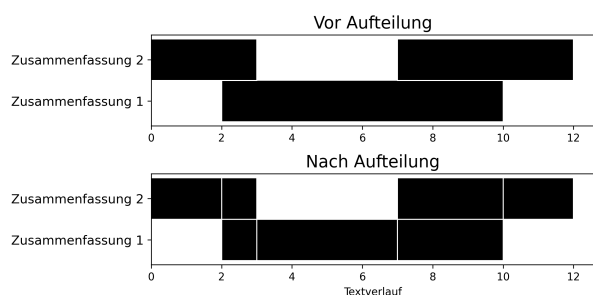


Abb. 2: Segmentierung von zwei Annotationsspannen in semantisch eindeutige SCUs. Spannen werden derart zerlegt, dass sich nur noch gleiche Spannen überhaupt überschneiden.

Die Auswertung der Zusammenfassungen mit der Pyramiden-Methode ist in Tab. 4 zu sehen. Nahezu alle semiprofessionellen Zusammenfassungen liegen dabei über dem Wert 0,70 (siehe Abb. 3), während die anderen Zusammenfassungen im Vergleich zum Mittelwert schlechter abschneiden (siehe Tab. 4).

Tab. 4: Pyramiden-Scores für unterschiedliche Zusammenfassungen. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,85	0,80	0,84	0,80
professionell	0,73	0,52	0,75	0,55
nutzer:innengeneriert	0,71	0,70	0,81	0,62

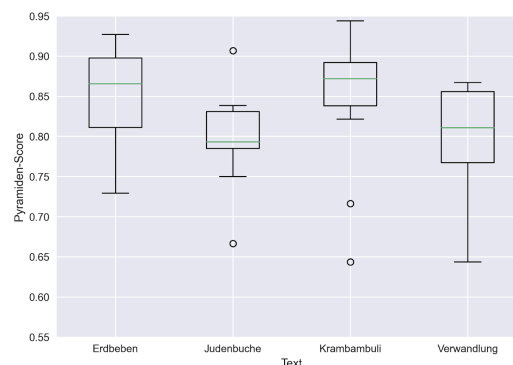


Abb. 3: Verteilung der Pyramiden-Scores für die semiprofessionellen Zusammenfassungen.

Ähnlichkeiten zwischen den Zusammenfassungen

Insgesamt wird so deutlich, dass eine Betrachtung auf Pyramiden-Ebene Unterschiede offenbart, die zwar denen der N-Gramm-Methoden ähnlich, jedoch nicht grundsätzlich anhand oberflächlichen maschinellen Textauswertungen (z.B. BERTScore) festzumachen sind.

Narrativität und Handlung

Wir wollen nun evaluieren, wie Erzählwürdigkeit, repräsentiert durch die Handlungszusammenfassungen, und Narrativität, repräsentiert durch unsere Narrativitätsgraphen, zusammenhängen. Wir überprüfen dafür, ob der Teil des Originaltextes, auf den sich die Zusammenfassungen beziehen, einen großen Narrativitätswert aufweist. Als erste Analyse berechnen wir dazu den Narrativitätswert der in der Zusammenfassung referenzierten Passagen. Wir setzen diesen ins Verhältnis zum erwarteten Gesamtscore, gegeben der Länge der in der Zusammenfassung enthaltenen Textstellen (in Ereignissen).⁵ Somit ergibt sich im Mittel ein Wert von 1,0 bei zufälliger Auswahl der Passagen. Ein Wert $> 1,0$, hingegen heißt, dass in der Zusammenfassung referenzierte Passagen mehr Narrativität aufweisen als nicht referenzierte Passagen. Dies ist tatsächlich für alle Zusammenfassungstypen der Fall (siehe Tab. 5).

Auch ein Vergleich von Ereignissen, die in Zusammenfassungen genannt werden, mit jenen die es nicht werden, bestätigt dies anhand der Narrativitätswerte: im Mittel 3,13 für genannte und 2,86 für nicht genannte Ereignisse.

Tab. 5: Faktoren des erwarteten Narrativitätswerts. Für die semiprofessionellen Zusammenfassungen wird der Mittelwert (inklusive der Standardabweichung) angegeben.

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	1,04±0,06	1,02±0,09	1,04 ±0,07	1,06±0,10
professionell	1,00	1,03	1,02	1,05
nutzer:innengeneriert	1,06	1,12	1,07	1,08

Für den Vergleich der Ausschläge der Narrativitätskurven verwenden wir den Gipfelprominenzfaktor. Dabei handelt es sich um ein Maß, welches die Wichtigkeit eines Ausschlags und damit seinen Wert im Vergleich zum umliegenden Kurvenverlauf quantifiziert. Für diesen Vergleich werden alle lokalen Maxima in der cosinusgeglätteten Narrativitätskurve (window size=50) berücksichtigt und für jedes lokale Maximum wird die Gipfelprominenz berechnet.⁶ Jedes Ereignis erhält nun, wenn es ein lokales Maximum darstellt, den Wert der Gipfelprominenz, andernfalls den Wert 0. Nun wird wie oben verfahren und der erwartete Prominenzwert mit dem tatsächlichen verglichen. Es wird der erwartete Wert, also die durchschnittliche Gipfelprominenz des Graphen, ins Verhältnis zur tatsächlich vorgefundenen Gipfelprominenz des betrachteten Segments gesetzt. Lokale Maxima sind durch das Smoothing relativ selten, dementsprechend ist die Streuung der Werte deutlich größer. Dies erschwert die Interpretation der Werte. Interessant aber ist, dass in einigen Fällen der Faktor deutlich über 1 liegt (vgl. Tab. 7), wobei dies für die nutzer:innengenerierten Zusammenfassungen durchweg der Fall ist. Abb. 4 veranschaulicht trotz der starken Varianz erkennbare Unterschiede zwischen den Originaltexten.

Tab. 6: Gipfelprominenzfaktoren

	Erdbeben	Judenbuche	Krambambuli	Verwandlung
semiprofessionell	0,91±0,52	0,68±0,65	0,90±0,57	1,45±0,78
professionell	0,09	1,15	1,42	1,38
nutzer:innengeneriert	1,23	1,75	1,28	1,1

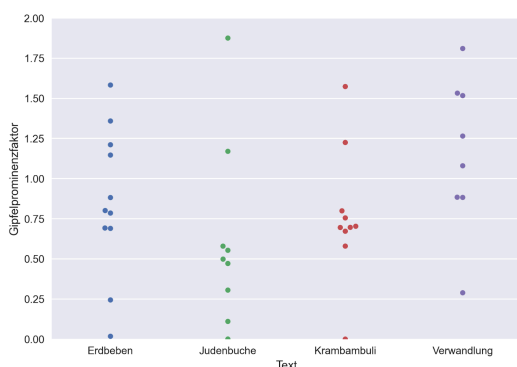


Abb. 4: Gipfelprominenzfaktoren der semiprofessionellen Zusammenfassungen

Narrativität und Handlung: Ein kurzes Fazit

Die vorgestellten Ergebnisse deuten darauf hin, dass die Nutzung von Zusammenfassungen für die weitere Arbeit mit Ereignissen und Ereignishaftigkeit produktiv ist. Hervorzuheben ist, dass wir einen Relevanzzusammenhang zwischen Ereignissen und Handlung nachweisen konnten, der auf einer Operationalisierung ersterer anhand der sprachlichen Oberfläche aufbaut. Damit kann unser Ereigniskonzept mit reduziertem Handlungsbezug anhand von handlungsbezogenen Informationen aus Zusammenfassungen weiterentwickelt werden. Die bereits umgesetzte, vergleichsweise erfolgreiche Automatisierung der Ereigniserkennung und damit der Narrativitätsverläufe wird nun in Bezug auf die handlungsbezogene Relevanz von Ereignissen erweiterbar, ohne dass Handlungsinformationen mühevoll manuell für die einzelnen Ereignisse bestimmt werden müssen. Dafür erscheint es vielversprechend, den Handlungsbezug von Zusammenfassungen weiter zu evaluieren und dabei ein Verfahren zu entwickeln, das besonders relevante Stellen identifizieren kann.

Danksagung

Dieser Beitrag entstand im von der DFG im Schwerpunktprogramm Computational Literary Studies (SPP 2207) geförderten Projekt „Evelautating Events in Narrative Theory“ (EvENT).

Fußnoten

1. Dabei hat der Ereignistyp, der Zustandsveränderungen beschreibt, den höchsten Wert, gefolgt von weiteren weniger ereignishaften Kategorien. Vgl. zu den Kategorien und zur Bestimmung der Ereignistypen die Annotationsrichtlinien Vauth & Gius (2021) und zu den Narrativitätskurven Vauth et al. (2021), für den Classifier Hatzel (2022).
2. Zur Diskussion der beiden narratologischen Ereigniskonzepte und ihrem Verhältnis vgl. Hühn (2014).
3. Trotz des mit diesen beiden Ansätzen erreichten hohen Handlungsbezugs haben 13% der Sätze keine entsprechend annotierte Spanne im Originaltext (18% bei den professionellen und 9% bei den semiprofessionellen Zusammenfassungen).
4. Die Daten wurden in Vauth & Gius (2022) publiziert.
5. Der Gesamtscore ist also jener, der bei der zufälligen Auswahl der Events aus dem Text im Durchschnitt zustande kommt. '1' bedeutet also, dass der Score zufällig ausgewählten Events entspricht, '2' heißt, dass der Score doppelt so hoch ist wie bei zufälligen Events.
6. Für die Berechnung wurde SciPy verwendet: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.peak_prominences.html#scipy-signal.peak_prominences
7. Für die Berechnung wurde SciPy verwendet: <https://docs.scipy.org/doc/scipy/reference/gene->

rated/scipy.signal.peak_prominences.html#scipy.signal.peak_prominences

Bibliographie

Arnold, Frederik, und Benjamin Fiechter. 2022. „Lesen, was wirklich wichtig ist. Die Identifikation von Schlüsselstellen durch ein neues Instrument zur Zitatanalyse“. In DHd2022. Potsdam, Deutschland. <https://doi.org/10.5281/zenodo.6327917>

Arnold, Heinz Ludwig, Hrsg. 2020. Kindlers Literatur Lexikon (KLL). Stuttgart: J.B. Metzler. <https://doi.org/10.1007/978-3-476-05728-0>.

Baroni, Raphaël. 2012. „Tellability“. In the living handbook of narratology, herausgegeben von Peter Hühn, John Pier, Wolf Schmid, und Jörg Schönert. Hamburg: Hamburg University Press. <http://hup.sub.uni-hamburg.de/lhn/index.php?title=Tellability&oldid=1577>.

Gius, Evelyn, und Michael Vauth. 2022. „Inter Annotator Agreement und Intersubjektivität“. In DHd2022, 147-151. Potsdam, Deutschland.

Groeben, Norbert. 1977. Rezeptionsforschung als empirische Literaturwissenschaft: Paradigma- durch Methodendiskussion an Untersuchungsbeispielen. Empirische Literaturwissenschaft. Bd. 1. Königstein/Ts.: Athenäum.

Hatzel, Hans Ole. 2022. Event Narrativity Classifier. Zenodo. <https://doi.org/10.5281/zenodo.6821142>.

Hühn, Peter. 2014. „Event and Eventfulness“. In the living handbook of narratology, herausgegeben von Peter Hühn, John Pier, Wolf Schmid, und Jörg Schönert. Hamburg: Hamburg University Press. <https://www.lhn.uni-hamburg.de/node/39.html>.

Iser, Wolfgang. 1976. Der Akt des Lesens. Theorie ästhetischer Wirkung. München: Fink.

Lin, Chin-Yew. 2004. „ROUGE: A Package for Automatic Evaluation of Summaries“. In Text Summarization Branches Out, 74-81. Barcelona, Spanien: Association for Computational Linguistics. <https://aclanthology.org/W04-1013>.

Miall, David, und Don Kuiken. 2001. „Shifting perspectives: Readers' feelings and literary response“. In New Perspectives on Narrative Perspective, herausgegeben von Willi Van Peer und Seymour Chatman, 289-301. Albany: SUNY Press.

Nenkova, Ani, und Rebecca Passonneau. 2004. „Evaluating Content Selection in Summarization: The Pyramid Method“. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 145-52. Boston, Massachusetts, USA: Association for Computational Linguistics. <https://aclanthology.org/N04-1019>.

Papineni, Kishore, Salim Roukos, Todd Ward, und Wei-Jing Zhu. 2002. „BLEU: a Method for Automatic Evaluation of Machine Translation“. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-18. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073135>.

Vauth, Michael, und Gius, Evelyn. 2021. „Richtlinien für die Annotation narratologischer Ereigniskonzepte“. Zenodo. <https://doi.org/10.5281/zenodo.5078174>.

Vauth, Michael, und Evelyn Gius. 2022. forTEXT/EvENT-Dataset: v.1.1 (Version v.1.1). Zenodo. <https://doi.org/10.5281/ZENODO.6406568>.

Vauth, Michael, Hans Ole Hatzel, Evelyn Gius, und Chris Biemann. 2021. „Automated Event Annotation in Literary Texts“. In CHR 2021: Computational Humanities Research Conference, 333-45. Amsterdam, Niederlande. http://ceur-ws.org/Vol-2989/short_paper18.pdf.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, und Yoav Artzi. „BERTScore: Evaluating Text Generation with BERT.“ In International Conference on Learning Representations. Online, 2020. <https://openreview.net/forum?id=SkeHuCVFDr>.

Offene Daten für die digitale Philosophie: Anforderungen an eine Datensammlung zur Philosophie und ihrer Geschichte

Heßbrüggen-Walter, Stefan

early.modern.thought.online@gmail.com
Universität Trier, Deutschland

Gegenstand meines Beitrags ist die Erörterung von Anforderungen an eine offene Datensammlung zur digitalen Philosophie und Philosophiegeschichte, insbesondere im Blick auf zwei Fragen. Einerseits ist aus der Sicht der digitalen Geisteswissenschaft zu fragen, welchen Kriterien eine solche Datensammlung genügen sollte, um valide Schlussfolgerungen zu erlauben. Dies betrifft sozusagen ihre "formale" oder "methodische" Seite. Andererseits ist aus der Sicht des Faches Philosophie zu fragen, welche Arten von Daten überhaupt für das Fach relevante Einsichten ermöglichen können. Dies betrifft die inhaltliche oder "materiale" Seite. Hier können beide Fragen natürlich nur exemplarisch erörtert werden, ihre Wichtigkeit für die Projektierung einer Datensammlung zur digitalen Philosophie und Philosophiegeschichte sollte aber auf der Hand liegen.¹

Zum Begriff der Datensammlung

Wir sprechen im folgenden von „Datensammlungen“,² weil die digitale Philosophie und Philosophiegeschichte, wie genauer zu zeigen wird, nicht nur Textdaten, sondern auch Metadaten über Text zu ihren Forschungsgegenständen zählt. Zudem kann die Philosophie des 20. Jahrhunderts und ihre Geschichte in reproduzierbarer Weise aus urheberrechtlichen Gründen weit überwiegend nur mittels „abgeleiteter Textformate“ (Schöch u. a.

2020) analysiert werden. Die Heterogenität dieser Datenformate legt es nahe, den generischen Begriff der Datensammlung anstatt spezifischerer Termini wie "digitale Textsammlung", "Corpus", "Kanon" (siehe dazu (Henny-Krahmer und Neuber 2017)) in Anschlag zu bringen.

Untersuchungsgegenstand

Analysieren möchte ich im folgenden zwei jüngere Studien zur digitalen Analyse von Philosophie im 20. bzw. 21. Jahrhundert (Malaterre u. a. 2021; Noichl 2021).³ Gegenstand sind zum einen die Philosophie insgesamt, deren Struktur anhand von Kozitationsanalysen erhellt werden soll,⁴ zum andern eine Teildisziplin der Philosophie, die Wissenschaftstheorie, deren diachrone Entwicklung anhand von topic models von acht einschlägigen Zeitschriften sichtbar gemacht werden soll.⁵ Die zugrundeliegenden Datensammlungen enthalten also einmal ausschließlich Metadaten und einmal Volltexte von Zeitschriftenartikeln mit korrespondierenden Metadaten, v. a. Erscheinungsjahr und die publizierende Zeitschrift. Damit ist schon eine erste Anforderung an eine offene Datensammlung zur digitalen Philosophie und Philosophiegeschichte benannt: Sie sollte nicht nur Volltexte, sondern auch Metadaten einschließen, selbst wenn die den Metadaten korrespondierenden Textträger nicht oder noch nicht digitalisiert sein sollten, sondern, wie bei Noichl, nur Angaben zu den erfassten Texten (wie der Aufsatztitel oder Abstract) sowie die Bibliographie zitierter Werke in die Datensammlung aufgenommen werden, da die eigentlichen Aufsätze selbst noch urheberrechtlich geschützt sind.

Zur Erstellung von Datensammlungen in der Philosophie: zwei Beispiele

Erster Schritt der Erstellung einer Datensammlung und demnach auch ihrer Bewertung ist nach Schöch 2017, 224 die Angabe, wie Gegenstand und Umfang eingegrenzt werden. Malaterre u. a. 2021, 2885 geben den Umfang ihrer Datensammlung mit 15897 englischsprachigen Aufsätzen an, die zwischen 1934 und 2017 in acht wissenschaftsphilosophischen Zeitschriften veröffentlicht worden sind. Die Volltexte wurden von JSTOR zur Verfügung gestellt. Dass die Vollständigkeit der Digitalisierung und die Korrektheit zugrundegelegten Metadaten mit Hilfe von weiteren Quellen überprüft wurden, ist nicht ersichtlich. Noichl 2021, 5092 nutzt als Ausgangspunkt der Erstellung der zugrundeliegenden Datensammlung die Fachbibliographie „Philpapers“ (Bourget und Chalmers o. J.). Die dort verzeichneten 1.782.816 Aufsätzen werden in zwei Schritten auf eine Datensammlung von insgesamt 68.152 Aufsätzen reduziert, indem zunächst Zeitschriften ausgeschlossen werden, die nach Meinung des Autors nicht zum fachlichen Kern der Philosophie zu zählen sind, allerdings ohne die Anzahl der damit ausgeschiedenen Aufsätze anzugeben. Für die verbliebenen Zeitschriften werden die in der Zitations-

datenbank „Web of Science“ enthaltenen Texte abgefragt. Aufsätze, die nicht mindestens viermal in anderen in „Web of Science“ enthaltenen Aufsätzen zitiert werden, werden ausgeschlossen. Es verbleiben 87.720 Datensätze. Aus dieser Teilmenge werden alle Datensätze entfernt, die nicht mindestens drei Zitationen enthalten, die auch in anderen Aufsätzen angeführt werden. Damit umfasst die zu analysierende Datensammlung am Ende Metadaten zu 68.152 Aufsätzen. Die zeitliche Erstreckung des erfassten Schrifttums wird nicht angegeben.

Festzuhalten ist zunächst, dass in beiden Analysen die verwendeten Datensammlungen nicht offen sind, die Ergebnisse somit nicht ohne weiteres überprüft bzw. reproduziert werden können. Dass als grundlegende Anforderung für die hier zu projektierende Datensammlung die Erfüllung der FAIR-Prinzipien zugrundegelegt ist, versteht sich eigentlich von selbst, soll aber hier dennoch ausdrücklich hervorgehoben werden.

Schöch 2017, 225 f unterscheidet weiter drei Modi der Festlegung von Datensammlungen: repräsentative Zufallsstichproben, "balancierte Sammlungen", in denen versucht wird, Objekte so auszuwählen, dass Kombinationen aller für die jeweilige Forschungsfrage einschlägigen Merkmale in der Sammlung vorhanden sind, sowie schließlich das Verfahren der „opportunistischen Auswahl“, die die Verfügbarkeit von Daten als wesentliches Auswahlkriterium an erste Stelle setzt.

Repräsentativität im statistischen Sinne setzt Bestimmung einer ‚Grundgesamtheit‘ voraus. Für Noichls Ziel, die ‚gegenwärtige Philosophie‘ als solche abzubilden ist eine solche Grundgesamtheit kaum konstituierbar. Selbst die auf Philpapers verzeichneten mehr als eine Million Aufsätze sind nicht als eine solche zu betrachten: Philosophie wird schließlich auch in Buchform publiziert. Auch aus inhaltlicher Sicht ist es zudem fraglich, ob die für die Bestimmung einer solchen Grundgesamtheit erforderliche Definition der Philosophie als Disziplin überhaupt möglich ist. Weitere Hindernisse für die Bestimmung der Grundgesamtheit selbst einer philosophischen Teildisziplin wie der Wissenschaftsphilosophie sind ebenfalls zu bedenken: selbst wenn es gelingen würde, ein solches Feld in operationalisierbarer Form einzugrenzen, müsste es auch in zeitlicher Hinsicht in einleuchtender Weise abgegrenzt werden. Dass der Beginn der Wissenschaftsphilosophie auf das Jahr 1934 festgelegt werden kann, ist aus fachlicher Sicht eine mit guten Gründen bezweifelbare Annahme.

Ob die von Malaterre et al. und Noichl vorgelegten Datensammlungen als ‚balanciert‘ gelten können, ist wohl ebenfalls eine strittige Frage. Wie man sie beantwortet, hängt davon ab, welche Merkmale als wesentlich für die behandelte Forschungsfrage anzusehen sind. Malaterre et al. gehen davon aus, dass nicht auf Englisch verfasste Texte für ihre Analyse vernachlässigbar sind bzw. der für deren Modellierung erforderliche Aufwand nicht notwendig ist.⁶ Die Sprache, in der ein Aufsatz abgefasst ist, wird also nicht als wesentliches Merkmal aufgefasst, sondern kann für die Untersuchung vernachlässigt werden. Noichls Daten lassen die diachrone Dimension außer acht, hier gilt also das Veröffentlichungsdatum sowohl des zitierenden wie des zitierten Textes nicht als wesentliches Merkmal, das für die Balance der zugrundegelegten Datensammlung erforderlich wäre.

Schlussfolgerungen

Aus diesen Befunden sind meines Erachtens zwei Schlussfolgerungen für die Ausgestaltung einer offenen Datensammlung für die Philosophie und Philosophiegeschichte zu ziehen: erstens sollte man sich wohl von dem Anspruch, mit einer Datensammlung die Disziplin als solche abzubilden, verabschieden. Ziel sollte vielmehr die Zusammenführung von Teildatensammlungen sein, die die Abhängigkeit von den sie leitenden Forschungsfragen offenlegen und damit auch das Gebiet bestimmen, innerhalb dessen aus den in ihnen enthaltenen Datensätzen valide Schlüsse gezogen werden können. Zweitens bedarf die Eingrenzung auf nur einen sprachlich-kulturellen Zusammenhang der forschungsbasierten Begründung und sollte nicht allein pragmatisch motiviert sein.

Nicht nur in methodischer, sondern auch in inhaltlicher Hinsicht kann man aus beiden Arbeiten jedoch wertvolle Hinweise erhalten, in welchen Hinsichten die Ergebnisse digitalbasierter Forschungen für die Philosophie relevant sein können. Dies betrifft zunächst die Unterteilung der akademischen Philosophie in Teildisziplinen, d. h. den Prozess ihrer Spezialisierung. Malaterre et al. legen eine solche Teildisziplin als ‚Analyseeinheit‘ zugrunde, nämlich die Wissenschaftsphilosophie. Noichl untersucht die Auffächerung der Philosophie in solche Spezialdisziplinen und -diskurse. Die Organisation von Forschungsdatensammlungen zur Philosophie und Philosophiegeschichte wird sich also auch an solchen Einheiten zu orientieren haben.

Zugleich wird die Frage zu beantworten sein, ob, und wenn ja in welchem Sinne, sich solche subdisziplinären Einheiten von gesamtphilosophischen Traditionen abgrenzen lassen. Noichl diskutiert z. B. auch die Unterscheidung zwischen ‚analytischen‘ und ‚kontinentalen‘ Traditionen der Philosophie. ‚Kontinentale‘ Traditionen wie die der Phänomenologie werden sich jedoch kaum als Teildisziplinen definieren, sondern eher als Teiltraditionen der Philosophie. Noichls Analyse spiegelt dies, da zur Abgrenzung der kontinentalen Philosophie von anderen Teilbereichen ein Kanon zitierter Autor:innen herangezogen wird (Noichl 2021, 5094).

Mit dem von Noichl gewählten Werkzeug der Koziationsanalyse lassen sich beide Dimensionen kaum voneinander abgrenzen. So wie Teildisziplinen Standardtexte zitieren, werden auch in philosophischen Traditionen gemeinsame Referenztexte als Bezugspunkte genutzt. Hier wird an einer inhaltlichen Analyse kein Weg vorbeiführen. Dass topic modeling in dieser Hinsicht eine hilfreiche Methode sein kann, zeigen Malaterre et al.: sie ermöglicht die Erschließung von Begriffskonstellationen und deren diachronen Verlauf. Während also die Identifikation von Traditionen und Schulbildungen wohl über die Betrachtung von Autor:innen und ihren Generationskohorten möglich sein dürfte, also durch Rekonstruktion von Kanonisierungs- und Dekanonisierungsprozessen von Personen, erlauben Verfahren der skalierten Erschließung von Inhalten Einblicke in die Kanonisierung und Dekanonisierung von Begriffen und deren Konstellationen.

Fazit: Anforderungen an philosophische Datensammlungen

Abschließend sollen die in diesem Beitrag entwickelten Anforderungen an eine Datensammlung digitaler Philosophie und Philosophiegeschichte kurz zusammengefasst werden. Digitale Philosophie benötigt Metadaten und Textdaten, die auffindbar, zugänglich, interoperabel und reproduzibel sind. Sowohl Textdaten als auch Metadaten sollten je nach Provenienz zumindest stichprobenhaft auf ihre Qualität hin überprüft werden. Die aus ihnen zu konstituierende Datensammlung sollte modular aufgebaut sein, um unterschiedlichen Forschungszielen, die den Teildatensammlungen zugrundeliegen, gerecht werden zu können. Datenquellen sind auf mögliche Verzerrungen und Auslassungen hin zu untersuchen. Diese sind, so weit sie sich pragmatisch aus dem zugrundeliegenden Forschungsziel der Teildatensammlung ergeben, zumindest explizit zu machen. Ein wichtiger Aspekt ist hierbei das Streben nach Multilingualität, um die globale Dimension der Philosophie angemessen abbilden zu können. Datensammlungen können dabei sowohl entlang von Teildisziplinen der Philosophie wie auch von Autor:innen, Epochen oder Traditionszusammenhängen konzipiert werden. Sie sollten es aber immer ermöglichen, auch die Entwicklung von Begriffen und Begriffskonstellationen – eines wesentlichen Mediums des Philosophierens – nachzuvollziehen.

Fußnoten

1. Mit der vorläufigen Klärung der hier als ‚material‘ bezeichneten Fragen ist natürlich, wie in der Begutachtung richtigerweise angemerkt, noch nicht alles gesagt, was aus fachphilosophischer Sicht zu Datensammlungen der Philosophie zu sagen wäre. Diese Debatten sollten jedoch zuerst innerhalb des Faches geführt werden und sind im disziplinären Zusammenhang der digitalen Geisteswissenschaften erst dann von Belang, wenn sie über das Fach hinausweisende Einsichten ermöglichen sollten (was wir naturgemäß erst wissen werden, wenn diese Debatten tatsächlich geführt worden sind).
2. Schöch 2017, 223 definiert sie als „Zusammenführung einzelner [...] Datensätze nach einer Einheit stiftenden Systematik“.
3. Der sich aus dieser Wahl des Gegenstandes ergebende Fokus auf die philosophische Zeitgeschichte ist kontingent: hier sind eben schon Untersuchungen mit Methoden der digitalen Geisteswissenschaften durchgeführt worden. Mutatis mutandis lassen sich jedoch die hier aufgeworfenen Fragen auf Datensammlungen zur Philosophiegeschichte übertragen.
4. Noichl 2021, 5091: „a visual representation of recent philosophy as a whole“.
5. Malaterre u. a. 2021, 2886: „an empirical basis for what might otherwise be informal claims about the discipline and its evolution in the past eight decades as reconstructed from the perspective of its major journals“.
6. Malaterre u. a. 2021, 2888: „Whenever journals were published in several languages, we retained only those

articles that were written in English due to algorithmic linguistic constraints.“

Bibliographie

Bourget, David, und David Chalmers. o. J. „PhilPapers: Online Research in Philosophy“. Zugegriffen 1. August 2022. <https://philpapers.org/>.

Henny-Krahmer, Ulrike, und Frederike Neuber. 2017. „EDITORIAL: Reviewing Digital Text Collections – RIDE“. *RIDE* 6. <https://doi.org/10.18716/ride.a.6.0>.

Malaterre, Christophe, Francis Lareau, Davide Pulizzotto, und Jonathan St-Onge. 2021. „Eight Journals over Eight Decades: A Computational Topic-Modeling Approach to Contemporary Philosophy of Science“. *Synthese* 199 (1): 2883–2923. <https://doi.org/10.1007/s11229-020-02915-6>.

Noichl, Maximilian. 2021. „Modeling the Structure of Recent Philosophy“. *Synthese* 198 (6): 5089–5100. <https://doi.org/10.1007/s11229-019-02390-8>.

Schöch, Christof. 2017. „Aufbau von Datensammlungen“. In *Digital Humanities: Eine Einführung*, herausgegeben von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein. Stuttgart: J.B. Metzler. <http://dx.doi.org/10.1007/978-3-476-05446-3>.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2020_006.

Offene Werkgenesen, Editionen und Archive. Versuch einer generischen Datenmodellierung

Bürgermeister, Martina

martina.buergermeister@onb.ac.at
Österreichische Nationalbibliothek, Österreich

Pektor, Katharina

katharina.pektor@onb.ac.at
Österreichische Nationalbibliothek, Österreich

Steindl, Christoph

christoph.steindl@onb.ac.at
Österreichische Nationalbibliothek, Österreich

Eigner, Johanna

johanna.eigner@onb.ac.at
Österreichische Nationalbibliothek, Österreich

Einleitung

Unser Beitrag zur Tagung *Open Humanities*, *Open Culture* präsentiert einen innovativen Vorschlag für die Datenmodellierung zukünftiger digitaler Werkeditionen. Das Modell ist hochgradig flexibel, indem es eine Diversität und komplexe Relationalität der Werkdaten zulässt, und zugleich stabil, wodurch es maximale Offenheit der Rezeption und Nachnutzbarkeit gewährleistet.

Entwickelt und erprobt wurde das Datenmodell im Zuge der Daten-Migration der Forschungsplattform *Handkeonline* (<https://handkeonline.onb.ac.at/>), bei der eine vorerst „geschlossene“ Modellierung von präbibliographischen Daten der verschiedenen, genetisch interpretierten Werkmaterialien aus dem Vorlass zusammen mit bibliographischen Daten der veröffentlichten Werke des Literaturnobelpreisträgers Peter Handke in eine „offene“ umgebaut werden musste – wir nennen diese Modellierung eine „offene Werkgenese“.

In unserer Präsentation werden wir zuerst kurz in die Voraussetzungen der Plattform *Handkeonline* (Punkt 2) und in den aktuellen Forschungsstand der semantischen Modellierung von bibliographischen und präbibliographischen Daten (Punkt 3) einführen. Danach wird anhand von zwei Beispielen der generische Modellierungsansatz für „offene Werkgenesen“ präsentiert (Punkt 4).

Forschungsplattform *Handkeonline*

Die Forschungsplattform *Handkeonline* wurde im Zuge eines FWF-Forschungsprojekts am Literaturarchiv der Österreichischen Nationalbibliothek Wien 2011–2015 entwickelt. Sie versteht sich als virtuelles Archiv, das sämtliche auf öffentliche und private Archive in Österreich, Deutschland und der Schweiz verstreute Vorlass-Materialien (wie Notizbücher, Bleistiftmanuskripte und Typskripte unterschiedlicher Textfassungen sowie Druckfahnen, aber auch annotierte Bücher, Fotos, Landkarten oder sogar Wanderstöcke) zum veröffentlichten Werk Peter Handkes zusammenführt. Die Materialien werden tabellarisch und in Form von Texten beschrieben und in eine entstehungschronologische Beziehung gesetzt, die in Paratexten kommentiert wird. Die werkgenetische Interpretation und Beschreibung erweitern das virtuelle Archiv zu einer Edition. Ergänzt wurde diese durch eine umfangreiche Bibliographie zur Primär- und Sekundärliteratur des über 140 Bücher sämtlicher Genres umfassenden Werks Handkes. Inhaltliche und formale Vielfalt öffnen die Plattform für ein breites Publikum. Sie wird nicht nur für die universitäre Forschung intensiv genutzt, sondern bietet Schüler*innen Unterstützung für Hausarbeiten, wird von Theatern für die Erarbeitung von Stückinszenierungen oder Begleitheften herangezogen und, wie die Aufregungen um den Nobelpreis gezeigt haben, auch von Journalist*innen konsultiert. (Der faktischen Materialdokumentation der Plattform kam etwa innerhalb der

kontroversen Debatte um Handkes "Serbien-Bücher" ein wichtige Rolle zu.)

Zentrales Anliegen des *Handkeonline*-Projekts war die Präsentation und Weitergabe des erarbeiteten bibliographischen, literaturwissenschaftlichen und archivalischen Wissens über Handke, die Entstehung seiner Werke, und damit zusammenhängend über die Beschaffenheit und Bedeutung der einzelnen Vorlass-Materialien. Die nachhaltige Modellierung dieser Forschungsdaten, welche erst eine langfristige Verfügbarkeit sowie Ausbau- und Anschlussfähigkeit des Wissens garantiert, wurde dabei vorderhand aus zeitökonomischen und budgetären Gründen vernachlässigt. Vor allem aber fehlte noch das geeignete konzeptuelle Modell das von Literatur-, Editions- oder Archiv- und Informationswissenschaft gemeinsam gedacht wird, und damit auch eine praktische Vorlage. Die Daten von *Handkeonline* wurden im Altsystem flach strukturiert und in einem proprietären Format vorgehalten. Die Folgen wurden bereits nach Projektende deutlich: die innovative Darstellung der Werkgenese war zwar Vorbild für andere Projekte, aber diese wurde weder in den Daten repräsentiert noch nachhaltig strukturiert. Deutlich zeigte sich das Problem in dem 2021, also nur sechs Jahre später begonnenen digitalen Editionsprojekt *Peter Handke Notizbücher* der Österreichischen Nationalbibliothek Wien und des Deutschen Literaturarchivs Marbach (<https://edition.onb.ac.at/handke-notizbuecher>): auf die wertvollen Daten über die Notizbücher konnte sogar intern nur mehr kompliziert über die dem Content-Management-System (Drupal) zugrunde liegende relationale Datenbank zugegriffen werden. Die damals getroffenen Zuordnungen der einzelnen Datenfelder konnten wegen des geschlossenen Datenformats teilweise nicht mehr vollständig rekonstruiert werden.

Aufgrund der Bedeutung von *Handkeonline* beschloss die Österreichische Nationalbibliothek die Daten-Migration in die erst später eingerichtete *Nachhaltige Infrastruktur für digitale Editionsprojekte an der Österreichischen Nationalbibliothek* (ÖNB-DE) und nützt diese Gelegenheit zum vollständigen Re-Design der Daten und Datenstruktur. Ziel ist die Transformation sämtlicher für *Handkeonline* generierte Beschreibungsdaten zum Werk Peter Handkes nach TEI/XML und die Bereitstellung eines werkgenetischen Katalogs als *Linked Open Data*-Datensatz. Diese Umsetzung liefert einen wesentlichen Beitrag zu einer Forschungskultur, die Vernetzung und Dynamik von Forschungsprozessen ins Zentrum rückt. Folgende Herausforderungen galt es dabei zu bewältigen:

- Einen Modellierungsansatz zu finden, mit dem präbibliographische und bibliographische Daten gemeinsam repräsentiert und in Beziehung gesetzt werden können.
- Die geschlossene in eine offene Werkgenese umzuwandeln und den Datenkatalog entsprechend den FAIR-Prinzipien (Findable, Accessible, Interoperable, Reuseable) zur Verfügung zu stellen.

Modelle für präbibliographische und bibliographische Daten

Zur Erschließung und Verbreitung von bibliographischen Daten haben sich im wissenschaftlichen Bereich verschiedene bibliographische Datenbanken (*Initiative for Open Citation* (i4oc.org)) etabliert, wie z.B. *Open Citation* oder *Google Scholar*, die sich allerdings auf die Erfassung bereits veröffentlichter Werke konzentrieren. Das gilt auch für eine Reihe weitverbreiteter Standardvokabulare für die grundlegenden Begriffe zur Dokumentation und Beschreibung bibliographischer Daten wie etwa *Dublin Core Element Set* und *Metadata Terms* (<https://www.dublincore.org/>) und *Schema.org*. Das *BIBFRAME*-Vokabular (<https://www.loc.gov/bibframe/docs/index.html>) wiederum ist ein vor allem im wissenschaftlichen Bereich gängiges (Austausch-)Format und man kann MARC- und RDA-Daten im Linked-Data-Standard RDF abbilden – dieses Vokabular ist jedoch ebenfalls ungeeignet für die Repräsentation präbibliographischer Werkmaterialien. Für das Projekt *Handkeonline* sind diese gängigen Vokabulare somit zu limitiert, da hier unveröffentlichte Werke und unterschiedlichste Werkvorstufen (Werkmaterialien) weder erfasst noch in ihrer teilweise komplexen Relationalität zu den publizierten Werken abgebildet werden können.

Eine differenzierte bibliographische Basiskodierung, wie sie Martina Gödel und Sebastian Zimmer (2017) auf der Basis der *Functional Requirements for Bibliographic Records* (FRBR) (Coyle 2016) für die Werke und vor allem den Zettelkasten von Niklas Luhmann gemacht haben (<https://niklas-luhmann-archiv.de/>), ist die Voraussetzung für weitere Forschungen zur Werkgenese. In dem von uns gewählten Ansatz spielt gerade dieser Aspekt deshalb von Beginn an eine Schlüsselrolle. Auch unsere Modellierung bibliographischer Daten basiert wie zuvor schon bei Stefanie Gehrke et al. (2016), Frederike Neuber (2016), Gödel und Zimmer (2017) oder Andreas Lüscho (2020) auf den Grundkonzepten von FRBR, doch wir gehen für unsere Anforderungen noch einen Schritt weiter:

Wir greifen, um die für unsere Abbildung der Werkgenesen notwendige Flexibilität (Werkdatendiversität und -referenzierbarkeit) bei gleichzeitiger Stabilität zu erreichen, die für eine Anschluss- und Ausbaufähigkeit der Daten notwendig ist, auf das im Museumsbereich gebräuchliche Format FRBR_{OO} zurück. FRBR_{OO} ist eine Harmonisierung des FRBR-*Entity-Relationship-Models* und des *Conceptual Reference Models* des *International Council of Museums CIDOC CRM* (<https://www.cidoc-crm.org/>). Letzteres wurde entwickelt, um die in der Museumsdokumentation relevanten Objekte und Konzepte (Ereignisse) zu beschreiben und institutionsübergreifend Daten austauschen zu können (Doerr, Crofts 2004, 15). FRBR_{OO} führt diese ereigniszentrierte Sichtweise nun auf Werke aller Art ein. Die Anwendung von FRBR_{OO} für die Modellierung von komplexen Werkgenesen ist, soweit wir die digitalen Editionen der letzten Jahre überblicken, neu; sie bringt viele Vorteile. FRBR_{OO} ermöglicht eine exakte Darstellung der komplexen Genese von Handkes Wer-

ken: Erstens, weil das Modell differenzierte Werkbegriffe bereitstellt, die uns erlauben, nicht nur die zahlreichen literarischen Genres, die Handke bedient (Prosa, Theater, Hörspiel, Film, Briefe, Interviews, Essays, Gedichte, Journale), sondern auch seine bislang unveröffentlichten, aber Werken entsprechenden Notizbücher zu formalisieren. Zweitens, weil FRBR_{OO} auf CIDOC-CRM basiert, ist es möglich auch individuelle, nichtbibliographische Materialien zu beschreiben. Drittens können mit FRBR_{OO} die Vorstufen eines Buches (*Manifestation Singleton*) von einer veröffentlichten Ausgabe (*Manifestation Product Type*) in den Daten klar unterschieden werden. Viertens schließlich erlaubt uns FRBR_{OO}, über das Ereignis *Expression Creation* die für eine Darstellung von Werkgenesen wichtige zeitliche und materielle Einordnung der einzelnen Fassungen.

Datenmodellierung von Handkeonline mit FRBRoo

Für die Modellierung von *Handkeonline* gilt es die vielfältigen Werkmaterialien aus dem Vorlass mit den publizierten Werken in Zusammenhang zu bringen. Peter Handkes Notizbüchern kommt dabei für die Genese seiner Werke ein besonderer Stellenwert zu: sie sind Materialsammlung oder Skizze für mehrere publizierte Werke; sie sind erste Textfassung für die gedruckte Notizauswahl in Journalen und darüber hinaus in ihrer Gesamtheit noch unveröffentlichtes Werk, das im Zuge des digitalen Editionsprojektes *Peter Handke Notizbücher* erstmals publiziert wird. Wir gehen deshalb von zwei unterschiedlichen Werkbegriffen aus: einmal von dem von Handke autorisierten veröffentlichten Werk mit seinen verschiedenen Vorstufen und einmal von einem von Handke verfassten, aber nicht als Werk autorisierten vorerst noch unveröffentlichten, fortlaufenden Werk – den Notizbüchern, die für sich stehen und zugleich Vorstufe sein können. Dabei geht es in der Reflexion des Werkbegriffs bei der Datenmodellierung immer um das Ausloten von Grenzen, hier zwischen Flexibilität (Diversität und Relationalität oder Referenzialität) und Stabilität.

Im Folgenden wird das neue Modellierungskonzept von *Handkeonline* erstens für veröffentlichte Werke und zweitens für Notizbücher exemplarisch skizziert. Die kursiven Ausdrücke und in eckigen Klammern angegebene Informationen beziehen sich auf die Spezifikationen laut dem FRBR_{OO}-Modell (Bekiari et al. 2015). Die Begriffe *Manifestation* und *Expression* bleiben unübersetzt (Arbeitsstelle für Standardisierung 2006).

Modellierung veröffentlichter Werke

Handkes Werke werden grundsätzlich über die Entität *Work* [1] erfasst. Die zugleich als Buch und Theaterstück oder Film publizierten Werke werden jedoch als *Complex Work* [F15] beschrieben, dessen Werkkomponenten (*Performance Work*, *Recording Work*, *Work*) über *has member* [R10] zueinander in Beziehung gesetzt werden. Zur Rekonstruktion der Werkgenesen wird jede Textfassungsversion (z.B. Vorfassung, Druckfahnen, Druck-

fassung etc.) der Buchpublikation als Entität vom Typ *Expression* [F2] angesehen. Die Erstausgabe wird als *Publication Expression* [F24] mit *Manifestation Product Type* [F3] modelliert.

Die anderen Textfassungsversionen ordnen sich relativ zu dieser chronologisch ein. Um die zeitliche Abfolge dieser einzelnen *Expression*-Entitäten in FRBR_{OO} beschreiben zu können, wird jeweils ein Ereignis *Expression Creation* [F28] definiert, das Angaben zum Entstehungszeitraum über die Entität *Time-Span* [E52] beinhaltet. Auf diese Art lässt sich die Werkhistorie als Abfolge von Ereignissen logisch rekonstruieren. Im Ereignis können auch Personen und Werke, die den Herstellungsprozess beeinflusst haben, repräsentiert werden. Es kann daher ein Unterschied zwischen Textfassungsversion und Notizbuch als selbständiges Werk in die Werkgenese integriert werden (*was influenced by* [P15]).

Alle *Expressions*, die das präbibliographische Werkmaterial repräsentieren, erzeugen (*created* [18]) Entitäten des Typs *Manifestation Singleton* [F4]. Bei Bedarf kann die *Expression* auch in einzelne *Expression Fragments* [F23] unterteilt werden, um die Textstadien in einer feineren Granularität abbilden zu können. Abbildung 1 veranschaulicht die eben besprochene werkgenetische Modellierung am Beispiel des 1981 als Buch erschienenen und 1982 bei den Salzburger Festspielen uraufgeführten Theaterstücks *Über die Dörfer* von Peter Handke.

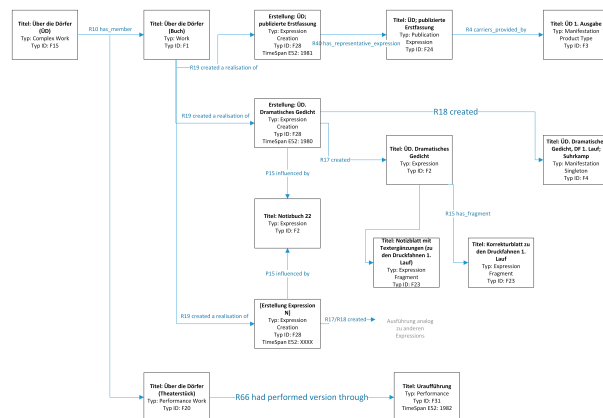


Abbildung 1: Beispielhafter Graph der werkgenetischen Beschreibung zum Werk *Über die Dörfer* von Peter Handke.

Modellierung Notizbücher und digitale Edition

Nicht veröffentlichte Notizbücher von Peter Handke werden im Modell durch die Entität *Individual Work* [F14] repräsentiert; die daraus realisierte *Self-Contained Expression* [F22] kann wiederum als Ereignis (*Expression Creation* [F28]) detailliert beschrieben werden. Schließlich wird das Notizbuch als *Manifestation Singleton* [F4] im Datensatz erfasst. Auf Werkebene werden die Notizbücher über *is logical successor* [R1] in Beziehung gesetzt. Für alle Notizbücher, die in der digitalen Edition *Handke Notizbücher* erstmals publiziert werden, werden Beziehungen auf Expression-Ebene (*is component of* [P148]) zur digitalen Edition hergestellt. Die digitale Edi-

tion wird als fortlaufendes Werk (*Serial Work* [F18]) repräsentiert, das kontinuierlich aktualisiert wird.

Vergleich zu *Handkeonline*

Im Vergleich zum Datenmodell von *Handkeonline* können mit dem vorgestellten System die Entitäten qualitativ miteinander in Beziehung gesetzt werden, was u.a. auch die erwähnte zeitliche Relation zulässt. Die ursprüngliche flache Struktur wird im neuen Modell durch ein Netzwerk repräsentiert, wodurch sich eine Bandbreite von neuen Auswertungsmöglichkeiten ergibt. Damit wird die Grundlage für ein durchsuchbares, interoperables Netzwerk geschaffen, welches das Auffinden von Ressourcen erleichtert und neue Forschungsfragen ermöglicht.

Ausblick

Das von uns entwickelte generische Datenmodell ist für spezifische Beschreibungen erweiterbar, die darin gefassten Daten frei nachnutzbar. Was mit *Handkeonline* begonnen wurde, kann in der Tiefenerschließung durch andere Forschungsprojekte erstmals umfassend nachgenutzt werden. Davon profitiert vor allem auch das Editionsprojekt *Peter Handke Notizbücher*.

Die hier präsentierte Modellierung unterstützt nicht nur das Verständnis der Werkgenesen von Peter Handkes Werken, sondern ist Vorbild für zukünftige DH-Projekte, die präbibliographische Daten zusammen mit bibliographischen Daten semantisch differenziert erfassen wollen.

Erst durch die hier besprochene Öffnung der Daten entsteht die Affordanz zur Repräsentation in vollem Umfang. Die Effekte sind vielfältig: Die Migration der isolierten Daten und deren Re-Modellierung nach anerkannten Standards machen das Material langzeitarchivfähig und schaffen die Grundlage für LOD. Durch den Einsatz von FRBR OO wird die Interoperabilität der vernetzten Informationen sichergestellt. Insgesamt werden durch die Öffnung neue Wege der Analyse geschaffen. Abfragen zu werkhistorisch relevanten Aspekten werden möglich: Wie die differenzierte Abfrage der Werkkomponenten oder zu aufführungsspezifischem Material oder sogar zu parallelen Entwicklungsprozessen; gleichzeitig wird auch eine Visualisierung von Stammbäumen zur Entstehung der Werke möglich. Wir erwarten mit Spannung, wie sich unsere Anwendung der formalen Semantik von FRBR OO in Zukunft auf die Reflexion über Werk und Werkgenesen auswirken wird.

Bibliographie

Arbeitsstelle für Standardisierung. 2006. *Funktionale Anforderungen an Bibliografische Datensätze. Abschlussbericht der IFLA Study Group on the Functional Requirements for Bibliographic Records*. Leipzig/Frankfurt a.M./Berlin: DNB. <https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr-deutsch.pdf> (zugegriffen: 01.08.2022).

Bekiari, Chrysoula et al. 2015. *FRBR object-oriented definition and mapping from FRBR ER, FRAD and FRSAD* (version 2.4). https://www.cidoc-crm.org/frbroo/sites/default/files/FRBRoo_V2.4.pdf (zugegriffen: 01.08.2022).

Coyle, Karen. 2015. *FRBR Before and After. A Look at Our Bibliographic Models*. Chicago: Ala Editions 2016. ISBN 978-0-8389-1364-2 (PDF). <http://192.248.73.38/bitstream/handle/94ous-1/926/RDA%20FRBR%20book%20for%20reading%20%20978-0-8389-1364-2.pdf> (zugegriffen: 01.08.2022).

Doerr, Martin und Patrick LeBoeuf. 2006. *FRBR object-oriented definition and mapping to FRBR ER (version 0.8.1). International Working Group on FRBR and CIDOC CRM Harmonisation*. https://archive.ifla.org/VII/s13/wg-frbr/FRBR_oo_V.0.8.1c.pdf (zugegriffen: 01.08.2022).

Gehrke, Stefanie, Pauline Charbonnier und Frunzeanu Eduard. 2016. „Biblissima - Semantic Web Application für Handschriften, Inkunabeln und historische Sammlungen - Zwischenbericht“ In *DHd 2016: Modellierung - Vernetzung - Visualisierung: Die Digital Humanities als fächerübergreifendes Forschungsparadigma*. <https://zenodo.org/record/4645060#.YuefAhzP2Uk> (zugegriffen: 01.08.2022).

Goedel, Martina und Sebastian Zimmer. 2017. „Niklas Luhmanns Werk- und Leskosmos - DH in der bibliographischen Dimension“ In *DHd 2017: Digitale Nachhaltigkeit*. <https://dh-abstracts.library.cmu.edu/works/10641> (zugegriffen: 01.08.2022).

Lüschow, Andreas. 2020. „Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane“ In *DHd 2020: Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. <https://zenodo.org/record/4621704#.YuedhRzP2Uk> (zugegriffen: 01.08.2022).

Neuber, Frederike. 2016. „Stefan George Digital“ In *DHd 2016: Modellierung - Vernetzung - Visualisierung: Die Digital Humanities als fächerübergreifendes Forschungsparadigma*. <https://zenodo.org/record/4645060#.YuefAhzP2Uk> (zugegriffen: 01.08.2022).

Offenheit durch
Dokumentation: Lose
Forschungsfäden im
"Online-Compendium
der deutsch-
griechischen
Verflechtungen"
zusammenführen

Soethaert, Bart

bart.soethaert@fu-berlin.de

Freie Universität Berlin, Deutschland

Pechlivanos, Miltos

m.pechlivanos@fu-berlin.de
Freie Universität Berlin, Deutschland

Das *Online-Compendium der deutsch-griechischen Verflechtungen* (ComDeG) ist ein laufendes Forschungs- und Publikationsprojekt im Open Access, initiiert und konzeptualisiert durch das Centrum Modernes Griechenland (CeMoG) unter Federführung der Professur Neogräzistik (FU Berlin).¹ Das ComDeG umfasst zum einen, in Kooperation mit dem Institut für griechisch-deutsche Beziehungen (EMES) der Nationalen und Kapodistrias-Universität Athen, wissenschaftliche Essays und Fallanalysen (Mikrogeschichten, Makrovorgänge, Metanarrative und Präsentationen)² sowie enzyklopädische Artikel und tabellarische Biogramme zu Akteuren der deutsch-griechischen Verflechtungen,³ die die deutsch-griechische Geschichte seit dem ausgehenden 18. Jahrhundert als europäisches Aktionsfeld transnationaler Interaktionen, Interpretationen, Übersetzungen und Transfers ausloten und erschließen. Zum anderen beinhaltet das dynamisch verknüpfte Informationsangebot die CeMoG-Wissensbasis mit angereicherten Indexeinträgen zu Personen und Institutionen, Wirkungs-orten, Kontaktzonen und Vermittlungspraktiken⁴ sowie bibliographische Sammlungen mit u.a. Forschungsliteratur zu den deutsch-griechischen Verflechtungen, zu deutsch-griechischen und griechisch-deutschen Übersetzungen,⁵ die ebenfalls mit allen Inhaltsbereichen des ComDeG verknüpft wurden.

Am kollaborativen Aufbau der Inhalte dieses multiperspektivischen Online-Sammelwerks ist ein breitgefächertes Netzwerk von Forscher:innen, primär aus Deutschland und Griechenland, beteiligt. Die Inhaltserstellung für das Compendium erfolgt auf der Grundlage einschlägiger Workshops, wozu Forscher:innen eingeladen werden, ihre Fachexpertise einzubringen, Fallgeschichten mit weiteren Expert:innen zu diskutieren und die wissenschaftlichen Erträge zur Veröffentlichung in das ComDeG aufzubereiten.⁶ Weitere Beiträge stammen von Forschenden an deutsch- bzw. griechischsprachigen Universitäten oder Forschungsinstitutionen mit einschlägiger Fachkompetenz wie etwa Instituten der Germanistik, der Neogräzistik und der Südosteuropa-Geschichte.⁷ Darüber hinaus bietet sich das ComDeG als geeignetes Repositorium für die (Teil-)Veröffentlichung von Forschungsergebnissen (etwa in der Form von Bibliographien, biographischen Profilen, enzyklopädischen Lemmata oder Mikropublikationen), die eine ähnliche thematische und methodische Perspektivierung vornehmen.⁸

Ein internationaler wissenschaftlicher Beirat steuert gemeinsam mit den beiden Herausgebern des Compendiums, Prof. Dr. Miltos Pechlivanos (Freie Universität Berlin) und Prof. Dr. Alexandros-Andreas Kyrtis (Nationale und Kapodistrias-Universität Athen) die inhaltliche Entfaltung und prüft die Qualität aller wissenschaftlichen Beiträge.⁹ Das ComDeG-Redaktionsteam ist hauptverantwortlich für die redaktionelle Aufbereitung und Vernetzung aller Inhalte, kuratiert und ergänzt die Datensammlungen (Bibliographie und CeMoG-Wissensbasis)

und koordiniert die Übersetzung aller Compendium-Inhalte, die sowohl in deutscher als auch in griechischer Sprache veröffentlicht werden.¹⁰ Neue Inhalte werden laufend hinzugefügt.

Ein freizugängliches Informationsangebot und Recherche-tool

Das ComDeG versteht sich als eine Brücke der Informationsvermittlung, der Zusammenarbeit und der Vernetzung, die darauf abzielt, eine gemeinsame deutsch-griechische Geschichtskultur zu ermöglichen. Seine Inhalte richten sich an eine möglichst breite deutsche und griechische Öffentlichkeit, die daran interessiert ist, ihr Wissen über die politischen, gesellschaftlichen und wirtschaftlichen bis hin zu wissenschaftlichen und kulturellen Verflechtungen vom ausgehenden 18. Jahrhundert bis in die jüngste Vergangenheit zu erweitern. Seit September 2020 steht der deutsch-griechischen Fachcommunity und der interessierten Öffentlichkeit auf comdeg.eu ein qualitätsgesichertes Informationsnetzwerk mit wissenschaftlichen Beiträgen zu deutsch-griechischen Verflechtungen und (bibliographischen und prosopographischen) Forschungsdaten zu deren historischen Akteuren im Open Access zur Verfügung. Die Register der CeMoG-Wissensbasis sowie die erweiterten Suchfunktionen in den bibliographischen Sammlungen unterstützen das Auffinden von passenden Inhalten im gesamten ComDeG¹¹ und ermöglichen eine personenbezogene bzw. thematisch eingegrenzte Recherche in den bereitgestellten Publikationen.

Das ComDeG verknüpft disparates Wissen, stellt überblickende Zusammenstellungen bereit und macht fokussierte Fallstudien für neue Fragestellungen anschlussfähig. Es bildet nicht nur ein breitgefächertes Forschungsnetzwerk, das Wissenschaftler:innen aus unterschiedlichen Disziplinen in ein Gemeinschaftsprojekt zusammenbringt, sondern generiert zugleich eine öffentlichkeitswirksame Sichtbarkeit für seine Forschungserträge. Es trägt wesentlich dazu bei, das Forschungsfeld der deutsch-griechischen Verflechtungen in seiner Größe und Vielfalt zu vermessen und seinen Untersuchungsstand zu dokumentieren. Avisierte Publikums- und Nutzer:innengruppen sind nicht nur Forscher:innen, denen das ComDeG als heuristisches Forschungsinstrument und domänenspezifische Publikationsplattform dient, sondern auch Lehrende und Studierende, Journalist:innen, Kulturarbeiter:innen und sonstige Interessierte (Soethaert 2020).

Offenheit als konzeptionelle Anforderung an die Entwicklung des ComDeG

Der öffentliche Angebotscharakter einer geisteswissenschaftlichen Online-Publikation im Open Access, wie das ComDeG, wird meistens unter vier Gesichtspunkten

thematisiert bzw. konkretisiert (vgl. AG Digitales Publizieren 2021, §§ 58-68 und §§ 79-81; Kleineberg und Kaden 2017; Open Knowledge Foundation 2015; Wissenschaftsrat 2022, 38-47): Zugänglichkeit und Nachnutzbarkeit (die Online-Veröffentlichung der Compendium-Inhalte unter der Open Access-Lizenz CC BY-NC-ND 4.0),¹² Auffindbarkeit (die Zitierbarkeit aller Inhaltsbereiche, der Einsatz von URLs und normierten Deskriptoren für alle interne Links, die Berücksichtigung von Suchmaschinenoptimierungs-Faktoren für wissenschaftliche Inhalte, vgl. Putnings 2017; Schilhan 2020; Schilhan, Kaier und Lackner 2021) und Gebrauchstauglichkeit (etwa durch die Bereitstellung von erweiterten Such- und Filteroptionen in allen Registern sowie die Einrichtung einer Zeitleiste als Facettensuche für die Compendium-Inhalte, vgl. Russell-Rose und Tate 2013; Tunkelang 2009).

Dem öffentlichen Start des ComDeG ging ein mehrjähriger Design- und Entwicklungsprozess voraus, in dem Offenheit als konzeptionelle Anforderung auch alle Aspekte in der informationstechnischen Konzeption des ComDeG als lebhaftes Publikationsprojekt anbelangte, sei es im Bereich der inkrementellen Inhaltserstellung und -publikation, in der Verwaltung, Anwendung und Ergänzung der Verschlagwortungsvokabulare (in der Regel mit GND-ID), oder in der Integration von fortwährend mit Zotero verwalteten bibliographischen Daten.¹³ Angesichts der angestrebten netzartigen Fortschreibung des Publikationsprojekts wurde schnell deutlich, dass das ComDeG sich nicht auf die Implementierung eines digitalen Äquivalents zu einer herkömmlichen Printpublikation mit vorstrukturierten Inhaltsanordnungen und festgeschriebenen Bezugnahmen einschränken ließe; vielmehr sollten in der digitalen Mediatisierung der lebhafteste Charakter des ComDeG sowie die Bildung und die fortschreitende Ansammlung seiner inneren Verbindungen Rechnung getragen werden.

Der konzeptionelle und funktionelle Designprozess des ComDeG sah sich in diesem Kontext mit einer praktischen Herausforderung konfrontiert: Wie kann die Offenheit vor dem Hintergrund eines stetigen Inhaltswachstums als konzeptionelle Anforderung längerfristig und ressourcenschonend ausgetragen werden, ohne dass neu hinzukommende Inhalte die redaktionelle Überarbeitung bestehender Informationsarrangements verursachen würden? Anders gesagt: Wie kann man im Verlauf dieses ‚Work-in-Progress‘ die Möglichkeit bereithalten, den einzelnen Beiträgen graduell mehr Kontext hinzuzufügen, ohne jeden einzelnen Beitrag neu editieren bzw. immer wieder weiter verknüpfen zu müssen? Wie können Verbindungen zu fokussierten Beiträgen angelegt werden, die (noch) nicht existieren? Und, nicht zuletzt, wie ermöglicht das ComDeG, dass lose Forschungsfäden in den dokumentierten Informationen an Komplexität und Kontext gewinnen können, indem der interessengeleitete Spürsinn anderer Forscher:innen auf anderweitige Informationscluster gelenkt wird (vgl. Krämer 1998, 79; 2007, 18-19)?

Mit dem Open Encyclopedia System (OES), das 2016-2020 parallel zum ComDeG am Center für Digitale Systeme der Freien Universität Berlin als standardisierte Plattform zur Erstellung, Publikation und Pflege von lemmabasierten Sammelwerken konzipiert und erstentwickelt wurde,¹⁴ bot sich die besondere Chance an, die

Leitprinzipien der OES-Systemarchitektur (Modularität, Offenheit, Datenintegrität und Schnittstellen; vgl. Apostolopoulos, Schimmel und Egilmez 2017, 382) auch im Hinblick auf ein wissens- und erkenntnisorientiertes Datendesign für das ComDeG zu prüfen und produktiv zu machen. Denn durch die Beteiligung des Centrum Modernes Griechenland am Projektkonsortium für die Entwicklung des OES ergab sich die Gelegenheit, einige Module (wie etwa die Bibliographie und die Verschlagwortung) neu zu überdenken und dabei die Offenheit, anders als primär auf Open Access, Open Content, Open Licenses und Open Data ausgerichtet, auch als tragfähiges Design- und Verknüpfungsprinzip aufzugreifen.

Funktionale Leistungseigenschaften der CeMoG-Wissensbasis

Die Funktionalisierung und Implementierung der Offenheit macht das besondere informationstechnische Merkmal des ComDeG aus und versetzt die Plattform in die Lage, ihr Potenzial nicht nur als Medium einer domänenspezifischen Wissenskommunikation sondern vor allem auch als heuristisches Forschungsinstrument für neue kontextbezogene Wissensgenerierung zu entfalten. Denn, im Vergleich zu anderen, bereits veröffentlichten OES-Anwendungen, in denen die jeweiligen Artikel *indexiert* werden,¹⁵ zeichnet sich das ComDeG durch die Verknüpfung *aller* Segmente der Plattform untereinander (Bibliographie, Wissensbasis, Compendium) aus. Die CeMoG-Wissensbasis bildet das strukturelle Rückgrat des ComDeG: ihre Einträge stellen nicht nur ein offenes aber kuratiertes Verschlagwortungsvokabular zu Personen und Institutionen für die bibliographischen Sammlungen sowie für die Essays und Artikel des Compendiums bereit, sondern werden durch neu hinzukommende Inhalte stetig um *neue* Einträge ergänzt. Darüber hinaus verstehen die Einträge der CeMoG-Wissensbasis sich als Datenverknüpfungsobjekte, die diverse Verweise auf ComDeG-Inhalte an einer Stelle darstellen und Lesepfade zwischen ihren verschiedenen Kontexten etablieren. Sie sind mit anderen Worten eigenständige Datenobjekte, die mit Attributen angereichert werden und über Relationen in Beziehung zu anderen Datenobjekten (wie etwa Artikel, Essays, Biogramme, bibliographische Datensätze) stehen.

Die Vorteile der Einrichtung solcher stabil adressierbaren und dennoch offen in ihren Verknüpfungen OES-Objekte machen sich leicht bemerkbar. Kommt ein neuer Essay oder Artikel hinzu, bildet deren Verschlagwortung *mit bestehenden Einträgen* eine implikative und kontextuelle Beziehung zu *anderen*, bereits damit verknüpften ComDeG-Inhalten, ohne dass an der Stelle umgekehrt ein expliziter Verweis angelegt werden muss. Das Datenverknüpfungsobjekt dokumentiert und bündelt in einer übersichtlichen Darstellung alle angelegten *und* neu hinzukommenden Verbindungen. Es erfüllt die wichtige Voraussetzung, dass Informationen zusammenkommen bzw. miteinander kombiniert werden können, um neue Erkenntnisse zu generieren. Kommt durch die Aufnahme eines neuen Essays oder Artikels auch ein

neuer Eintrag zu einer Person oder Institution zu Stande, wird dieser ebenfalls mit biographischen Kurzinformati-
onen (und womöglich mit einer GND-Referenzierung) so-
wie mit relevanten Verweisen auf (bestehende oder neue)
Einträge in den bibliographischen Sammlungen ausge-
stattet.

Die Einträge der CeMoG-Wissensbasis stellen folglich
durch die Verschlagwortung in den jeweiligen Segmen-
ten immer schon vernetzte Datenobjekte dar, die expli-
zit gemachte Beziehungen abbilden und/oder lose For-
schungsäden aus den jeweiligen ComDeG-Segmenten
festhalten. Gerade weil sie aus mehreren Kontexten her-
aus erreichbar sind und die verfügbaren Verweise jeweils
an einer Stelle akkumulieren, können sie andererseits
auch auf diverse Zusammenhänge hinweisen. Das ist
ihre Kernleistung. Infolgedessen (und anders als in gän-
gigen Indexierungsverfahren) verfügen diese eigenstän-
digen OES-Objekte über die Kapazität, auf etwaige Infor-
mationslücken im Compendium hinzuweisen, z.B. wenn
in den einzelnen Darstellungen zu der betroffenen Per-
son oder Institution (noch) kein Verweis auf einen ent-
sprechenden Artikel bzw. weiterführenden Essay vorliegt.

The screenshot shows the 'WISSENSBASIS' section of the 'Online-Compendium der deutsch-griechischen Verflechtungen'. It features a navigation bar with links to 'Compendium', 'Wissensbasis', 'Bibliographie', 'Archiv', 'Über uns', and 'Griechisch'. Below the navigation bar, the entry for 'Alexander Steinmetz' is displayed, including his birth and death dates (1879 (Hamel) - 1973 (Traunstein)). A table titled 'Nachweise im Compendium' lists various references, including articles and essays. Below this, a section titled 'Nachweise in der Bibliographie' shows a list of entries with their respective counts: 'DEUTSCH-GRIECHISCHE VERFLECHTUNGEN' (11 Einträge), 'DEUTSCH-GRIECHISCHE ÜBERSETZUNGEN' (5 Einträge), 'GRIECHISCH-DEUTSCHE ÜBERSETZUNGEN' (62 Einträge), 'GRIECHISCHSPRACHIGE LITERATUR' (24 Einträge), and 'DEUTSCHSPRACHIGE LITERATUR' (45 Einträge).

Abb. 1 Der Eintrag zu Alexander Steinmetz in der CeMoG-Wissensbasis¹⁶

Die zuverlässigen und beständigen Identifizierungen
der CeMoG-Wissensbasis sichern dem ComDeG sei-
nen nachhaltigen Ausbau aus redaktionstechnischer
Sicht zu, ermöglichen die sukzessive Netzwerkbildung
affiner Inhalte und eröffnen nicht zuletzt im Frontend
nicht-lineare Navigationswege bzw. Lesepfade durch die
Segmente des ComDeG. Die Einträge der CeMoG-Wis-
sensbasis signalisieren nicht zuletzt über ihre noch aus-
stehenden Nachweise, dass wir nicht über ein (alles

umfassendes) *Online-Compendium der deutsch-griechi-
schen Verflechtungen* verfügen, sondern immer daran
arbeiten. Es handelt sich um ein kollaboratives Unterfan-
gen, das ein solches Compendium als Möglichkeit ver-
handelt, als Perspektive weiterentwickelt und als Projekt
fortschreibt, ohne den synthetisierenden Punkt jeweils
erreichen oder alle Erwartungen vollumfänglich erfüllen
zu können.

Das narrative Vorgehen des ComDeG orientiert sich an
einem Modus fragmentarischer Geschichtsschreibung,
der nach einem Montageprinzip Momente und Konstella-
tionen der deutsch-griechischen Verflechtungen auf-
schließt, bestimmte Navigationspunkte (wie etwa Perso-
nen, Institutionen, Medien, Objekte, Orte, Kontaktzonen
und Vermittlungspraktiken) dichter Beschreibungen
unterzieht, ohne mit den Narrativisierungen von Mikro-
geschichten, Makrovorgängen und Metanarrativen ein
einheitliches oder erschöpfendes Bild anzustreben (vgl.
Pechlivanos 1995; Büttner und Kim 2022). Die Essays
des ComDeG erzählen Geschichten (*stories*) über Ge-
schichte (*history*); sie stellen als Textkorpus eine An-
sammlung von einzelnen Geschichten oder Zusammen-
führungen dar, die keinesfalls einen abgeschlossenen
Sinn ergeben.

Den originellen Beitrag, den das Centrum Modernes
Griechenland mit der Integration der CeMoG-Wissens-
basis in das ComDeG für die Indexierung von dyna-
mischen Artikelbeständen mit dem Open Encyclopedia
System geleistet hat, liegt einerseits in der „strukturellen
Ausdifferenzierung des Publikationsobjektes in unter-
schiedlich verarbeitbare und aktualisierbare Teile“ (vgl.
Kaden 2016, 19), andererseits in der Ausarbeitung und
Implementierung eines tragfähigen Konzepts für die
fortwährende Informationsorganisation und die prakti-
sche Ausgestaltung jener Leistungseigenschaften sol-
cher Online-Sammelwerken, die sie als Werkzeug für die
Forschung nutzbar machen. Das ComDeG ist weniger
ein Compendium (im Sinne eines Handbuchs) als ein
Vektor, ein vorwärts gerichtetes Publikationsprojekt, das
Änderungen und Ergänzungen offen gegenübersteht,
nicht abgeschlossen ist und nie einen fertigen, in sich ab-
geschlossenen Zustand erreichen wird.

Fußnoten

1. <https://comdeg.eu/>. Rollen der Beiträger zu diesem Vortrag (nach der CRediT-Taxonomie): Prof. Dr. Mil-
tos Pechlivanos (Conceptualization), Dr. Bart Soetha-
ert (Conceptualization, Writing – original draft, review &
editing). Alle in den Fußnoten angegebenen Links wur-
den am 14. Dezember 2022 abgerufen.
2. <https://comdeg.eu/compendium/essays/>
3. <https://comdeg.eu/compendium/artikel/>
4. <https://comdeg.eu/wissensbasis/>
5. <https://comdeg.eu/bibliographie/>
6. <https://www.cemog.fu-berlin.de/compendium/aktivitaeten/>
7. <https://comdeg.eu/compendium/autorinnen/>
8. [https://www.cemog.fu-berlin.de/compendium/mit-](https://www.cemog.fu-berlin.de/compendium/mitmachen/und)
[machen/und https://comdeg.eu/projekt/mitmachen/](https://comdeg.eu/projekt/mitmachen/).
In diesem Zusammenhang ist insbesondere auf den
signifikanten Informationszuwachs hinzuweisen, der
bisher im Compendium durch Kooperationen mit the-

matisch affinen Projekten erreicht wurde. Hierzu zählen mitunter das Projekt von Prof. Dr. Ulrich Moennig (Universität Hamburg) zu „Griechischen Doktorand(in)en an der Universität Hamburg von der Gründung der Universität 1919 bis 1941“ (<https://comdeg.eu/compendium/essay/101014/>) und das Projekt „Akteure deutsch-griechischer Übersetzungskulturen: eine Standortbestimmung aus kollektiv-biographischer Perspektive“, das am Centrum Modernes Griechenland der Freien Universität Berlin in Zusammenarbeit mit dem Institut für Griechisch-Deutsche Beziehungen der Nationalen und Kapodistrias-Universität Athen erarbeitet wurde: <https://comdeg.eu/compendium/essay/103006/>.

9. <https://www.cemog.fu-berlin.de/compendium/beteiligte/>

10. <https://www.cemog.fu-berlin.de/compendium/koordination/>

11. Um eine optimale Durchsuchbarkeit aller Segmente des ComDeG zu gewährleisten, wurde die frei verfügbare Suchtechnologie Apache Solr (<https://solr.apache.org/>) integriert.

12. Vgl. die „Open-Access-Grundlagen“ auf der Plattform open-access.network sowie die Open-Access-Definition in der BOAI-Grundsatzerklärung: „There are many degrees and kinds of wider and easier access to [research] literature. By ‚open access‘ to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.“ (Budapest Open Access Initiative 2002)

13. Als Werkzeug für die strukturierte Erfassung bibliographischer Daten wird das freie, quelloffene Literaturverwaltungsprogramm Zotero (<https://www.zotero.org/>) eingesetzt. Für die Einbindung der Zotero-Sammlungen in das OES-Redaktionssystem wurde eine Programmierschnittstelle (API) mit lesendem und schreibendem Zugriff auf die bibliographischen Datensammlungen eingerichtet sowie ein User-Interface, das, auf dieser Schnittstelle aufbauend, den Datenaustausch zwischen den Bibliographien und dem Redaktionssystem dokumentiert. Die bibliographischen Daten, die mit Normeinträgen zu Personen und Institutionen verknüpft werden können und über Relationen in Beziehung zu Compendium-Inhalten stehen, werden in einem eigens dafür konzipierten Bereich „Bibliographie“ ausgeliefert und verfügen über umfassende Such- und Filterfunktionen.

14. Die OES-Systemarchitektur und -Komponenten basieren auf dem Open Source Content Management System WordPress (<https://wordpress.com/de/>) und wurden als WordPress-Plugin mit zugehörigem Theme und einem projektspezifischen Datenmodell, über welches die Beitragstypen, weitere OES-Datenobjekte sowie die Relationen zwischen den Datenfeldern definiert werden, umgesetzt. Für die Implementierung von benutzerdefinierten Eingabemasken und Beziehungsfeldern nutzt OES das Open Source Plugin Advanced Custom Fields (ACF, <https://www.advancedcustomfields.com/>). Der Quellcode von OES steht zur akademischen Verwendung und Erweiterung durch Dritte unter einer GPLv2-Lizenz auf GitHub zur Verfügung: . Die Erst-

entwicklung von OES erfolgte im Rahmen des von der DFG geförderten Projekts „Von 1914-1918-online zum Open Encyclopedia System“ (2016-2020): . OES wird kontinuierlich weiterentwickelt, nicht zuletzt in enger Kooperation mit dem Exzellenzcluster „Temporal Communities. Doing Literature in a Global Perspective“. Betrieb und Pflege der Redaktions- und Publikationsumgebung für das ComDeG wird durch die Freie Universität Berlin gesichert.

15. <https://www.open-encyclopedia-system.org/use-cases/>

16. <https://comdeg.eu/namennormdatei/nnd-alexandersteinmetz-2/>

Bibliographie

AG Digitales Publizieren, Hg. 2021. "Digitales Publizieren in den Geisteswissenschaften: Begriffe, Standards, Empfehlungen." In *Zeitschrift für digitale Geisteswissenschaften / Working Papers*, 1. Wolfenbüttel. https://doi.org/10.17175/wp_2021_001 (zugegriffen: 14. Dezember 2022).

Apostolopoulos, Nicolas , Christoph Schimmel und Ilker Egilmez. 2017. "Open Encyclopedia System. Open Source Software for Open Access Encyclopedias." In *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium for Information Science (ISI 2017)*, hg. von Maria Gäde, Violeta Trkulja und Vivien Petras, 380-385. Berlin: Werner Hülsbusch. http://isi2017.i-b.hu-berlin.de/ISI_17_ONLINE_FINAL.pdf (zugegriffen: 14. Dezember 2022).

Bardi, Alessia und Paolo Manghi. 2014. "Enhanced Publications: Data Models and Information Systems." *Liber Quarterly* 23(4), 240-273. <https://doi.org/10.18352/lq.8445> (zugegriffen: 14. Dezember 2022).

Budapest Open Access Initiative. 2002. "Declaration." 14. Februar 2002. <https://www.budapestopenaccessinitiative.org/read/> (zugegriffen: 14. Dezember 2022).

Büttner, Urs und David D. Kim. 2022. "Globalgeschichten der Literaturen. Ein Methodenprogramm." In *Globalgeschichten der deutschen Literatur. Methoden – Ansätze – Probleme*, hg. von Urs Büttner und David D. Kim, 1-32. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05786-0_1 (zugegriffen: 14. Dezember 2022).

Kaden, Ben. 2016. "Zur Epistemologie digitaler Methoden in den Geisteswissenschaften." In *Berliner Beiträge zu Digital Humanities*. <https://doi.org/10.5281/zenodo.50623> (zugegriffen: 14. Dezember 2022).

Kleineberg, Michael und Ben Kaden. 2017. "Open Humanities? ExpertInnenmeinungen über Open Access in den Geisteswissenschaften." *LIBREAS. Library Ideas* 32: 1-31. <https://doi.org/10.18452/19096> (zugegriffen: 14. Dezember 2022).

Krämer, Sibylle. 1998. "Das Medium als Spur und als Apparat." In *Medien, Computer, Realität. Wirklichkeitsvorstellungen und Neue Medien*, hg. von Sibylle Krämer, 73-94. Frankfurt am Main: Suhrkamp. <https://www.geisteswissenschaften.fu-berlin.de/we01/institut/mitarbeiter/emeriti/kraemer/PDFs/>

Aufsätze/Das-Medium-als-Spur-und-Apparat-1998-_50_.pdf (zugegriffen: 14. Dezember 2022).

Krämer, Sibylle. 2007. "Was also ist eine Spur? Und worin besteht ihre epistemologische Rolle? Eine Bestandsaufnahme." In *Spur. Spurenlesen als Orientierungstechnik und Wissenskunst*, hg. von Sibylle Krämer, Werner Kogge und Gernot Grube, 11-33. Frankfurt am Main: Suhrkamp. https://www.geisteswissenschaften.fu-berlin.de/we01/institut/mitarbeiter/emmeriti/kraemer/PDFs/Aufsätze/Was-also-ist-eine-Spur-2007-_113_.pdf (zugegriffen: 14. Dezember 2022).

Kyrtis, Alexandros-Andreas und Miltos Pechlivanos, Hgg. 2020-. *Compendium der deutsch-griechischen Verflechtungen*. <https://comdeg.eu/compendium/> (zugegriffen: 14. Dezember 2022).

Open Knowledge Foundation. 2015. "Open Definition 2.1." In *The Open Definition. Defining Open in Open Data, Open Content and Open Knowledge*. <https://opendefinition.org/od/2.1/en/> (zugegriffen: 14. Dezember 2022).

Pechlivanos, Miltos. 1995. "Literaturgeschichte(n)." In *Einführung in die Literaturwissenschaft*, hg. von Miltos Pechlivanos, Stefan Rieger, Wolfgang Struck und Michael Weitz, 170-181. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-03544-8_15 (zugegriffen: 14. Dezember 2022).

Putnings, Markus. 2017. "6g. Die Rolle der Metadaten - Indexierung und Sicherung der Auffindbarkeit." In *Praxishandbuch Open Access*, hg. von Konstanze Söllner und Bernhard Mittermaier, 311-320. Berlin: De Gruyter. <https://doi.org/10.1515/9783110494068-036> (zugegriffen: 14. Dezember 2022).

Russel-Rose, Tony und Tyler Tate. 2013. "Chapter 7 - Faceted Search." In *Designing the Search Experience. The Information Architecture of Discovery*, 167-218. Waltham, MA: Morgan Kaufmann. <https://doi.org/10.1016/B978-0-12-396981-1.00007-0> (zugegriffen: 14. Dezember 2022).

Schilhan, Lisa. 2020. "Sichtbarkeit und Auffindbarkeit wissenschaftlicher Publikationen." In *Publikationsberatung an Universitäten: Ein Praxisleitfaden zum Aufbau publikationsunterstützender Services*, hg. von Karin Lackner, Lisa Schilhan und Christian Kaier, 237-258. Bielefeld: transcript Verlag. <https://doi.org/10.1515/9783839450727-013> (zugegriffen: 14. Dezember 2022).

Schilhan, Lisa, Christian Kaier und Karin Lackner. 2021. "Increasing Visibility and Discoverability of Scholarly Publications with Academic Search Engine Optimization." *Insights* 34(6), 1-16. <http://doi.org/10.1629/uksg.534> (zugegriffen: 14. Dezember 2022).

Soethaert, Bart. 2020. "Neue Recherche- und Publikationsplattform für die historische Erforschung der deutsch-griechischen Beziehungen. Launch der digitalen Plattform für das *Online-Compendium der deutsch-griechischen Verflechtungen* (ComDeG)." In *L.I.S.A. Wissenschaftsportal der Gerda Henkel Stiftung*. 2. Oktober 2020. https://lisa.gerda-henkel-stiftung.de/?nav_id=9436 (zugegriffen: 14. Dezember 2022).

Tunkelang, Daniel. 2009. *Faceted Search*. San Rafael, CA: Morgan & Claypool. <https://doi.org/10.1007/978-3-031-02262-3> (zugegriffen: 14. Dezember 2022).

Wissenschaftsrat, Hg. 2022. *Empfehlungen zur Transformation des wissenschaftlichen Publizierens zu Open Access*. Köln. <https://doi.org/10.57674/fyrc-vb61> (zugegriffen: 14. Dezember 2022).

Open and Closed AI. Eine Kunstkritik künstlich generierter Bilder

Bell, Peter

peter.bell@uni-marburg.de
Philipps-Universität Marburg, Deutschland

Einleitung

Unter dem Hashtag #aiart erscheinen täglich tausende künstlich erzeugte Bilder auf sozialen Medien und digitalen Marktplätzen z.B. als NFTs, flankiert von Kommentaren und Artikeln, welche die Qualität und Kreativität dieser Bildproduktionen diskutieren.

Der Vortrag reflektiert diese Entwicklungen, indem er die Bilder und ihre Genese kritisch hinterfragt, wertet und die Prozesse analysiert. Dass Digital Humanities parallel zum wissenschaftlichen Rezensionswesen auch eine Kritik von Software entwickeln sollte, zeigt sich z.B. in der Gründung des CKIT Rezensionsjournal das aus dem AK digitale Kunstgeschichte und nfdi4culture entstanden ist. Im Zusammenhang generativer künstlicher Intelligenz weitet sich diese Softwarekritik in Form einer Werkkritik aus. Aufgrund der raschen Entwicklung innerhalb der Computer Vision und der generativen Modelle gibt es bislang wenige Reflexionen über deren Schöpfungshöhe (Gary et al. 2022) jenseits der informatischen Überbietungslogiken und der stetigen Diskussion über bias von AI.

OpenAI und closed access. Technikkritik

Digitale Kunstkritik setzt bereits ein, wo die Frage gestellt wird, ob dieses popkulturelle Phänomen Gegenstand der Kunstgeschichte oder der Medienwissenschaften ist oder – wie öffentlich viel diskutiert – es sich überhaupt um Kunst handeln kann. Die Frage ist leicht zu beantworten, denn in dem Moment, wo über Kunst oder nicht Kunst entschieden werden muss, bedarf es unvermeidlich der Methoden der Kunstgeschichte und Kunstkritik. #Aiart könnte auf einer rein ästhetisch-visuellen Ebene mit konventionellen Methoden der Kunstwissenschaft bewertet werden. Dadurch, dass die Bilder nicht von Menschen sondern in erster Linie von einem mathematischen Modell erzeugt werden, bietet

sich an eine operative und technikzentrierte Medienkritik in die Kunstkritik zu integrieren. Die hier durchzuführende kritische Reflexion über AI Art findet somit in der digitalen Kunstgeschichte als Teil der Digital Humanities statt. Dies geschieht in Form Medienarchäologie, in der das maschinelle Lernen und die verschiedenen Modelle mit- und gegeneinander arbeitender neuronaler Netze untersucht werden. Dabei zeigt sich eine Vielzahl an Architekturen von GANs und Transformer-Netzwerken¹ deren Ergebnisse teils zusätzlich noch durch verschiedene Trainingsdatenbanken² variiert werden können. Die 2014 mit generative adversarial networks (GAN) begonnene Entwicklung hat sich in den letzten drei Jahren deutlich beschleunigt und zu immer realistischeren Bildern geführt. Während einige Modelle selbstständig Bilder erzeugen, ist die Eingabe via Bildbeschreibung quasi als Kommando (prompt) zum Standard der künstlichen Bildproduktion geworden. Diese Texteingabe ist möglich, da in DALL-E ein Sprachmodell (GPT-3) mit Algorithmen zur Bilderzeugung kombiniert ist. Auch durch dieses Zusammenkommen von Bildverarbeitung/Visualisierung und Computerlinguistik besteht hier für die Digital Humanities ein großes Potential an Forschung, Anwendung und Kritik. Daneben verbinden sich durch die Bildsynthese die Forschungsfelder Computer Vision und Computer Grafik noch enger, so dass ein transdisziplinärer bildwissenschaftlicher Gegenstand vorliegt. Eine erste Phase von ‚Kunstkritik‘ ist den Architekturen schon eingeschrieben, da ständig entschieden wird, welche Bilder den Trainingsdaten ähneln oder durch andere Kriterien als angemessen erscheinen.

DALL-E der Stiftung OpenAI lässt sich durch sieben Gründe nicht als ‚open‘ bezeichnen, da der Code ist bislang nicht publiziert worden und die Daten anhand derer das maschinelle Lernen durchgeführt wurde (oder werden) nicht bekannt sind. Bekannt ist lediglich, dass es sich um im Internet gescrapte Bild- und Textkombinationen (z.B. Bild und Bildunterschriften) handelt. Zudem gab es zunächst keinen freien Zugang zur Nutzung, sondern nur eine Registrierung über eine Warteliste und keine Transparenz über die Zugangsvergabe. Seit 20.07.2022 ist DALL-E monetarisiert, indem nach einer gewissen Anzahl unentgeltlicher Bilder, dann für weitere Bilderzeugungen bezahlt werden muss. Ähnlich wie bei Midjourney erscheinen diese Kosten allerdings relativ moderat. Die generierten Bilder gehören alle openai. Selbst wenn ein eigener Bildupload die Grundlage bildete oder das Kommando eine Innovation/Schöpfungshöhe besitzt. Allerdings räumt openai seit Juli 2022 eine Nutzung der Bilder auch zu kommerziellen Zwecken ein. Gleichzeitig wehren sich Künstler*innen international gegen die unfreiwillige Verwendung ihrer Bilder als Trainingsdaten, da ihr Stil so sehr einfach reproduzierbar wird.

Die Offenheit von DALL-E wird zuletzt durch eine teils rigide Content Policy eingeschränkt, wodurch Bilder mit sensiblen Inhalten oder Deepfakes verhindert werden sollen. Andere Modelle wie Stablediffusion sind offen im Hinblick auf open source, open access und freier Texteingabe, was wiederum Kritik aufgrund der Missbrauchsmöglichkeiten erzeugt.

Eine kritische Reflexion darüber, ob OpenAI mit dieser Herangehensweise nicht ihrer eigenen Stiftungscharta³ widerspricht, kann nur angerissen werden. Die

Geschlossenheit der kommerziellen Systeme wie DALL-E, IMAGEGEN und PARTI ist allerdings nur ein Aspekt, der eine wissenschaftliche Erforschung und Kritik erschwert. Der Aufbau konkurrenzfähiger Modelle an Universitäten scheitert oft am Aufbau der Trainingsdatensätze sowie der Infrastruktur respektive Rechenleistung für das maschinelle Lernen. Die akademische Forschung hat somit Schwierigkeiten selbst unabhängige Architekturen zu erstellen und kann gleichzeitig mangels Verständnisses der führenden Modelle keine fundierte Kritik zu deren Funktionsweise artikulieren. Modelle wie Stablediffusion (Rombach 2022) zeigen hingegen, dass durchaus konkurrenzfähige Architekturen im akademischen Kontext entstehen können. Insgesamt besteht allerdings die Herausforderung die Verfahren der oft als black box beschriebenen Netze zu interpretieren. Dies ist u.a. möglich durch den Vergleich der unterschiedlichen Modelle und Trainingsdatensätzen oder Ergebnissen aus unterschiedlichen Phasen der Bildgenerierung, sowie durch Prompt Engineering als experimentelle Methode. Bei letzterer kombiniert und ergänzt man in Versuchsreihen die Satzteile einer Befehlszeile in unterschiedlichen Reihenfolgen. Dabei können auch Marker gesetzt werden, z.B. indem im Kommando dazu aufgefordert wird, Personen einzufärben oder auf andere Weise Objekte zu markieren, um zu sehen, ob das Modell nicht nur Bilder des Konzepts kopiert, sondern ein tieferes Verständnis des Konzepts hat. Auch kann untersucht werden, wie ein Bild durch Zugabe oder Umformulierung von Text realistischer wird. An dieser Stelle zeigt sich auch wie subjektiv die Kriterien der Bewertung sind. Welche Grundlage hat die Wahrnehmung ein Bild realistischer oder die Umsetzung (man könnte auch sagen Inszenierung) eines prompts gelungener zu finden?

Kunstkritik

Neben der Medienarchäologie zu neuronalen Netzen und dem maschinellen Lernen muss hier also eine zweite Form der Auseinandersetzung gefunden werden, die sich methodisch an klassische Kategorien und Werte der Kunstkritik anlehnt. Diese Werte sind nicht essentialistisch zu verstehen, sondern relativistisch. So kann ein visueller Turing-Test durchgeführt werden, indem gefragt wird, ob das Bild für computergeneriert oder eine Fotografie bzw. von Menschen digital oder konventionell erstellte Grafik gehalten wird. Hier zeigt sich, dass in DALL-E 2 viele Bilder erzeugbar sind (Photograph of a house in the Bauhaus style; Abb. 1), die in diesem Turing-Test bestehen, während andere unbefriedigende Ergebnisse liefern (17th century landscape painting). Das Problem ist in vielen Fällen nicht die Imitation eines Stils, sondern die Form der dargestellten Dinge also die Detailschilderungen (im Fall der niederländischen Landschaftsmalerei sind es monströse Kühe). Die Bildgeneratoren insbesondere DALL-E 2 sind intelligent im Sinne von Mimikry. Sie können die Oberflächen von Dingen, Stilen, Texturen und Beleuchtungen überzeugend wiedergeben ohne ein tieferes Verständnis der Gegenstände zu haben. Die für die Kunstgeschichte und Kunstkritik so wichtigen Pole, Kunst und Natur, haben auch bei der künstlichen Bildgenerierung eine Bedeutung. Das antrainierte „Weltwissen“ führt dazu, dass die Modelle auch in der Natur

vorhandene Proportionen internalisieren – also visuelle Prinzipien der Kompositionalität befolgen. Entsprechend fällt es beispielsweise dem Algorithmus schwer Elefanten zu erzeugen, die kleiner als Schildkröten sind. Dadurch erscheint auch der Namensgeber Salvatore Dali für DALL-E eher unpassend, weil dessen radikal unnatürlichen Schöpfungen innerhalb seiner surrealen Kunst mit DALL-E kaum zu reproduzieren und noch weniger von diesem Modell zu erfinden wären.

Auf diese Weise sollen Kriterien bzw. die so genannten Rubriken der Kunstkritik (Vogt 2010) wie Bilderfindung, Komposition, Ausdruck, Stil und Dekor, Naturnähe (im Sinne von real-world-data) und Kunstrezeption untersucht werden. Auf ikonologischer Ebene führt die Bildkritik direkt in gesellschaftliche Zusammenhänge. Denn auch bei den neuen Generationen von Bildgeneratoren zeigen sich der Bias, der in Bezug auf das maschinelle Lernen bereits viel besprochen wurde (z.B. Stereotype Frauen- und Männerberufe sowie nicht-weiße Kriminelle werden wie selbstverständlich erzeugt), während die vollständige Unwissenheit bzgl. einiger kanonischer Werke der Kunst- und Kulturgeschichte auch die mangelnde Historizität der Trainingsdaten aufzeigt.

Anhand der exemplarischen Beurteilung der synthetischen Bilder soll es weniger um Urteile der einzelnen Bilder gehen, als um die Aufstellung von Kriterien und Erkenntnisse über die Potentiale von aiart insgesamt. Dabei zeichnet sich ab, dass wir es mit einer Form von Mimikry zu tun haben, durch die Oberflächen und Stile immer besser imitiert werden können, während komplexere Kompositionen und Konzepte scheitern. Die Forschungen zu aiart entstanden im Rahmen des SPP „Das digitale Bild“ im Projekt „Bildsynthese als Methode des kunsthistorischen Erkenntnisgewinns“ (2019–2022). In Ausblick und Diskussion soll auch darauf eingegangen werden, welche weiteren Anwendungsfelder sich in der Kunstgeschichte für die künstlich generierten Bilder aufzeigen lassen.



Abb. 1: Mit Dall-E 2 erzeugte Bauhaus Architekturen.

Fußnoten

1. StyleGAN, VQGAN+CLIP, GAUGAN, crayon, DALL-E 1 und 2, Imagegen, Midjourney, Stablediffusion.
2. z.B. ImageNet, OpenImage, CelebA.
3. OpenAI Charter: <https://openai.com/charter/>

Bibliographie

Chen, Weiwen, Mohammad Shidujaman, und Xuelin Tang (2020). *AiArt: Towards Artificial Intelligence Art*. In The 12th International Conference on Advances in Multimedia.

Esser, Patrick, Robin Rombach, und Björn Ommer (2021). *Taming Transformers for High-Resolution Image Synthesis*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12873–12883.

Kohle, Hubertus, und Stefan Germer (1991). „Spontaneität und Rekonstruktion. Zur Rolle, Organisationsform und Leistung der Kunstkritik im Spannungsfeld von Kunsttheorie und Kunstgeschichte“. Buchbeitrag. Wiesbaden. <https://archiv.ub.uni-heidelberg.de/artdok/109/>.

Marcus, Gary, Ernest Davis, und Scott Aaronson (2022). *A very preliminary analysis of DALL-E 2*. arXiv, 2. Mai 2022. <https://doi.org/10.48550/arXiv.2204.13807>.

Offert, Fabian, und Peter Bell (2022). *Generative Digital Humanities*. In CHR, pp. 202–212.

Oppenlaender, Jonas (2022). *Prompt Engineering for Text-Based Generative Art*. preprint arXiv:2204.13988.

Rombach, Robin, et al. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2022, pp. 10674–85, <https://doi.org/10.1109/CVPR52688.2022.01042>.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, et al. (2022). *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. arXiv. <https://doi.org/10.48550/arXiv.2205.11487>.

Sparkes, Matthew (2022). *AI Art Tool Covertly Alters Requests*. New Scientist 255, Nr. 3397 (30. Juli 2022): 10.

Vogt, Margrit (2010). *Von Kunstworten und -werten: Die Entstehung der deutschen Kunstkritik in Periodika der Aufklärung*. De Gruyter, Berlin.

Open Culture im
Museum: „Skandal-
KULTUR reloaded:
Literarische Affären
INTERAKTIV erkunden“
als digitales
Ausstellungsexperiment

Bamberg, Claudia

bamberg@uni-trier.de
Universität Trier, Deutschland

Lambertz, Michael

lambertz@uni-trier.de
Universität Trier, Deutschland

Petkov, Radoslav

petkov@uni-trier.de
Universität Trier, Deutschland

Zusammenfassung

Die digitale Vermittlung von kulturellen Inhalten gewinnt für Museen und Gedenkstätten immer mehr an Bedeutung (u.a. Kohle 2021); sie hat durch die Pandemie, nicht zuletzt durch gesonderte Ausschreibungen etwa der Bundeskulturstiftung,¹ nochmals einen Aufschwung bekommen (Richard et al. 2022). Aber auch schon davor existierten Förderkonzepte, die ihren Fokus explizit auf experimentelle Formate digitaler Kulturvermittlung legten, so z.B. die Reihe „experimente#digital“ der Aventis Foundation, die gezielt experimentelle Digitalkonzepte in Museen fördert.²

Dabei geht es nicht nur darum, innovative Inszenierungen im Museum zu kreieren, sondern auch um die Entwicklung von Vermittlungsformen, die allein für den digitalen Raum bestimmt sind und sich auch abseits des analogen Museumsortes virtuell erfahren lassen – und die damit neue Möglichkeiten und Chancen sowohl der Aufbereitung als auch der Verbreitung von kulturellen und wissenschaftlichen Inhalten eröffnen (Franken-Wendelsdorf 2019). Die Frage nach ‚Open Culture‘ bildet dabei einen Dreh- und Angelpunkt: ist mit ihr im Museumskontext doch erstens ein freier und transparenter Zugriff auf die Daten sowie zweitens eine Usability gemeint, die verschiedene Zielgruppen adressiert. Drittens sollte es – zumindest im Idealfall – auch darum gehen, Open Culture als Sujet der virtuellen Inszenierung mitzureflektieren und neu zu verhandeln.

Wie dies umgesetzt werden kann, was hierbei zu beachten ist und wo die Herausforderungen liegen, möchte der geplante Vortrag am Beispiel des digitalen Museumsprojekts „Skandal-KULTUR reloaded – Literarische Affären INTERAKTIV erkunden“ vorstellen und zur Diskussion stellen. Das Projekt wurde am Freien Deutschen Hochstift – Frankfurter Goethe-Museum in Kooperation mit dem Trier Center for Digital Humanities durchgeführt und von der Aventis Foundation gefördert.³ Die inhaltliche und gestalterische Konzeption ist weitgehend abgeschlossen (s. Abbildung), die noch offenen Fragen der Umsetzung, die insbesondere das Thema „Interaktivität“ betreffen, möchte der Vortrag diskutieren. Die Plattform wird zur DHd 2023 freigeschaltet.

Skandal und Open Culture

Das Thema Skandal eignet sich für die Reflexion von Open Culture besonders. In Skandalen verdichten und konzentrieren sich Konflikte und Veränderungsprozesse einer Gesellschaft (z.B. im Umgang mit Minoritäten, Wer-

tewandel etc.). Sie sind Indizien von offenen Gesellschaften und für deren Aushandlungsprozesse und Selbstverständnis unverzichtbar (Blasberg 2009). Anders gesagt: Dort, wo es keine Skandale gibt bzw. diese unterdrückt oder deren Auslöser:innen bestraft werden, kann es auch keine offenen Debatten geben – Modernisierungsprozesse bleiben notwendig aus. Dabei war und ist die Sensibilität für das, was als ‚Skandal‘ wahrgenommen und definiert wird, in von der Zensur betroffenen Gesellschaften meistens sehr hoch, in offenen Gesellschaften dagegen deutlich geringer.

Skandale gelten als zentrale Kommunikationsphänomene der Moderne: Sie vermessen das Spannungsfeld zwischen öffentlich Sagbarem und Unsagbarem, zwischen öffentlich Zeigbarem und Nicht-Zeigbarem stets neu und bewegen sich dabei auf der brisanten Grenze zwischen Herabwürdigung, Beschämung und Bloßstellung einerseits sowie satirischer Polemik und sprachlich-medialer Virtuosität andererseits (Friedrich 2011, Roßbach 2020). Neben kurzfristigen, heute allgegenwärtigen Medien‚skandalen‘, wie sie insbesondere in den Social Media mit allen fragwürdigen Grenzüberschreitungen (Hate Speech, Fake News etc.) inszeniert werden (Pörksen 2012), gibt es Skandale mit langfristiger Wirkung, die an gesellschaftlichen Grundfesten rütteln und die Grenzen des Sag- und Zeigbaren erweitern (Blasberg 2009).

Skandal-KULTUR reloaded

Der Rückblick auf historische Situationen kann somit Spiegel und Anstoß sein für eine vertiefende Auseinandersetzung mit auch heute noch brandaktuellen Themen und Problemen wie Ausgrenzung, Antisemitismus, Religions- und Moralfragen, neue Lebens- und Liebeskonzepte, Homophobie etc. Das digitale Museumsprojekt, das hier vorgestellt und in seiner Konzeption und Methodik zur Diskussion gestellt werden soll, widmet sich Literaturskandalen zwischen den beiden großen Revolutionen 1789 und 1848 (Gelz et al. 2014); dabei stehen die genannten Konflikte im Fokus, deren Themen bis heute von großer Aktualität sind: Um ‚verbotene‘ Liebe, Selbstmord und Jugendgefährdung wird in Goethes Briefroman „Die Leiden des jungen Werthers“ (1774) und den zeitgenössischen Reaktionen heftig gestritten. Einen Skandal um Fragen von Sexualität und Gender sowie von neuen, provokativen Ausdrucksmöglichkeiten in der Kunst entfachten Friedrich Schlegels romantischer Roman „Lucinde“ (1799) und die romantische Zeitschrift „Athenaeum“ (1798–1800). So konnte man z.B. in Friedrich Schlegels Roman „Lucinde“ die poetische Inszenierung des Rollentauschs von Mann und Frau beim Liebespiel finden („Eine unter allen [Situationen] ist die Witzigste und die Schönste: wenn wir die Rollen vertauschen und mit kindischer Lust wetteifern, wer den andern täuschender nachäffen kann, ob dir die schonende Heftigkeit des Mannes besser gelingt, oder mir die anziehende Hingebung des Weibes [...]“, Schlegel 2020, S. 19), und im 272. Athenaeumsfragment heißt es: „Warum sollte es nicht auch unmoralische Menschen geben dürfen, so gut wie unphilosophische und unpoetische? Nur antipolitische oder unrechtliche Menschen können nicht geduldet werden“ – beides war für die Zeitgenossen eine

unerhörte Provokation. Am Skandal um Karl Sessas Posse „Unser Verkehr“ (1812/1815) lassen sich antisemitische Angriffe bis hin zu den pogromartigen sog. „Hepp-Hepp-Krawallen“ (1819) verfolgen. Die Affäre um Heinrich Heines Judentum und Graf August von Platens Homosexualität (1826–1830) eignet sich für Analysen rhetorischer Strategien zur Ausgrenzung von Minoritäten. Schließlich ist ein Religionsskandal Thema: Karl Gutzkows Roman „Wally, die Zweiflerin“ (1835) wurde als heftiger Angriff auf Moral und Religion gewertet – der Autor musste sogar ins Gefängnis.

Konzeption

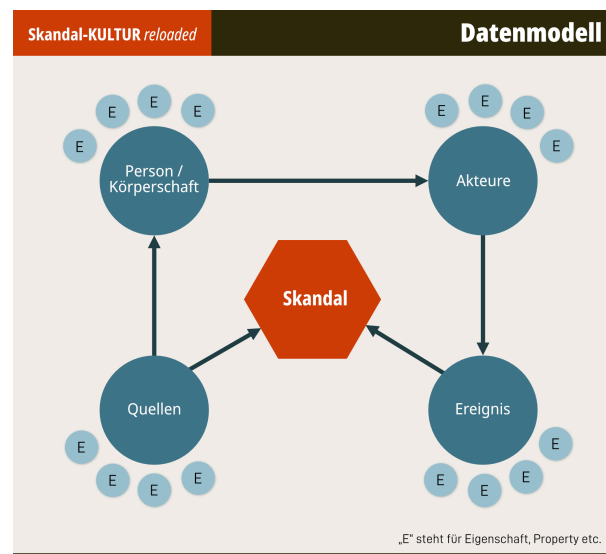
Das Hauptziel der digitalen Wissensvermittlung im rein virtuellen Museum war es, die Skandale für ein breites Publikum möglichst verständlich und real zu machen, ohne dabei den wissenschaftlichen Fokus zu verlieren. Das umfasst zum einen die Themen der Skandale, die Akteure und den historischen Kontext, aber auch die Form der Kommunikation.

Die Akteure der Skandale, die im Projekt thematisiert werden – insbesondere die Frühromantiker:innen – galten bereits als Medienvirtuosen, die die Spielarten des Druckmediums voll ausreizten (z.B. im Billet und der Anzeige) und gegen ihre literarischen Gegner, insbesondere die ältere Generation, einsetzten (vgl. Oesterle 2017). Zugleich sind die von ihnen skandalisierten Themen damals wie heute aktuell, so dass es in der Konzeption des Projekts eine Zielsetzung war, ein gemischtes Publikum für ihre Aktualität zu sensibilisieren und diese erlebbar zu machen. Um das Erleben kreativ und innovativ zu gestalten, bieten sich zwei grundsätzlich Alternativen einer stilistisch tonalen Umsetzung an, sowohl vom Design her als auch von der User Experience insgesamt. Einerseits besteht die Möglichkeit, durch das Gestalten einer historisch anmutenden Umgebung eine Atmosphäre zu schaffen, ähnlich wie sie damals herrschte. Eine zweite Möglichkeit, das Szenariums in die heutige Zeit mit heutigen Formen der Medialität zu übertragen, erschien dem Projektteam reizvoller, weil dadurch die Parallelen zur Skandalkultur heute sehr deutlich werden können. Ein wesentlicher Unterschied zwischen damals und heute ist der Faktor Zeit und Beschleunigung medialer Kommunikation. Während früher zwischen den Botschaften der Parteien und der Gegenrede Tage, Wochen oder gar Monate vergingen, sind es heute oft nur noch wenige Minuten (s. zur Entwicklung der Massenmedien und zum zunehmenden Tempo der Kommunikation Jäckel 2011). Gleichwohl waren die Texte des Disputs nicht weniger emotional aufgeladen und impulsiv, wie es heute der Fall ist, was zusätzlich für eine Umsetzung der Plattform in heutigem Stil spricht.

Technische Umsetzung

Die Datenerfassung wurde mit der virtuellen Forschungsumgebung „Forschungsnetzwerk und Datenbanksystem“ FuD (<https://fud.uni-trier.de>) umgesetzt. Die Datenstruktur ist auf die konkreten Anforderungen des Projektes angelegt: Skandale als zentrale Objekte

sind verknüpft mit zugehörigen Ereignissen, diese wiederum mit Akteuren, welche wiederum mit Personen bzw. Körperschaften verlinkt sind, die ihrerseits verschiedenen Skandalparteien angehören. Insbesondere werden eine Reihe an Informationen wie Quellen, Orte, Bilder etc. zu den jeweiligen Objekttypen erhoben. Das Modell wird in die FuD-Logik überführt und in einer relationalen Datenbank abgebildet. Durch die Flexibilität von FuD und das agile Datenmodell, das wir definiert haben (siehe Abbildung), sind wir in der Lage, jederzeit in dessen Komplexität einzugreifen und individuelle Anpassungen bzw. Erweiterungen vorzunehmen.

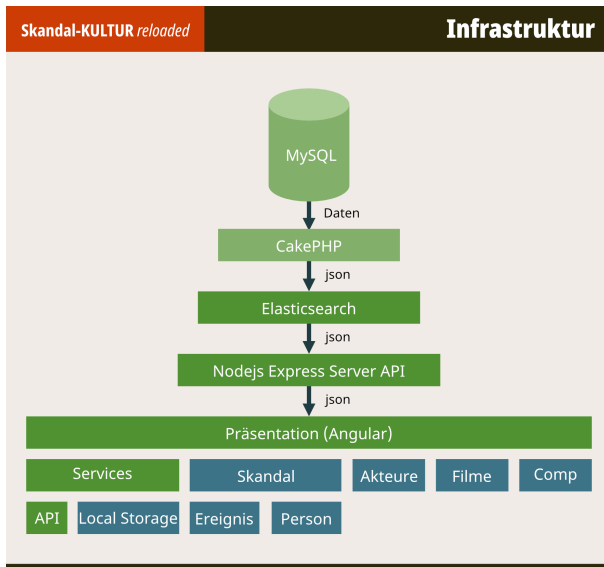


Datenmodell

Während das Design (Details s.u.) keineswegs typisch ist, sind die Ansichten der Weboberfläche eher konventionell aufgebaut. Bei der Online-Präsentation greifen wir auf bereits am TCDH etablierte Strukturen und Prozesse zurück und setzen auf neue, moderne und nachhaltige Technologien. Dabei bauen wir auf unsere Expertise aus der Arbeit mit Elasticsearch (<https://www.elastic.co>), NodeJS (<https://expressjs.com>), PHP (<https://www.php.net>), Angular (<https://angular.io>) und setzen auf JSON (<https://www.json.org>) als Austauschformat.

Nachdem die Daten in der Forschungsumgebung erfasst und freigegeben sind, werden diese in deren logische Einheiten exportiert, in das vordefinierte JSON Objekt Modell überführt und in Elasticsearch importiert. Die Weboberfläche wird modular, generisch und projektspezifisch mit dem Front-End-Framework Angular umgesetzt. Der Einsatz von Angular gibt uns die Möglichkeit, auf bereits gut erprobte Methoden zu setzen.

Als Verbindungsgrundbaustein zwischen Front-End und Elasticsearch setzen wir einen NodeJS Express Server als Open API ein, mithilfe dessen über vordefinierte Suchanfragen die RESTful API von Elasticsearch abgefragt wird (siehe Abbildung).

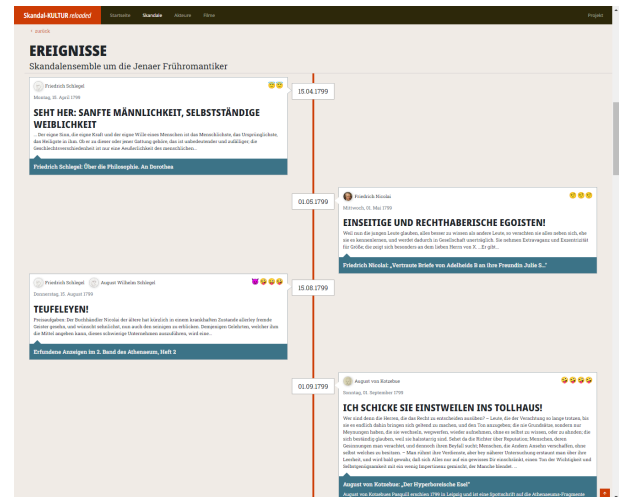


Infrastruktur

UX, Usability und Design

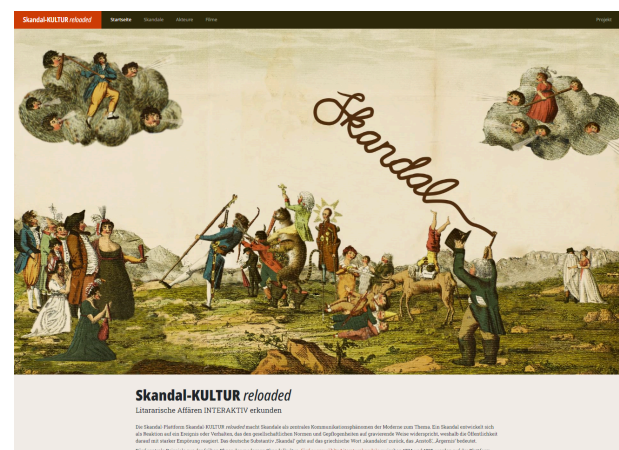
Bezüglich der Usability gelten auf der Plattform alle Regeln, die es auch sonst beim Aufbau einer Webapplikation zu beachten gilt. Seitenbesucher:innen sollen sich schnell zurechtfinden und in Bezug auf ihre Navigationsgewohnheiten nicht gestört werden. Es wurde auf bekannte Gestaltungsregeln und -merkmale zurückgegriffen und es wurden keine exotischen oder gar neu erfundenen Gestaltungselemente genutzt (zu grundlegenden Usabilitygrundsätzen vgl. Jacobsen et al. 2019).

Hinsichtlich der User Experience hingegen wurde versucht, bestimmte für gewöhnlich erwünschte emotionale Reaktionen von v.a. wissenschaftlich traditionell geprägten Nutzer:innen bewusst nicht zu erfüllen. Im Gegenteil sollen diese eine gewisse Irritation in Bezug auf die ästhetische Ansprache der gesamten Applikation erfahren.⁴ Das ist notwendig, damit im Anschluss ein Reflexionsprozess in Gang gesetzt werden kann. Zum einen darüber, wie eine Variante dieses Schlagabtauschs in der heutigen Zeit mit den heutigen Medien ausgetragen werden könnte, zum anderen darüber, was heute in den (sozialen) Medien tagtäglich passiert und sich ebenso zu Skandalen steigern kann. Der Zielgruppe junger Menschen wird die Ästhetik vertrauter sein, wobei sich auch hier spätestens beim genaueren Lesen eine gewisse Diskrepanz zwischen Sprache, Sprachduktus und bestimmten Inhalten auf der einen und dem Darstellungsstil auf der anderen Seite zeigt.



Auszug aus einem Skandalverlauf

Die ästhetische Ansprache besteht hauptsächlich im Design und der grafischen Aufmachung der Seite. In allen Belangen wurden diese in Anlehnung an sensationisierende Newsseiten, wie man sie etwa in der Boulevardpresse oder Yellow Press findet, sowie an Social-Media-Seiten gestaltet: Hierzu gehören eine schrille Farbgebung, übertrieben große fette Lettern, YouTube-Videos, Personendarstellungen in Form von Avatarbildchen, Ereignisse als Posts, ausgeschmückt mit vielen Emoticons usw. All dies soll auf heutige skandalträchtige Onlinemedien hinweisen. Inhaltlich wurde gleichfalls versucht, das Wortgefecht auf die Spitze zu treiben. Dafür haben die Mitarbeiter:innen des Projektes verbal drastische Zitate für die Überschriften und Textausschnitte der Skandalereignisse ausgewählt und als Schlagzeilen formuliert. Über Emoticons⁵ sollen die Stimmung des Beitrages sowie die emotionale Verfassung der Skandalakteure verdeutlicht werden.



Startseite von „Skandal-KULTUR reloaded“

Vom Seitenaufbau her gibt es neben der Startseite eine Überblickseite der Skandale. Von dort aus gelangt man zur Einzelansicht des gewählten Skandals, die das Thema erläutert und die Skandalparteien mit den zuge-

hörigen Akteuren sowie weiterführende Quellen auflistet. Diese Seite entspricht vom Design her einer Newsseite in einer Boulevard-Onlinezeitung. Von dort aus kann man dann über einen (unübersehbar großen) Button zum Skandalverlauf navigieren. Hier sind ähnlich einer vertikalen Infinite-Scroll⁶ Zeitleiste links und rechts die jeweiligen Beiträge (und Reaktionen) mit Akteuren, O-Ton und Kommentierung in chronologischer Reihenfolge gelistet. Zudem gibt es zu jeder Akteurin / jedem Akteur einen Steckbrief, der sich öffnet, sobald man den Avatar anklickt. Ferner präsentiert eine Seite alle Akteure mit den zugehörigen Skandalen. Eine zusätzliche Seite präsentiert darüber hinaus über YouTube eingebettete Videos, die im Rahmen eines Praxisseminars an der Goethe-Universität Frankfurt entstanden sind und in denen Studierende die historischen Skandale aktualisiert haben. Die Videos bieten gegenüber dem systematischen Einstieg der Skandalübersichtsseite einen thematisch emotional motivierten Zugang zum Skandal, zu dem man über einen Link navigieren kann.

Brisanz der Themen und Risiken der Interaktivität

Die Gegenstände der historischen Skandale haben nicht im Geringsten an Aktualität und Brisanz verloren. Die gefährlichen Konsequenzen von systematischer ‚Hate Speech‘ und von ‚Fake News‘ als absichtlich emotionalisierende Formen der (unwahren) Kommunikation sind insbesondere bei öffentlicher antisemitischer Hassrede, aber auch homophober Hate Speech unumstritten. Bei der Darstellung dieser Themen in einem Museumskontext, bei dem es um einen neuen, aktualisierenden Blick auf historische Ereignisse geht, besteht indes immer die Gefahr, auf der Ebene der Präsentation gesellschaftliche und persönliche Toleranzschwellen zu übertreten und die Distanzierung von den Inhalten zu schwach zu markieren – zumal wenn man sich im Design der Social Media bewegt – und dann auch selbst in der Kritik zu stehen.

Darum bedarf es hierfür eines wohl durchdachten und kritisch geprüften Konzepts. Das gilt auch für das Anliegen von „Skandal-KULTUR reloaded“, bei dem nach der Fertigstellung des Grundkonzepts nun auch eine interaktive Erkundung der thematisierten Literaturskandale entwickelt wurde. Die Nutzer:innen können sich beispielsweise Personensteckbriefe anschauen oder Ereignisse im zeitlichen Verlauf in den Fokus nehmen. Das Kommunikationsmodell ist dabei in historischer Betrachtung wechselseitig (zwischen den Skandalparteien) und es ist für uns von außen beobachtbar. Jedoch wird es den Besucher:innen der virtuellen Ausstellung nicht ermöglicht, auf der Plattform sozial zu interagieren, sie können sich also nicht in das Skandalgeschehen einmischen. Die Inhalte und somit der Zustand der Applikation erfahren keine Veränderung im Nachhinein, wie dies bei offenen sozialen Medien der Fall wäre.

Im Kontext von Social Media könnte dies zwar sehr spannende (und spannungserzeugende) Wirkungen zeigen. Man stelle sich vor, die/der Besucher:in könnte im Ereignisverlauf die Beiträge liken oder disliken, teilen sowie diese und ganze Skandale kommentieren. An-

dere würden diese Kommentare up- und downvoten, und es könnten durch algorithmische Auswertungen Interessen- und meinungsbezogene Netzwerke (Stichwort Filterblase, vgl. u.a. Pariser 2011) und Hierarchien der Museumsbesucher:innen entstehen. Indem man sich beispielsweise über Google, Facebook oder E-Mail registriert, könnte man mit anderen (ähnlich gesinnten) Nutzer:innen und den Akteuren in Kontakt kommen und so Follower finden – die Möglichkeiten sind vielseitig.⁷ Bei einer solchen Erweiterung stünde jedoch neben rechtlichen Gründen und Regelungen im Internet der Aufwand der Moderation und nachhaltigen Pflege in keinem Verhältnis zum Umfang des Projektes. Auf inhaltlicher Ebene würde solch ein digitales historisches Social-Media-Museum mit ziemlicher Sicherheit einen Skandal in den realen heutigen Medien auslösen.

Dennoch bleibt die Idee des digitalen Museums mit mehr sozialer Interaktion der Besucher:innen. So möchte der Vortrag nach einer Präsentation des Projekts danach fragen, wie interaktiv solche Unternehmen sein sollten, wo die inhaltlichen und gestalterischen Grenzen der medialen Aktualisierung liegen und welche Strategien für eine ‚kontrollierte‘ Interaktivität entwickelt werden können. ‚Open Culture‘ im Kontext des Ausstellungsmachens bedeutet immer auch, solche Fragen offen auszutragen, um Szenographien zu entwickeln, die möglichst viele Besucher:innen erreichen und dabei zugleich das Wertesystem einer offenen Kultur achten.

Fußnoten

1. So ist etwa die Ausschreibung „dive in – Programm für digitale Interaktionen“ im Rahmen der Pandemie entstanden, s. https://www.kulturstiftung-des-bundes.de/de/projekte/erbe_und_vermittlung/detail/dive_in_programm_fuer_digitale_interaktionen.html.
2. Vgl. <https://www.aventis-foundation.org/kultur/experimentedigital/>
3. Projekthomepage: <https://tcdh.uni-trier.de/de/projekt/skandal-kultur-reloaded>
4. Diese Irritation ist nicht zu verwechseln mit der durch unerwartetes Seitenverhalten (Usability).
5. Um als ‚Social Media‘ wahrgenommen zu werden, muss in gewissem Maße nach dessen Regeln gespielt werden. Die Auswahl der Emoticons wird somit nötig, auch wenn sie sich nicht in den historischen Medien befinden. Sie wurden auf der Grundlage einer Auswertung der Quellen hinzugefügt, wobei die Bestimmung des emotionalen Erregtheitslevels der Autoren natürlich immer eine subjektive Beurteilung aus heutiger Sicht bleibt.
6. Der Infinite-Scroll Mechanismus entspricht einer Art Endlosschleife im Design. Es muss vom User keine bewusste Entscheidung getroffen werden, um weitere Inhalte zu laden. Üblicherweise wird dies in Social-Media-Apps benutzt, was „eine längere Verweildauer bedeutet, dass mehr Werbung betrachtet wird“ (Yablonski 2020, 100). Diese Ähnlichkeit an gewohnte Mechanismen ist eine Entscheidung des Designkonzepts, da sie bei der Benutzung das Gefühl unterstützt, sich in entsprechendem Medium zu befinden.

7. Eine interessante Übersicht und kritische Auseinandersetzung von UX-Design-Möglichkeiten, wie sie heute in Apps umgesetzt werden, um letztlich finanziellen Gewinn zu generieren, wie deren psychologisch Funktionsweise ist und welche Risiken sie bergen, findet man bei Yablonski 2020: 97-109.

Bibliographie

Blasberg, Cornelia. 2009. „Skandal. Politische Pragmatik, rhetorische Inszenierung und poetische Ambiguität.“ In *Amphibolie – Ambiguität – Ambivalenz*, 269-290. Würzburg: Königshausen und Neumann.

Coquelin, Mathieu und Stephan Ruhmannseder. 2019. „Wir gegen die anderen? Zum Umgang mit Hate Speech in Zeiten von Fake News und Verschwörungsideologien.“ In *Mythen, Ideologien, und Theorien. Verschwörungsglaube in Zeiten von Social Media*, hg. von der Landesarbeitsgemeinschaft Mobile Jugendarbeit/Streetwork Baden-Württemberg e. V./Jugendstiftung Baden-Württemberg, 21 -31. Vaihingen an der Enz: Printmedien Karl-Heinz Sprenger.

Franken-Wendelsdorf, Regina. 2019. *Das erweiterte Museum. Medien, Technologie und Internet*. Berlin / Boston: De Gruyter.

Friedrich, Hans-Edwin, Hg. 2011. *Literaturskandale*. Frankfurt am Main: Peter Lang.

Gelz, Andreas, Dietmar Hüser und Sabine Ruß-Sattar, Hg. 2014. *Skandale zwischen Moderne und Postmoderne. Interdisziplinäre Perspektiven auf Formen gesellschaftlicher Transgression*. Berlin: De Gruyter.

Ilbrig, Cornelia. 2018-2020. „Die Geburt der Frühromantik aus dem Geiste des Skandals.“ In *Internationales Jahrbuch der Bettine-von-Arnim-Gesellschaft. Forum für die Erforschung von Romantik und Vormärz*: 131-158.

Jäckel, Michael. 2011. „Die Entwicklung der (Massen-)Medien.“ In *Medienwirkungen*, 33-60. Wiesbaden: VS Verlag für Sozialwissenschaften.

Jacobsen, Jens und Lorena Meyer. 2019. *Praxisbuch Usability & UX: was jeder wissen sollte, der Websites und Apps entwickelt*. Bonn: Rheinwerk Verlag.

Kohle, Hubertus. 2019. „Museen digital. Eine Modernisierungsperspektive für Gedächtnisinstitutionen.“ In *Zeitschrift für Wissenschaft und Kunst in Bayern* 3: 36-37.

Krug, Steve. 2014. *Don't make me think! : web & mobile usability – das intuitive Web*. Frechen: Mitp.

Oesterle, Günter. 2017. „Romantische Satire und August Wilhelm Schlegels satirische Virtuosität.“ In *Aufbruch ins romantische Universum: August Wilhelm Schlegel*, hg. von Claudia Bamberg und Cornelia Ilbrig, 70-82. Göttingen: Göttinger Verlag der Kunst.

Pariser, Eli. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

Pörksen, Bernhard. 2012. *Der entfesselte Skandal. Das Ende der Kontrolle im digitalen Zeitalter*. Köln: Halem.

Richard, Birgit, Jana Müller und Niklas von Reischach, Niklas, Hg. 2022. *Interaktion – Emotion – Desinfektion. Kunst und Museum in Zeiten von Corona*. Frankfurt am Main: Campus Verlag.

Rößbach, Regina. 2020. *Der Literaturskandal. Akteure, Verläufe und Gegenstände eines Kommunikationsphänomens*. Berlin: Frank & Timme.

Schlegel, Friedrich. 2020. *Lucinde. Studienausgabe*, hg. von Stefan Knödler. Stuttgart: Reclam.

Yablonski, Jon. 2020. *Laws of UX. 10 praktische Grundprinzipien für intuitives, menschenzentriertes UX-Design*. Heidelberg: O'Reilly.

Open Data in musikphilologischen Projekten: Herausforderungen, Strategien, Potenziale

Kepper, Johannes

kepper@edirom.de

Universität Paderborn, Deutschland

Münzmay, Andreas

andreas.muenzmay@uni-paderborn.de

Universität Paderborn, Deutschland

Abstract

Datenmanagement ist eine zentrale Fragestellung für sämtliche Projekte im Bereich der Digital Humanities (Altenhöner et al. 2020). Warum also ein gesonderter Blick auf DH-Projekte mit einem musikeditorischen Hintergrund? Die Publikationspraxis musikwissenschaftlicher Editionen unterscheidet sich im Kern deutlich von derjenigen in anderen Disziplinen. Hieraus ergeben sich fachspezifische Herausforderungen, die durchaus Einfluss auf die Konzeption bestehender digitaler Musikeditionsprojekte nehmen. Im Rahmen des Vortrags wollen wir versuchen, diese Besonderheiten herauszuarbeiten und anhand des von der Mainzer Akademie der Wissenschaften und der Literatur geförderten Projekts "Beethovens Werkstatt" Ansätze zur Umsetzung von Open Data auch im Bereich der Musikphilologie aufzuzeigen. Potenziale ergeben sich dabei sowohl auf inhaltlicher Seite, also in Bezug auf die erarbeiteten Daten, wie auf methodischer Seite, d.h. in Bezug auf die entwickelten Tools.

Zur Ausgangslage: Publikationspraxis und Wertschöpfungsketten musikwissenschaftlicher Editionen

Während es etwa in der Literaturwissenschaft eine Fülle verschiedener Publikationsformen gibt – Studienausgaben, Leseausgaben, kritische Ausgaben usw. –,

konnte sich in der Musikwissenschaft nie eine entsprechende Vielfalt wissenschaftlicher Ausgabentypen entwickeln. Schon die ersten musikalischen Gesamtausgaben formulierten in ihren Vorworten den Doppelanspruch als Ausgaben 'für Wissenschaft und Praxis gleichermaßen' (vgl. Fellerer 1980:185; Kepper 2011:180). Die Ursache waren schon damals finanzielle Erwägungen: Die manuelle Erstellung der Druckvorlagen für Notenstich bzw. -druck war schon immer ein sehr zeitintensiver Prozess. Breitkopf und Härtel etwa hatten 1869 insgesamt 14 Notenstecher in Dienst (Hase 1968:386), die in 1867 eine Jahresleistung von "über 5000 Platten" hatten (Hase 1968:394). Setzt man diese Zahlen ins Verhältnis, landet man bei etwa 357 Platten pro Notenstecher. Dieser Zeitbedarf von etwa einem Arbeitstag pro Seite Notentext blieb im Notenstich bis zur Verbreitung des computerbasierten Notensatzes gegen Ende des 20. Jahrhunderts recht konstant (Popp 2017). Dem standen bereits im 19. Jahrhundert recht geringe Auflagenhöhen gegenüber, und auch heutige Gesamtausgaben haben oft Subskribentenzahlen im unteren dreistelligen Bereich. Die Bände dieser Gesamtausgaben sind i.d.R. sehr aufwendig ausgestattet und werden entsprechend teuer verkauft – oft genug ebenfalls im unteren dreistelligen Bereich. Die Inhalte für diese Bände werden zu meist durch öffentlich geförderte Projekte erstellt und den Verlagen kostenfrei, teils sogar mit einem zusätzlichen Druckkostenzuschuss zur Verfügung gestellt. Auf dieser Basis liegt die eigentliche Wertschöpfung für die Verlage dennoch an nochmals anderer Stelle: Erst mit sogenannten praktischen Ausgaben für den professionellen wie nichtprofessionellen musikalischen Gebrauch wird ein größerer Markt jenseits von Bibliotheken und Musikwissenschaft adressiert. Außerdem eröffnen der Vertrieb von Leihmaterial und v.a. die Lizenzierung von Aufführungen bzw. Rundfunksendungen wichtige Einnahmequellen. Aufgrund der bereits erwähnten hohen Produktionskosten ist es üblich, diese praktischen Ausgaben mit möglichst geringem Aufwand aus den wissenschaftlich-kritischen Ausgaben zu destillieren und, beispielsweise mit dem schillernden Label "Urtext", zu vertreiben. Dafür wird oft der bereits vorhandene Notentext mit einem neuen Vorwort versehen, der Kritische Bericht – wenn überhaupt berücksichtigt – auf wenige Seiten kondensiert, und so mit minimalem Aufwand eine Ausgabe erstellt, die immer noch den aus Marketingsicht günstigen Anspruch erhebt, auf neuesten wissenschaftlichen Erkenntnissen zu basieren. Dass diese Wissenschaftlichkeit gerade aus den entfallenden Teilen resultiert, ist aus dieser Perspektive unerheblich. Viel wichtiger ist es aus Verlagssicht, dass sich das Notenbild schon der wissenschaftlichen Ausgaben nicht durch ein (vermeintliches) Übermaß diakritischer Auszeichnung für eine praktische Nachnutzung disqualifiziert, sondern mit einem vertretbaren Aufwand an diese Zweitverwertung angepasst werden kann. Solche wissenschaftsfremden Prämissen sind prägend für Ausgaben, die dem Paradigma "für Wissenschaft und Praxis gleichermaßen" unterworfen sind.

Nicht erst angesichts der sich stark verändernden gesetzlichen Rahmenbedingungen (v. a. das im Juli 2021 in Kraft getretene Zweite Open Data Gesetz der Bundesregierung verpflichtet ab 2024 zur Veröffentlichung aller mit Bundesfinanzierung erhobenen Forschungsda-

ten als Open Data) und der bereits entsprechend veränderten Vorgaben der Drittmittelgeber (vgl. z. B. die DFG-Information Nr. 25 "Konkretisierung der Anforderungen zum Umgang mit Forschungsdaten in Förderanträgen" vom 14.3.2021:

https://www.dfg.de/foerderung/info_wissenschaft/2022/info_wissenschaft_22_25/index.html) für den Umgang mit Forschungsergebnissen erscheint diese Wertschöpfung fragwürdig.

Es dürfte offensichtlich sein, dass Musikverlage bislang kaum als treibende Kraft für Open Data / Open Culture in Erscheinung getreten sind. Anders sieht es bei vielen DH-Projekten aus. Die freie Zugänglichkeit der Forschungsdaten spätestens zum Ende des Förderzeitraums ist inzwischen weitgehend obligatorisch. Auch Förderinstitutionen im Bereich Musikwissenschaft wie die Union der Akademien in Deutschland als Förderer der meisten musikphilologischen Langzeitprojekte in Deutschland legen insbesondere bei Neuvorhaben inzwischen verstärkt Wert auf Open Access – bei bestehenden Projekten stehen dem i.d.R. die bestehenden, langfristigen Verlagsverträge im Wege. Damit stehen die Geldgeber an der Seite der meisten Projekte im Umfeld der Digital Humanities. Im Sinne des ratchet effect (Tomasello 2009) arbeiten diese Projekte oft in der Annahme, im Zuge ihrer Forschung Daten zu erstellen, die auch für nach- oder anders gelagerte Forschungsfragen hilfreich sein können. Hierfür ist es aber zwingend notwendig, nicht nur die abgeschlossenen Ergebnisse zu publizieren, sondern auch die ihnen zugrundeliegenden Rohdaten. Zum Wandel der förderpolitischen Rahmenbedingungen gesellt sich dabei ein arbeitskultureller, vor allem aber auch ein Wandel editionstheoretischer Herangehensweisen vor dem Hintergrund der Digitalität, die a) die Statik und Linearität von (musikalischem) Text zugunsten netzförmiger, aber auch dynamischer, prozessualer Darstellungsweisen erweitert (Droese/Münzmay 2015; Stadler 2019), b) editorische und historische Objekte in hybriden Textgebilden zusammenführt (die häufig komplette digitalisierte RePublikationen von Kulturobjekten einbinden, die digital 'beschriftet' und vernetzt werden; Münzmay 2018) und c) Multimodalität editorisch handhabbar macht. Neben die herkömmliche Befassung mit dem Werktext treten neue musikphilologische Gegenstände, wie etwa phonographische Quellen (Orcalli 2017; Münzmay/Siegert 2019; Pasdzierny 2019) oder textgenetische bzw. Schreibprozesse (Kepper/Sänger 2017; Kepper/Cox 2021).

Beethovens Werkstatt

Das etwa in der Mitte seiner auf 16 Jahre angelegten Laufzeit stehende Projekt Beethovens Werkstatt reiht sich ein in eine Riege musikeditorischer Projekte, die bewusst mit den Konventionen der Musikphilologie, d. h. mit Konzepten wie 'Historisch-kritische Werkedition' und 'Gesamtausgabe', brechen und so neue Perspektiven aufzuzeigen versuchen. Dies wird in der Projektkonzeption unmittelbar deutlich: Wie kein anderes musikwissenschaftliches Akademieprojekt versteht sich Beethovens Werkstatt als methodisches Grundlagenforschungsprojekt, das ganz bewusst schon im Antrag auf eine Auflistung der zu behandelnden Werke verzichtet, sondern

lediglich für bestimmte Fragestellungen mit Beispielen versieht. Ziel ist demnach die Untersuchung dieser philologischen Fragestellungen (Darstellung von Varianz, Schreibprozessen und Revisionsprozessen; Fassungsvergleiche; Skizzeneditionen), nicht das starre Abarbeiten von Werklisten. In vielen Fällen ist dafür die Betrachtung von teils stark begrenzten Auszügen vollkommen ausreichend – die thematisierten Werke werden also nicht in Gänze thematisiert. Angesichts der eingangs geschilderten üblichen Verwertungsketten im Musikbereich ergibt sich daraus quasi zwangsläufig eine weitere Besonderheit des Projekts: Der bewusste Verzicht auf die Zusammenarbeit mit einem Musikverlag. Auch wenn in bestimmten Modulen des Projekts durchaus vollständige Werke im Sinne genetischer Textkritik erarbeitet werden, so dass eine musikpraktische Nutzung zumindest einiger Projektdaten nach entsprechender Aufbereitung durchaus vorstellbar wäre, ist dies gerade nicht die Zielsetzung – auch, weil das Projekt aller Voraussicht nach nicht in entsprechendem Maß an der Wertschöpfung beteiligt würde, um den damit verbundenen Aufwand zu kompensieren. Stattdessen werden sämtliche Projektdaten – XML-Daten nach dem Standard der Music Encoding Initiative MEI, SVG-Daten sowie die im Projekt entwickelte modulare Forschungs- und Visualisierungssoftware „VideApp“ – frei zur Verfügung gestellt. Damit steht es Verlagen offen, nun allerdings bei eigenem Aufwand entsprechende Ausgaben auf dieser Datengrundlage zu erstellen. Für das Projekt ergibt sich daraus kein echter Nachteil. Ein gründliches Verlagslektorat – was wohl der aus Projektsicht attraktivste Vorteil einer Zusammenarbeit wäre – ist ohnehin längst nicht mehr bei allen Musikverlagen bzw. Verlagsverträgen selbstverständlich.

An die Stelle der Ausrichtung auf Verlagspublikationen tritt eine enge und wechselseitige Verzahnung von inhaltlicher Erschließung des Gegenstands, Datenerstellung, -aufbereitung und -pflege sowie Entwicklung der Software, die eine Darstellung und Interaktion mit diesen Daten erlaubt. Auch hier zeigt sich eine Eigenheit des Projekts, die es stärker in eine Reihe mit DH-Grundlagenforschungsprojekten als mit herkömmlichen Musikeditionen stellt: Die erstellten Daten sind für sich genommen nicht ausreichend, um die Editions Inhalte zu transportieren. Zumindest auf absehbare Zeit ist nicht davon auszugehen, dass Standardsoftware wie der „DFG-Viewer für musikalische Quellen“ oder die diesem zugrunde liegende Verovio-Rendering-Bibliothek für MEI die Projektdaten in Gänze sinnvoll auswerten können. Vielmehr bedarf es auch der im Projekt erstellten Software, da erst diese die spezifischen Datenmodelle und Konzepte zur genetischen Edition von Musik greifbar macht.

Beethovens Werkstatt verfolgt damit einen integralen Ansatz für eine verschränkte, Hand-in-Hand-gehende Daten-Erarbeitung und Toolentwicklung. Dieser Ansatz ist dabei inhaltlich notwendig – auch mit viel Fantasie und ohne äußere Vorgaben fällt es schwer, sich eine auch nur ansatzweise vergleichbare Umsetzung der Projektergebnisse außerhalb digitaler Medien vorzustellen. Es handelt sich also um eine genuin digitale Edition im von Patrick Sahle beschriebenen Sinne (Sahle 2017:239).

Aus diesem grundlegend digitalen Ansatz folgen aber andere Anforderungen, als sie in musikwissenschaftlichen Editionsprojekten sonst üblich sind. Neben Über-

legungen zur Publikation der Daten bedarf es auch entsprechender Überlegungen zum Umgang mit der Forschungssoftware. Beethovens Werkstatt verfolgt von Beginn an die Strategie, sämtliche Software als Open Source in öffentlichen Repositorien auf GitHub zu entwickeln. Abgesehen von terminologischen Diskussionen, deren Zwischenschritte auf dem Weg zu einer ersten veröffentlichten Version im Glossar des Projekts zunächst in einem nur intern zugänglichen Wiki-System festgehalten werden, findet auch die gesamte inhaltliche Arbeit des Projekts als Open Data öffentlich statt (Open Access). Über parallele Continuous Integration-Pipelines werden stabile Zwischenstände dieser Arbeit öffentlich, tagesaktuelle Stände hingegen zunächst nur für den internen Gebrauch in leicht zugänglicher Form aufbereitet und zur Verfügung gestellt. Unabhängig davon kann die Arbeit des Projekts über die entsprechenden Repositorien jedoch jederzeit vollständig und frei zugänglich nachvollzogen werden. Ein völlig freier „Blick in die Werkstatt“ ist damit jederzeit möglich. Auch in dieser Meta-Perspektive wird also das „deiktische Prinzip“ des Projekts umgesetzt: Statt teils umständlicher Erläuterungen versucht die sog. VideApp, genetische Prozesse möglichst anschaulich zu visualisieren. Entsprechend direkt ist der Zugang zur Arbeit des Projekts (der freilich der gängigen Logik von Forschungsförderung folgend dennoch in regelmäßigen Berichten dokumentiert wird).

Damit bietet Beethovens Werkstatt ein ideales Anschauungsbeispiel zur Komplexität digitaler musikwissenschaftlicher Editionen, die sich u. a. aus nicht-linearen bzw. dynamischen Inhalten bzw. Visualisierungen historischer Objekte und darin manifester genetischer Prozesse sowie der Multimodalität der zur Umsetzung notwendigen Daten speist. Für einzelne der genannten Bereiche gibt es best practices, an denen sich Projekte orientieren können: Für den Bereich der Editionsdaten ist eine revisionssichere Publikation beispielsweise bei Zenodo gut möglich. Für den Bereich der Softwareentwicklung lassen sich Continuous Integration / Continuous Delivery -Workflows mit Docker und GitHub Actions umsetzen. Was aber, wenn die genutzte Software sowohl integraler Bestandteil der Edition ist als auch zeitgleich unabhängig von diesen Daten zur Nachnutzung in anderen Kontexten nutzbar sein soll? Bislang gibt es keine etablierten Standards, die sich einfach für neue Projekte adaptieren ließen, und auch Beethovens Werkstatt hat in diesem Feld wiederholt neue Ansätze erprobt. Der Vortrag wird diese Erfahrungen dokumentieren und problematisieren, was für eine weitergehend standardisierte Lösung notwendig wäre.

Bibliographie

Altenhöner, Reinhard / Blümel, Ina / Boehm, Franziska / Bove, Jens / Bicher, Katrin / Bracht, Christian / Brand, Ortrun / Dieckmann, Lisa / Effinger, Maria / Hagen, Malte / Hammes, Andrea / Heller, Lambert / Kailus, Angela / Kohle, Hubertus / Ludwig, Jens / Münzmay, Andreas / Pitroff, Sarah / Razum, Matthias / Röwenstrunk, Daniel / Sack, Harald / Simon, Holger / Schmidt, Dörte / Schrader, Torsten / Walzel, Annika-Valeska / Wiermann, Barbara (2020): „NFDI4Culture - Consortium for research data on material and immaterial cultural heritage“,

in: Research Ideas and Outcomes 6 (Juli 2020). <https://doi.org/10.3897/rio.6.e57036>

Droese, Janine / Münzmay, Andreas (2015): „Pfade im editorischen Netz. Überlegungen zur Pragmatik des editorischen Hyperlinks am Beispiel der Comédie en vaudevilles Annette et Lubin (1762)“, in: Editio. Internationales Jahrbuch für Editions-wissenschaft 29: 85–102. <https://doi.org/10.1515/editio-2015-007>.

Fellerer, Karl Gustav (1980): „Werk – Edition – Interpretation“, in: Martin Bente (ed.): Musik, Edition, Interpretation. Gedenkschrift Günter Henle, München: Henle 180–192.

Hase, Oskar von (1968): „Breitkopf & Härtel. Gedenkschrift und Arbeitsbericht. Zweiter Band: 1828 bis 1918“, Wiesbaden: Breitkopf und Härtel.

Kepper, Johannes (2011): „Musikedition im Zeichen neuer Medien. Historische Entwicklung und gegenwärtige Perspektiven musikalischer Gesamtausgaben“, Norderstedt: BoD.

Kepper, Johannes / Pugin, Laurent (2017): „Was ist eine Digitale Edition? Versuch einer Positionsbestimmung“, in: Acquavella-Rauch, Stefanie / Münzmay, Andreas (eds.), Digitalität in der Musikwissenschaft, Themenheft, Musiktheorie. Zeitschrift für Musikwissenschaft 32/4: 347–363.

Kepper, Johannes / Sänger, Richard (2017): „Encoding Genetical Processes“, in: Giuliano Di Bacco, Giuliano / Kepper, Johannes / Roland, Perry D. (eds.): Music Encoding Conference Proceedings. 2015, 2016 and 2017: 37–44 <https://bibdorm.bsb-muenchen.de/bv/BV045900855>.

Kepper, Johannes / Cox, Susanne (2021): „Encoding Genetic Processes II“, in: Münnich, Stefan / Rizo, David (eds.): Music Encoding Conference Proceedings 2021. 19–22 July, 2021 University of Alicante (Spain): On-site & Online 85–95 <https://hcommons.org/deposits/objects/hc:45962/datastreams/CONTENT/content>.

Münzmay, Andreas / Siegert, Christine (2019): „Phonographischer Text, Interpretation und Aufführungsmaterial als kritisch edierbarer Sachzusammenhang. Ein Beitrag zur Theorie der Edition von Klangdokumenten“, in: Editio. Internationales Jahrbuch für Editions-wissenschaft 33: 10–30. <https://doi.org/10.1515/editio-2019-0002>.

Orcalli, Angelo (2017) „Recorded Music: from the Ethics of Preservation to the Critical Editing“, in: Orcalli, Angelo / Cossettini, Luca (eds.): Sounds, Voices and Codes from the Twentieth Century. The critical editing of music at Mirage. Udine: Mirage: 3–81 <http://mirage.uniud.it/content/sounds-voices-and-codes-twentieth-century-critical-editing-music-mirage>.

Pasdzierny, Matthias (2019): „Tonband, Partitur, Aufführung. Medien- und musikphilologische Überlegungen zur Edition von Bernd Alois Zimmermanns Requiem für einen jungen Dichter“, in: Betzwieser, Thomas / Schneider, Markus (eds.): Aufführung und Edition. Editio Beihefte 46. Berlin: De Gruyter 219–233.

Popp, Susanne (2017): „Musikdrucke“, in: Appel, Bernhard R. / Emans, Reinmar (eds.): Musikphilologie. Grundlagen – Methoden – Praxis, Laaber: Laaber 70–87.

Sahle, Patrick (2017): „Digitale Edition“ in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): Digital Humanities. Eine Einführung, Stuttgart: J. B. Metzler 223–249.

Stadler, Peter (2019): „Musikwissenschaft und Digital Humanities“, in: Hentschel, Frank (ed.): Historische Musikwissenschaft. Gegenstand – Geschichte – Methodik. Laaber: Laaber 330–339.

Tennie, Claudio / Call, Josep / Tomasello, Michael (2009): „Ratcheting up the ratchet: on the evolution of cumulative culture“, in: Philosophical Transactions of the Royal Society B 364: 2405–2415 [10.1098/rstb.2009.0052](https://doi.org/10.1098/rstb.2009.0052).

Open Science Prinzipien und interdisziplinäre Kollaboration in D-WISE: Zwischen Hermeneutik und Digitaler Methode in der Diskursanalyse

Eiser, Isabel

isabel.eiser@uni-hamburg.de
Projekt D-WISE, Empirische Kulturwissenschaften,
Universität Hamburg, Deutschland

Fischer, Tim

tim.fischer@uni-hamburg.de
Projekt D-WISE, Informatik, Language Technology
Group, Universität Hamburg, Deutschland

Schneider, Florian

florian.schneider-1@uni-hamburg.de
Projekt D-WISE, Informatik, Language Technology
Group, Universität Hamburg, Deutschland

Koch, Gertraud

gertraud.koch@uni-hamburg.de
Projekt D-WISE, Empirische Kulturwissenschaften,
Universität Hamburg, Deutschland

Biemann, Chris

chris.biemann@uni-hamburg.de
Projekt D-WISE, Informatik, Language Technology
Group, Universität Hamburg, Deutschland

Petersen Frey, Fynn

fynn.petersen-frey@uni-hamburg.de
Projekt D-WISE, Informatik, Language Technology
Group, Universität Hamburg, Deutschland

Das D-WISE Projekt

Das Verbundprojekt D-WISE (www.dwise.uni-hamburg.de) ist ein 2021 gestartetes BMBF gefördertes interdisziplinäres Kooperationsprojekt an der Universität Hamburg zwischen den Geisteswissenschaften, vertreten durch das Institut für Empirische Kulturwissenschaft, und dem Fachbereich Informatik, vertreten durch die Arbeitsgruppe Language Technology. D-WISE entwickelt eine prototypische Arbeitsumgebung, die D-WISE Tool Suite (DWTS), in der sowohl innovative KI-Verfahren für die Analyse von multimodalen Daten entwickelt als auch hermeneutische Analysen der wissenssoziologischen Diskursanalyse digital unterstützt und erweitert werden.¹ Dabei wird in dem Projekt herausgearbeitet, zu welchen Zwecken, Zeitpunkten und in welcher Form Digital Humanities (DH)-Verfahren in qualitative diskursanalytische Wissensproduktion analytisch sinnvoll eingebunden werden können. Die erkenntnistheoretische Reflexion und Weiterentwicklung hermeneutischer Methoden in der Nutzung (halb-)automatisierter diskursanalytischer Forschungsprozesse sind integraler Bestandteil innerhalb der konzeptuellen Forschung des D-WISE Projekts. Es sollen Fragen danach beantwortet werden, wie Automatisierung und DH-Methoden sinnvoll in qualitative diskursanalytische Ansätze und Wissensproduktion integriert werden können, welche bestehenden Methoden und Tools dafür übernommen werden können und welche dafür neu entwickelt werden müssen.

Das D-WISE Projekt zielt darauf ab jene Herausforderungen zu adressieren, denen sich Forschende im Hinblick auf methodologische Ansätze aus den Digital Humanities (DH) und der Diskursanalyse sowie dem zunehmenden Umgang mit digitalen Tools und offenen Korpora, bestehend aus heterogenen, multimodalen und großen Datenmengen gegenübersehen. Dabei zielt das Projekt auch darauf ab, den Mangel an digitalen Lösungen für die Analyse von Pluralität von Bedeutungen in multimodalen Materialien zu beheben sowie Dokumentations- und Reflexionsprozesse diskursanalytischen Arbeitens und die Weiterentwicklung der Diskursanalyse als Methode selbst zu unterstützen.

In Co-Creation Ansätzen zwischen Informatik und Geisteswissenschaften wird die diskursanalytische Arbeitsweise durch digitale Methoden erweitert und an digitale Modalitäten angepasst. Das Projekt innoviert dabei sowohl die informatische KI-Technologie kontextorientierter Embedding-Repräsentationen als auch hermeneutische Arbeitsweisen zur wissenssoziologischen Diskursanalyse. Dabei steht auch die Überbrückung der Lücke zwischen strukturellen Mustern, die mit digitalen Methoden aufgedeckt werden, und interpretativen Prozessen menschlicher Bedeutungsproduktion im Mittelpunkt des kollaborativen Ansatzes der Empirischen Kulturwissenschaften und Computerlinguistik innerhalb des D-WISE-Projekts. Dabei legt der D-WISE-Ansatz einen besonderen Akzent auf die Interaktion zwischen Mensch und Algorithmus bzw. Maschine, um KI-Forschungssysteme zu verbessern, die sich auf den menschlichen Erkenntnisprozess auswirken (Oeste-Reiß et al. 2021: 221, 149); Als Datenanalysewerkzeug mit Komponenten des maschinellen Lernens wird die DWTS zukünftig einen interaktiven und wechselseitigen Prozess

bieten, in dem maschinell erstellte Vorschläge von Annotator:innen akzeptiert, abgelehnt oder korrigiert werden können, wodurch das maschinelle Lernen mit der Zeit verbessert wird (Yimam et al. 2017, Koch et al. 2022: 77).

Computergestützte Diskursanalyse und ihre Entwicklungspotenziale

Anthropologisch und soziologisch orientierte Diskursanalysen haben zum Ziel diskursive Konstruktionen von Wirklichkeit, von Wissen oder von Macht-/Wissensbeziehungen zu untersuchen, wobei sie von der Annahme ausgehen, dass der Sprachgebrauch in diskursiven Praktiken eben jene Gegenstände konstituiert, die sie als Wissen behandelt. Grundlage digitaler Ansätze in der Diskursanalyse sind Kodierungsprozesse in Anlehnung an die Grounded Theory: Annotieren und Extrahieren von Zitaten aus Transkripten, Erstellen von Codes, Sub-Codes und Parent Codes und der Identifizierung von Konzepten oder Themen. Die Analyse von Diskursen erfolgt in zirkulären Prozessen der Suche, Auswahl, Analyse und Interpretation von Forschungsdaten, unterstützt durch Literaturarbeit, und wird iterativ durchgeführt, bis die Forschungsfrage beantwortet ist.

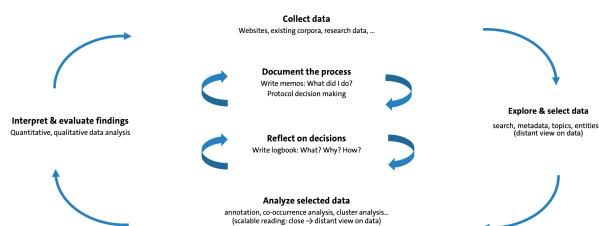


Abb. 1: Schaubild des D-WISE Projektziels der Unterstützung und Erweiterung soziologisch und anthropologisch orientierter Diskursanalysen, © D-WISE

Die Verwendung von digitalen Werkzeugen in der Analyse von Diskursen ist weit verbreitet: Ob Organisation, Kodierung, Annotation oder Analyse von Forschungsmaterial – digitale Tools bieten eine nützliche Ergänzung in der qualitativen und quantitativen Datenanalyse. Lizenzbasierte und kostenpflichtige qualitative Datenanalysetools wie MAXQDA, Atlas.ti oder NVivo sind bemüht umfangreiche ‚All-in-one-Lösungen‘ zu bieten und verschiedenen Anforderungen gleichzeitig gerecht zu werden. Zur Verfügung stehende Open Source Tools hingegen konzentrieren sich oftmals auf wenige spezialisierte Funktionen, die dafür teilweise sehr gut beherrscht werden.² Catma oder WebAnno sind entsprechende Beispiele für offene Werkzeuge, die jedoch noch viele Wünsche für die qualitative Diskursanalyse offenlassen – Lücken, die mit D-WISE geschlossen werden sollen (Koch et al. 2022: 81; Gius et al. 2021; Eckart De Castilho et al. 2016).³ Dabei konzentriert sich das Projekt vornehmlich auf drei Aspekte, in denen Entwicklungspotenzial im Be-

reich digitaler Tools und semi-automatisierter Funktionen identifiziert wurden: Analysen, Kodierungen und Annotationen von multimodalen Daten, die Text, Bild, Audio und Video beinhalten; das Crawlen, Filtern und Management von großen digital zu Verfügung stehenden Datenmengen („big data“); sowie Prozesse der Dokumentation und Reflexion von diskursanalytischen Verfahren und Methoden sowie der Verwendung und des Einflusses digitaler Tools.

Digitale Diskursanalyse und Hermeneutik in D-WISE

Mit der zunehmenden Integration von IT-Werkzeugen und -Infrastrukturen in die geisteswissenschaftliche Forschung kommt es zu ontologischen Veränderungen in der Wissensproduktion; neue soziale, ethische und/oder politische Konstellationen werden verfestigt, gestört oder geschaffen (Koch 2018: 71). Die Interaktion mit und Interpretation von Datenmodellierung oder Mustererkennung als grundlegende Operationen in den Digital Humanities können die Perspektive von Geisteswissenschaftler:innen auf ihre Quellen verändern (Schwandt 2020: 19). Die zunehmende Integration solcher strukturellen Ansätze in den hermeneutischen Ansätzen diskursanalytischer Forschung kann als sich verfestigende Infrastruktur für die Sozial- und Kulturforschung verstanden werden. Dies wirft methodologische und epistemologische Fragen hinsichtlich der Erforschung sozialer Wirklichkeit und der Pluralität von Bedeutung auf und bringt die Herausforderung mit sich, dass Diskurs- und Arbeitspraktiken in Standardformen gepresst werden (Koch 2018: 70; Koch et al. 2022: 73). Die wachsende Hybridität von manuellen und digitalen Herangehensweisen im Forschungsalltag macht eine Neuverhandlung von einer Hermeneutik notwendig, die „das Dazwischen“ gegenwärtiger geisteswissenschaftlicher Praktiken problematisiert (Fickers und Tatarinov 2022: 7, 11).⁴

Im Mittelpunkt dieser Entwicklung steht auch im D-WISE Projekt die Frage, wie Menschen und digitale Strukturen interagieren und auf welche Weise diese Interaktion menschlichen und methodologischen Bedürfnissen dient (Koch et al. 2022: 70). Unterschiede in Forschungsdesign und Methodik (quantitatives bzw. maschinengestütztes ‚distant reading‘ und qualitatives bzw. individuelles ‚close reading‘ von Korpora) sowie in den Ambitionen (Auffinden allgemeiner wissenschaftlicher Gesetze versus Produktion originärer subjektiver Interpretationen in den Geisteswissenschaften) schaffen neue Herausforderungen und Entwicklungspotenziale für das Forschungs- und Tool-Design (Fickers und Tatarinov 2022: 5).

Die Digital Humanities sind davon geprägt eine vielfältige und nach wie vor im Bestehen begriffene Forschungslandschaft zu sein, die eine Reihe von geisteswissenschaftlichen und informatischen Forschungen, Methoden und Technologien umfassen. Dabei kommt immer wieder die Frage auf, inwiefern die Geisteswissenschaften durch ihre Interaktion mit Technologie, Medien und digitalen Methoden Teil an diesen haben (Schlicht 2021: 26 zit.n. Svensson; Fickers und Tatarinov 2022: 6-7). Während komplexe Mensch-Computer-Inter-

aktionen zunehmend in Forschungsprozesse eingebunden werden, können diese zumeist nur in phänomenologisch-deskriptiver und ethisch-normativer Hinsicht ausreichend verstanden werden (Fritz et al. 2020: 3). Teil dabei entstehender Reflexionsprozesse sind auch Herangehensweisen wie die sogenannten „screwneutics“, wobei in die digitalen Werkzeuge als Mittel der explorativen Hermeneutik eingetaucht wird und durch spielerisches Herumprobieren und „screwing around with data“ (Fickers und Tatarinov 2022: 9 zit.n. van Zundert 2016 und Ramsay 2014) Erkenntnisse gewonnen, ‚user stories‘ und Kritiken formuliert werden gegenüber angewandten Methoden und Tools (Koch et al. 2022: 73-76; Schwandt 2020: 20).⁵

Für die Entwicklung zentraler Innovationen für die qualitative Datenanalyse werden in D-WISE in einem konzeptuellen Rahmen der ‚manuellen Analyse‘ bereits vorhandene Features und Tools sowie deren technische und methodische Nutzbarkeit von Projektinternen und -externen Forscher:innen aus verschiedenen Bereichen der Geisteswissenschaften und der Informatik erprobt. Funktionalitäten wie das Kodieren, Dokumentieren oder Visualisieren werden in Co-Creative Prozessen auf ihre Nützlichkeit für Diskursanalyse im Allgemeinen sowie sinnvolle Implementierung für die Tool Suite getestet, eruiert und reflektiert. Diese Integration von Erfahrungen und Feedbackschleifen von Wissenschaftler:innen mit verschiedenen Forschungsschwerpunkten macht die DWTS robust für disziplinübergreifende diskursanalytische Ansätze.

Diese wechselseitige Arbeitsweise, die sich in der Entwicklung der D-WISE Tool Suite manifestiert hat, wird als methodischer Ansatz zur Kombination von manuellen und digitalen Analysemethoden in der qualitativen und quantitativen Sozial- und Kulturforschung angewandt. Kern dieser konzeptuellen Forschung beinhaltet auch die epistemologische Reflexion über Relevanz und Validität der erhobenen Daten, der angewandten Methoden sowie verwendeten Tools und hermeneutischen Prozesse. In diesem wechselseitigen Ansatz von manueller und digitaler Arbeit sollen die User und der Prozess der Diskursanalyse mit digitalen Lösungen, Algorithmen und KI effektiv unterstützt werden und jene Forschungsaspekte ausgemacht werden, in denen hermeneutische Interpretationen notwendig werden. Damit einher geht eine Reflexion darüber, wie digitale Tools und Methoden diskursanalytische Forschungsprozesse generell verändern. Direkt im Forschungsprozess entstanden, legt eine solche Zirkulation von Agency in der Mensch-Computer-Interaktion den Grundstein für diesen Ansatz von wechselseitiger Beeinflussung, Anreicherung und entsprechenden Reflexionsprozessen.⁶ Während für die Entdeckung und Produktion von Wissen, von Innovation, Integration und Wiederverwendung von Daten eine gute Verwaltung zentral ist, wird die D-WISE Tool Suite den Anforderungen der Prinzipien der Open Science angepasst, um den Forderungen nach einem größtmöglichen und inklusiven Nutzen von Daten, Kollaboration und ihren Werkzeugen gerecht zu werden (Wilkinson et al. 2016; Schlicht 2021: 29).⁷

Neben der Tool Entwicklung findet sich der Wert des Projekts in der Konzeptualisierung selbst. Zentraler Aspekt ist dabei der hermeneutische Ansatz in D-WISE:

Die epistemologische Reflektion und die Weiterentwicklung hermeneutischer Methoden und semi-automatisierter Prozesse. Mit diesem methodisch-konzeptuellen Ansatz will das Projekt die Lücke schließen zwischen manuellen und digitalen diskursanalytischen Ansätzen: Dies beinhaltet auf forschersicher Projektseite einerseits den interdisziplinären Brückenschlag durch Co-Creation Konzepte sowie andererseits die Überbrückung auf technischer und methodisch-diskursanalytischer Ebene: zwischen strukturellen Mustern, die durch digitale Methoden erkannt werden sowie hermeneutischen und interpretativen Prozessen der menschlichen Bedeutungsgebung und Wissensproduktion – zwischen qualitativen Ansätzen und ‚close readings‘ und quantitativen Ansätzen des ‚distant readings‘. Teil dessen ist auch die Reflexion der Veränderung empirischer Forschung durch digitale Technologien.

Dabei dient die Tool Suite als Bindeglied zwischen manueller und digitaler sowie quantitativer und qualitativer Analyse; Durch Hin- und Anleitung zu proprietären Tools, durch semi-automatisierte Kodierungen, Visualisierungen und Mappings können Lücken zwischen quantitativer und qualitativer Analyse überbrückt werden. Durch Weiterentwicklungen von semi-automatisierter Annotation und der Produktion von Sub-Korpora oder Code-Gruppen werden qualitative Verfahren beschleunigt und es kann sich schrittweise einem axialen Kodieren und einer qualitativen Feinanalyse angenähert werden. Der forschersiche Schwerpunkt in D-WISE zu der Frage, welche digitalen Tools und Funktionen sinnvoll eingesetzt werden können und wo hingegen der Mensch mit hermeneutischer und interpretativer Leistung sowie mit Skalierungen und Medienwechsel eingreifen muss, ist der methodisch-theoretische und konzeptuelle Rahmen für die Erforschung dieser Überbrückung von quantitativ und qualitativ, von ‚distant‘ und ‚close reading‘, von digital und manuell geleiteten sowie hermeneutisch-interpretativen Zugängen zu Diskursanalyse.

Die D-WISE Tool Suite als Open Source Software

Der Prototyp der DWTS wird als webbasierte Open Access Serverarchitektur entwickelt und auf Basis bereits erfolgreich etablierter DH-Methoden und digitalen Tools und deren Reflektion und Evaluation konzipiert.⁸ Die Tool Suite ist projekt-zentriert aufgesetzt, und soll kollaboratives Arbeiten und alle diskursanalytischen Schritte unterstützen. Gestartet im Mai 2021, befindet sich die Tool Suite in einem frühen prototypischen Stadium, in dem zunächst erste grundlegende Funktionen und erwartete Standards umgesetzt wurden, wie das Suchen, Filtern, Kodieren, Annotieren sowie Dokumentieren.⁹

Der im Rahmen des D-WISE Projektes entwickelte Prototyp der D-WISE Tool Suite wird angelegt als Free and Open Source Software (FOSS) zur digitalen Unterstützung qualitativer Datenanalyse und Diskursanalysen (Schlicht 2021: 34). Eines der Ziele ist es dabei die DWTS so zu entwickeln, dass sie entsprechende Kriterien in Bezug auf Klimafreundlichkeit, Nachhaltigkeit, Langlebigkeit und Offenheit erfüllt (Jentsch und Porada 2020:

91).¹⁰ Dabei wird auch den FAIR-Leitprinzipien für wissenschaftliches Datenmanagement gefolgt, die entwickelt wurden, um den Umgang mit Forschungsdaten, die Ausweitung des Geltungsbereiches von Open Access sowie eine breite Implementierung der Open Science Prinzipien zu unterstützen.¹¹ Die Idee der Tool Suite im Sinne eines Werkzeug- oder Baukastens strebt an, Tools und Features sowohl innerhalb der Tool Suite anzubieten als auch die Verwendung externer Tools zu ermöglichen und anzuleiten. Dabei sollen, wo immer möglich, extern wie intern, Open-Source-Features, -Tools und -Lösungen implementiert werden (Fischer et al. 2023). Vor allem proprietäre Tools kommen dabei zum Einsatz, die Ansprüche an die Arbeitsweise der qualitativen Diskursanalyse erfüllen und digitale Verfahren entsprechend sinnvoll einbinden. Aus der Implementierung jener Open Science Prinzipien ergeben sich folglich Fragen nach Inklusion und Gleichberechtigung; Sie sollen zu einer höheren Sichtbarkeit und Auffindbarkeit führen sowie schnellere Verbreitung von Forschungsergebnissen und freien Zugang unabhängig von institutioneller Zugehörigkeit ermöglichen (Schlicht 2021: 28, 29 zit.n. Drucker 2009). Die Wahl von freier und quelloffener Software (FOSS) stellt zudem sicher, dass der Quellcode dieser Werkzeuge stets nachvollzogen werden kann, von Nutzer:innen veränderbar ist und Verarbeitungsschritte dokumentiert, reproduzierbar sowie interoperabel sind – im Idealfall über einen längeren Zeitraum über die Projektphase hinaus (Jentsch und Porada 2020: 92).¹²

Schlussbemerkung: Offener Zugang und Interdisziplinärer Austausch

Kooperatives und interdisziplinäres Arbeiten sowie Fragen nach Open Access gewinnen in den Digital Humanities zunehmend an Bedeutung. Vor allem geisteswissenschaftlich ausgerichtete Forschungsteams interdisziplinärer Projekte können einen wichtigen kritischen Beitrag leisten, um mittels der Verwendung von Werkzeugen, Paradigmen und Konzepten digitaler Technologien, die Idee von digitalen Tools, deren Zugänglichkeit und Instrumentalität neu zu überdenken. Durch die kontinuierliche Zusammenarbeit zwischen Informatik und Geisteswissenschaften wird im D-WISE Projekt sichergestellt, dass in iterativen Zirkeln der Software-Entwicklung die Umsetzung von inklusiven und möglichst unbefangenen Wissens- und Untersuchungsparadigmen unterstützt werden (Koch et al. 2022: 71; Schlicht 2021: 46; zit.n. Liu 2012: 501-502). Dieser Forschungsprozess ist technisch als zyklischer Prozess entwickelt, somit flexibel auf iteratives Arbeiten anpassbar und wiederverwendbar und entspricht selbst einem hermeneutischen Zirkel, der durch wiederholtes Hinterfragen und ständige Erweiterung des Wissensstandes immer wieder zu neuen Fragen, Erkenntnissen und damit zu einem tieferen Verständnis des Phänomens führt. Kritisch reflektiert werden dabei neben der Entwicklung der Tool Suite, auch Prozesse der Wissensproduktion und das zugänglich machen von Wissen und Werkzeugen, angewand-

ten Methoden und Fragestellungen. Die Geisteswissenschaften dienen dabei, wie Liu es formuliert, als „Vektor“ für den Import fremder Wissensparadigmen: Sie bringen neue Phänomen-Ebenen ein durch beispielsweise quantitativ definierte Strukturen, Formen und Zyklen; in den Analyse- und Interpretationsverfahren werden durch digitale Technologien Modellierungen eingeführt und in der Wissensproduktion wird das Repertoire von Ergebnispräsentationen um Visualisierungen, Programmen oder Datenbanken erweitert (Schlicht 2021: 47; zit.n. Liu 2009: 27).

Fußnoten

1. Das Projekt orientiert sich an soziologisch und anthropologisch ausgerichteten Diskursanalysen, vor allem an dem von Reiner Keller aufgesetzten Forschungsprogramm „Wissenssoziologische Diskursanalyse“, siehe z.B. Keller 2011.
2. Digitale Tools unterstützen durch (halb-)automatisierte Verfahren die methodisch geleitete Reduktion größerer Datenmengen oder dienen der Identifikation potenziell relevanter Daten für die manuelle Bearbeitung, Analyse und Interpretation.
3. Als weitere Tools zu nennen sind beispielsweise Ant-Conc, Voyant, New/s/leak, sentText oder neue Projekte wie Label Sleuth und InsightNet.
4. Fickers und Tatarinov glauben, dass, „indem die digitale Hermeneutik den intellektuellen Raum zwischen dem ‚Unbekannten‘ und dem ‚Vertrauten‘ erforscht, nimmt sie eben jenen Raum ein, den der Wissensphilosoph Hans-Georg Gadamer als ‚Ort‘ der Hermeneutik identifiziert hatte - nämlich ihr Dazwischen.“ (Fickers und Tatarinov 2022: 11; zit.n. Gadamer).
5. ‚User stories‘ sind durch Endnutzer:innen formulierte Software-Anforderungen gerichtet an Software Developer zur Weiterentwicklung zentraler Funktionen.
6. In der Mensch-Computer-Interaktion unterstützt einerseits der Mensch computationale Vorgänge und verbessert das maschinelle Lernen und digitale Tools und andererseits unterstützen digitale Tools menschliche Forschungsprozesse, indem sie diese z.B. beschleunigen oder vereinfachen.
7. Der Begriff Open Science verbindet als Oberbegriff verschiedene Ideen rund um den digitalen und freien Zugang.
8. Neben den von externen Fellows eingebrachten Erfahrungen mit digitalen Tools wird in internen Tool-Analysen vor allem MAXQDA zur Testung und Evaluation herangezogen.
9. Eine Demo der Tool Suite sowie erste Evaluationen, bearbeitete Korpora und Ergebnisse werden Anfang 2023 erwartet.
10. Die Autoren weisen auf die Auswahlkriterien in dem Leitfaden „Software Evaluation Criteria-based Assessment“ hin, veröffentlicht von dem Software Sustainability Institute.
11. Die FAIR Leitprinzipien besagen, dass die Organisation von Daten so ausgeführt werden sollte, dass Daten „Findable, Accessible, Interoperable, and Reusable“ sind (Schlicht 2021: 33-34; Wilkinson 2016: 2).
12. Quellcodes und Verarbeitungsschritte werden über Github dokumentiert (<https://github.com/uhh-it/dwts>). Die Tool Suite kann von Nutzer:innen über eigene Server

installiert werden, auch um so einen sicheren Umgang mit sensiblen Daten zu ermöglichen.

Bibliographie

- Eckart De Castilho, Richard, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank und Chris Biemann. 2016. „A Web-Based Tool for the Integrated Annotation of Semantic and Syntactic Structures.“ In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. <https://aclanthology.org/W16-4011>.
- Fickers, Andreas und Juliane Tatarinov. 2022. *Digital History and Hermeneutics: Between Theory and Practice*. De Gruyter. <https://doi.org/10.1515/9783110723991>.
- Fischer, Tim, Isabel Eiser, Florian Schneider, Fynn Petersen-Frey, Chris Biemann, Gertraud Koch. 2023. „D-WISE – Wissenssoziologische Diskursanalyse“. *DHd2023 Luxemburg/Trier. Open Humanities – Open Culture. Konferenzabstracts*.
- Fritz, Alexis, Wiebke Brandt, Henner Gimpel und Sarah Bayer. 2020. „Moral Agency without Responsibility? Analysis of Three Ethical Models of Human-Computer Interaction in Times of Artificial Intelligence (AI).“ *De Ethica* 6(1): 3–22. <https://doi.org/10.3384/de-ethica.2001-8819.20613>.
- Gius, Evelyn et al. Catma 6 (Version 6.3). Zenodo, Online (2021).
- Jentsch, Patrick und Stephan Porada. 2020. „From Text to Data: Digitization, Text Analysis and Corpus Linguistics.“ In *Digital Methods in the Humanities: Challenges, Ideas, Perspectives*, hg. von Silke Schwandt, 89–128. Bielefeld: transcript Verlag. <https://doi.org/10.14361/9783839454190-004>.
- Keller, Reiner. 2011. *Wissenssoziologische Diskursanalyse: Grundlegung eines Forschungsprogramms*. Wiesbaden: VS Verlag für Sozialwissenschaften. <https://doi.org/doi:10.1007/978-3-531-92058-0>.
- Koch, Gertraud. 2018. „The Ethnography of Infrastructures. Digital Humanities and Cultural Anthropology.“ In *Cultural Heritage Infrastructures in Digital Humanities*, hg. von Agiatis Benardou, Erik Champion, Costis Dallas, and Lorna M. Hughes, 63–81. Abingdon, Oxon: Routledge. <https://doi.org/doi:10.4324/9781315575278-5>.
- Koch, Gertraud, Chris Biemann, Isabel Eiser, Tim Fischer, Florian Schneider, Teresa Stumpf, und Alejandra Tijerina García. 2022. „D-WISE Tool Suite for the Sociology of Knowledge Approach to Discourse.“ In *Culture and Computing*, hg von Matthias Rauterberg, 68–83. Cham: Springer International Publishing. https://doi.org/doi:10.1007/978-3-031-05434-1_5.
- Oeste-Reiß, Sarah, Eva Bittner, Izabel Cvetkovic, Andreas Günther, Jan Marco Leimeister, Lucas Memmert, Anja Ott, Bernhard Sick, Kathrin Wolter. 2021. Hybride Wissensarbeit. *Informatik Spektrum* 44(3): 148–152. <https://doi.org/doi:10.1007/s00287-021-01352-0>.
- Schlicht, Helene. 2020. „Open Access, Open Data, Open Software?: Proprietary Tools and Their Restrictions.“ In *Digital Methods in the Humanities: Challenges, Ideas, Perspectives*, hg. von Silke Schwandt, 25–58. Bielefeld: Bielefeld University Press. <https://doi.org/doi:10.1515/9783839454190-002>.

Schwandt, Silke. 2020. *Digital Methods in the Humanities: Challenges, Ideas, Perspectives*. Bielefeld: transcript Verlag/ Bielefeld University Press. <https://doi.org/10.14361/9783839454190>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3(1): 1-9. <https://doi.org/10.1038/sdata.2016.18>.

Yimam, Seid Muhie, Steffen Remus, Alexander Panchenko, Andreas Holzinger, and Chris Bieermann. 2017. "Entity-Centric Information Access with the Human-in-the-Loop for the Biomedical Domains." In *Proceedings of the Biomedical NLP Workshop associated with RANLP*, 42-48. <https://doi.org/10.26615/978-954-452-044-1#006>.

Oral History auf dem Weg zu Big Data: menschliche und maschinelle Annotation lebensgeschichtlicher Interviews im Vergleich

Egger, Nils

nils.egger@soziologie.uni-muenchen.de
LMU München, Deutschland

Franken, Lina

lina.franken@soziologie.uni-muenchen.de
LMU München, Deutschland

Möbus, Dennis

dennis.moebus@fernuni-hagen.de
FernUniversität in Hagen, Deutschland

Schmid, Florian

florian.schmid@gsi.uni-muenchen.de
LMU München, Deutschland

Oral History sieht sich durch die Digitalisierung großen Veränderungen ausgesetzt. Angesichts der digitalen Aufnahmetechnologien stellte sich schon vor Jahren die Frage nach der Archivierung der „born digital“ Quellen. Diese Herausforderung bietet aber auch neue Chancen und Perspektiven: etwa die automatische Spracherkennung digitaler Audiosignale, computergestützte Transkription und die komfortable Suche in Online-Repositories (Leh 2015, Leh 2018, Gref/Köhler/Leh 2017). In den letzten Jahren konnten große Fortschritte im Bereich

der automatischen Spracherkennung gemacht werden, weshalb auch die Zahl digitaler Transkripte deutlich angestiegen ist (Köhler et al. 2019). Mit Oral-History.Digital entsteht zurzeit das größte Portal zur Archivierung und Präsentation lebensgeschichtlicher Interviews in digitaler Form im deutschsprachigen Raum – damit sind auch die historische und qualitative Forschung auf dem Weg in die „Big Data“ (Graham et al. 2015).¹ Um die wachsende Quellenbasis erschließen zu können, bedarf es perspektivisch automatisierter Verfahren, die das klassisch hermeneutische Arbeiten ergänzen. Mit den Digital Humanities haben Verfahren maschinellen Lernens Einzug in die Oral History gehalten, um Texte systematisch inhaltlich zu analysieren.

Als Heuristik zur Erschließung großer Textkorpora hat sich etwa das Topic Modeling etabliert (Graham et al. 2015, Lemke/Wiedemann 2016, Adelmann et al. 2019). Mit diesem Verfahren werden über Wahrscheinlichkeitsrechnung Gruppen von Wörtern extrahiert, die miteinander in Zusammenhang stehen. Gut trainierte Topic Models ermöglichen zum Beispiel die Extraktion thematischer Zusammenhänge aus kompletten Sammlungen lebensgeschichtlicher Interviews und vermitteln einen ersten inhaltlichen Überblick (Hodel et al. 2022). Darüber hinaus können bei explorativer Durchsicht der Ergebnisse unerwartete Phänomene an die Oberfläche gespült werden, die in den vielschichtigen Interviews allzu leicht verschütt gehen (Möbus 2022). Gerade für Sekundäranalysen ist das Verfahren deshalb vielversprechend (Franken 2022).

Allerdings wird im Machine Learning noch zu selten qualitativ evaluiert, um die automatisch generierten Ergebnisse zu validieren (Dobson 2021). Des Weiteren mangelt es in den DH an systematischen Studien, die computationale und menschliche Inhaltserschließung vergleichen, auch wenn erste Ansätze bestehen (Andorfer 2017, Baumer 2017, Fechner/Weiß 2017, Andresen et al. 2020). Schließlich ist die Hemmschwelle zum Einstieg in die Verwendung digitaler Methoden bei qualitativ Forschenden besonders hoch (Franken 2020). Um diesen Desiderata und Vorbehalten zu begegnen, wurde das Potential des Topic Modeling für die Aufbereitung größerer Datenmengen nach Grounded Theory durch die Autor:innen systematisch getestet. In einem am Turing-Test orientierten Versuchsaufbau haben wir im Rahmen eines zweitägigen Workshops lebensgeschichtliche Interviews verschlagwortet – einmal klassisch qualitativ (also manuell), einmal auf Grundlage von Topic Modeling (also maschinell). Leitendes Erkenntnisinteresse des Workshops war, wie das maschinelle Verfahren qualitative Analysen bereichern kann, welche Unterschiede entstehen, wenn Interviewmaterialien durch ein Topic Modeling strukturiert gesichtet werden und wie sich Perspektiven auf (unbekannten) Text unterscheiden. Topic Modeling wurde als Verfahren gewählt, weil es einen intuitiven Zugang zu großen Textmengen ermöglicht und zudem mit Interviewtranskripten gut zurechtkommt (was viele computerlinguistische Verfahren aufgrund der abweichenden Satzstruktur bisher leider nur begrenzt tun). Es ging uns dabei nicht darum, Grounded Theory und Topic Modeling als vergleichbar zu setzen, sondern zu prüfen, wie sich die beiden Methoden ergänzen und wie sich Sinnstiftungsprozesse je nach Zugang ausgestalten. Die schlussendlichen Annotationen an den Transkripten wurden mit

im Vorfeld vorbereiteten Projekten in Catma (Gius et al. 2022) durch die Gruppen umgesetzt.

Im Mittelpunkt des Workshops stand das konkrete Arbeiten am Textkorpus aus unterschiedlichen Perspektiven sowie die gemeinsame Reflexion dieser Perspektiven. Der Beitrag stellt die Ergebnisse dieses Experiments vor und diskutiert die Mehrwerte sowie weitere Anschlussmöglichkeiten. Dafür wurden insgesamt sechzehn Teilnehmende in vier interdisziplinäre Gruppen mit jeweils gemischten Kompetenzen und Vorwissen unterteilt, sodass beispielsweise Teilnehmende mit Erfahrung in qualitativer Forschung und Teilnehmende, die bereits mit Topic Modeling vertraut waren, zusammenarbeiteten. Insgesamt setzten sich die Gruppen sehr heterogen aus Studierenden, Promovierenden und Promovierten zusammen.

Gearbeitet wurde parallel in den Gruppen mit einem umfangreichen, digital erschlossenen Korpus von Transkripten lebensgeschichtlicher Interviews zur Sozialgeschichte des 20. Jahrhunderts, dem Bestand Lebensgeschichte und Sozialkultur im Ruhrgebiet (LUSIR). Dieser geht zurück auf das erste großangelegte Oral-History-Projekt in Deutschland, das von Lutz Niethammer und Alexander von Plato 1981 bis 1988 durchgeführt wurde (Niethammer 1983). Die Interviews sind transkribiert und mit Timecodes versehen und können über das Archiv "Deutsches Gedächtnis" online eingesehen und angehört werden.² Für das dem Experiment zugrunde liegende Sample wurden 166 Volltexte herangezogen. Durch die Laufzeit der Interviews von bis zu acht Stunden hat das Korpus einen Umfang von 3,7 Millionen Wörtern, wovon nach Stopwordbereinigung gut 700.000 übrigbleiben. Das Topic Modeling wurde mit Mallet, allerdings mit Hilfe des Gensim-Wrappers in Python, umgesetzt, nutzt somit LDA mit Gibbs-Sampling als Inferenzalgorithmus. Nach der Evaluation verschiedener Modelle und umfangreichem Parametertuning konnte festgestellt werden, dass eine Aufteilung der Interviews in kürzere Einheiten (Chunks) zu 25 Sätzen inhaltlich konsistente und aussagekräftige Topics hervorbringt, als optimale Topic-Anzahl hat sich 50 herausgestellt (vgl. ausführlich: Hodel et al. 2022, 188f., 194f.).

Als Sample hatten die Organisator:innen im Vorfeld in einer Pilotstudie zahlreiche Interviewpassagen gesichtet, die mit arbeitsspezifischen Topics gelabelt waren, und ein Sample zusammengestellt. In zwei der vier Gruppen wurden drei ausgewählte Interviewpassagen zunächst manuell annotiert (Kategorienvergabe oder Codierung), während die beiden anderen Gruppen mit einer Erschließung der noch unbekannten Texte durch das vortrainierte Topic Modeling starteten. Danach tauschten die Gruppen die Rollen, um die Unterschiede systematisch vergleichen zu können. Als Fragestellung wurde das Thema „Arbeit“ in all seinen Facetten gesetzt, um vergleichbare Ergebnisse zu erhalten. Der exemplarische Zugang wurde gewählt, weil der arbeitsspezifische Wandel im 20. Jahrhundert gesellschaftlich relevant ist (Stichwort: vom Normalarbeitsverhältnis zu prekärer und/oder entgrenzter Arbeit). In Oral-History-Interviews finden sich fast durchgängig Aussagen zur persönlichen Arbeitsbiografie und Einschätzungen zum Wandel des eigenen Arbeitsumfeldes. Sie sind deshalb als Daten für die computationale Analyse zum Thema Arbeit besonders vielversprechend.

Die qualitative Annotation erfolgte auf Grundlage der Grounded Theory (Glaser/Strauss 2010 [1967]; Charmaz 2014). Diese ist als Analysemethode in der qualitativen Sozialforschung und den empirischen Kulturwissenschaften weit verbreitet und auch für die Erweiterung qualitativer Forschungsprozesse gut geeignet (Franken 2022). Sie ermöglicht ein induktives Vorgehen, das aus dem Material heraus Bedeutungen und Kontexte erschließt und besonders in offenen und selektiven Annotationen als Prozess (Franken/Koch/Zinsmeister 2020) Teil des Erkenntnisprozesses ist. Allerdings wurde das Methodensetting im Workshop nicht in seiner vollen Komplexität realisiert, sondern für den experimentellen Aufbau auf den Schritt des offenen Annotierens reduziert. Dabei werden Textabschnitte gelesen und aus dem hermeneutischen Sinnerschließen heraus Kategorien gebildet. Diese werden direkt an einzelne Textstellen vergeben, so dass das Kategoriensystem bei der Texterschließung nach und nach wächst und strukturiert wird (Holton 2007). Weitere Schritte, wie das theoretische Sampling, auf dessen Grundlage zentrale Quellen aus Korpora ausgewählt werden (Morse 2007), oder das anschließende selektive Annotieren, mit dem das Kategorienset je nach Erkenntnisinteresse wieder reduziert wird, wurden aufgrund der begrenzten Zeit im Workshop nicht realisiert. Auch eine die Analyse begleitende Verschriftlichung von Gedanken in Memos (Lempert 2007) wurde nicht umgesetzt, da eine umfassende Auswertung der im Vorfeld ausgewählten Textstellen nicht Ziel war, sondern der Vergleich der unterschiedlichen Texterschließungen im Mittelpunkt stand. Konkret war der Arbeitsauftrag an die Gruppenmitglieder, die Textstellen zu lesen und offene Kategorien zu vergeben, um die im Text enthaltenen Inhalte möglichst gut zu beschreiben. Im Anschluss an einzeln umgesetzte Annotationen vergaben die Gruppen dann in einer Diskussion gemeinsame Kategorien an den Textstellen, abstrahierten also von der individuellen Interpretation.

Die Analyse der Topic Models erfolgte zunächst auf globaler Ebene. Dazu standen in vorher vorbereiteten Jupyter-Notebooks verschiedene Funktionen zur Verfügung: Topiclisten mit variabler Keyword-Anzahl, Balkendiagramme, welche die Topic-Verteilung zeigen, und Heatmaps, die einerseits die globale Verteilung der Topics auf die Interviews, andererseits die Verteilung der Topics über die in Textpassagen zerteilten Interviews im zeitlichen Verlauf zeigen. Zunächst wurden die Ergebnisse des Topic Modelings - also je eine Wortliste zu jedem der fünfzig Topics - einzeln gesichtet und dann in der Gruppe diskutiert. Topics mit Bezug zum Thema „Arbeit“ wurden anschließend mit einem Schlagwort gelabelt. So wurde dem Topic „chef“, „büro“, „angestellt“, „abteilung“, „thysen“, „sekretärin“, „abteilungen“, „arbeit“, „herren“, „damen“ [...] von einer Gruppe etwa die Kategorie „Anstellung/Verwaltung“ zugewiesen, dem Topic „betriebsrat“, „gewerkschaft“, „betrieb“, „kollegen“, „gewerkschaften“, „betriebsräte“, „vorsitzend“, „wählen“, „metall“, „belegschaft“ [...] die Kategorie „Interessenvertretung/Betriebsperspektive“. Anschließend wurden mit Hilfe des Jupyter-Notebooks stichprobenartig Interviewpassagen ausgegeben, die den relevanten Topics zugeordnet waren, um die Qualität der Topics qualitativ zu überprüfen. Im nächsten Schritt gab die Workshopleitung die vorher ausgewählten Interviewpassagen be-

kannt, um den Vergleich mit den qualitativ arbeitenden Gruppen sicherzustellen.³ Die Teilnehmenden sollten beurteilen, ob sie anhand ihrer in der freien Exploration gesammelten Eindrücke auch zu diesem Sample gelangt wären. Abschließend wurden die Oberbegriffe der stärksten Topics einer Textpassage (Chunk) in Catma an die Interviewtranskripte getagged. Die Vergabe der Topics für die jeweiligen Interviewpassagen konnten die Teilnehmenden ebenfalls dem Jupyter-Notebook entnehmen.

Insgesamt stehen aus den Gruppen acht verschiedene Schlagwort-Sets pro Chunk zum Vergleich zur Verfügung (vier Gruppen, pro Gruppe rein menschliche Schlagwörter sowie Oberbegriffe für die Topics). Im Ergebnis lässt sich feststellen, dass für ein Topic, also eine Wortliste, zwar unterschiedliche, aber dennoch vergleichbare Oberbegriffe gewählt wurden. Beispielsweise wurden für das bereits erwähnte Topic, für das eine Gruppe den Begriff „Anstellung/Verwaltung“ vergab, von den anderen Gruppen die Oberbegriffe „Positionen in einer Firma“, „Organisation von Arbeit“ und „Arbeitsorganisation“ festgelegt. In den rein menschlich vergebenen Schlagwörtern für eine Textstelle ergeben sich ebenfalls Unterschiede, die jedoch grundsätzlich auf ein vergleichbares Textverständnis schließen lassen. Die Unterschiede der gewählten Kategorien bestehen hier eher in der thematischen Schwerpunktsetzung. Während eine Gruppe „Technische Entwicklung“ benannte, wählte eine andere Gruppe „Wandel Arbeitstechnik“ für den gleichen Textabschnitt. Bei den durch eine Gruppe benannten „Arbeitskonflikte[n]“ annotierte eine andere Gruppe „Arbeitsbedingungen“, nahm also ebenfalls ähnliche Interpretationen vor, wenn auch in geringerer Deutlichkeit. In der Annotation wurden für die maschinengenerierten Topics wesentlich allgemeinere Kategorien vergeben, die menschliche Annotation erfolgte zielgenauer und hat in verschiedenen Dimensionen die den Forschenden bekannten Kontexte (etwa zu historischen Ereignissen) einbezogen.

Im Anschluss an die Arbeit in den Gruppen wurde in einer Abschlussdiskussion das unterschiedliche Vorgehen verglichen und diskutiert. Besonders sticht hervor, dass die Teilnehmenden - unabhängig von Qualifikationsstufe und Vorwissen - übereinstimmend der Meinung waren, dass Topic Modeling die qualitative, textnahe Arbeit vorbereiten und anleiten kann. Es ermöglicht, so eine Teilnehmerin, das „Springen“ zwischen Nähe und Distanz und damit einen anders informierten Umgang mit dem Quellenmaterial. Die Teilnehmenden waren sich einig, dass die aus dem Topic Modeling erzeugten Ergebnisse ihren Interpretationsvorgang angeregt haben. Das Verfahren wurde auch als „kreative Methode“ bezeichnet. Mit ihrer Einschätzung stimmen die Teilnehmenden damit bisherigen theoretisch-konzeptionellen Überlegungen (Jacobs/Tschötschel 2019; Nelson et al. 2018) zu.

Die Gruppen, die zuerst die Topic Models betrachtet hatten, gingen gezielter an den qualitativen Arbeitsschritt. Allerdings wurde mehrfach darauf hingewiesen, dass parallel oder im Anschluss „Rücksprache mit dem Text gehalten“ werden müsse. Denn allein aus den Mustern des Topic Modelings könne man sich nicht erschließen, was die Interviewpartner:innen gemeint haben. Zudem würden sich hierdurch nur beschreibende Kategorien entwickeln, die um analytische Kategorien er-

gänzt werden müssten, wie sie für den qualitativen Interpretationsprozess typisch sind. Die Notwendigkeit der Verbindung von qualitativen und maschinellen Schritten wurde also mehrfach betont. Wie es ein Teilnehmer zusammenfasste: „Topic Modeling ist eine Suchmaschine, bei der ich Parameter gut beeinflussen kann. Es eignet sich, um Themenkomplexe zu finden, die ich mir danach anschauen kann.“ Dennoch kann Topic Modeling als der Grounded Theory in vielen Punkten entsprechend verstanden werden, da es ohne vorher gebildete Kategorien an Text herangeht und Cluster von Bedeutungen aus dem Text selbst herausstellt. Es eignet sich also für induktives Vorgehen, wie Kitchin (2014, 5) es in seiner *data driven science* fordert und auch Salganik als *empirically driven theorizing* (2018, 61) vorschlägt.

Zur weiteren Bewertung der im Workshop generierten Schlagwörter wurde im Nachgang eine Befragung unter 13 Bachelor-Studierenden der Soziologie sowie mit einigen Monaten Abstand auch unter 10 der Teilnehmenden des Workshops selbst durchgeführt. Über die Schlagwortmengen und Chunks hinweg entstanden so 1.148 Bewertungen. Es sollte die Passung der acht verschiedenen Mengen an Schlagwörtern zu jedem Textabschnitt bewertet werden. Dazu lasen die Teilnehmenden sieben Textabschnitte und bewerteten die jeweiligen Schlagwörter auf einer Skala von eins bis fünf, wobei bei den Teilnehmenden des Workshops die Schlagwörter der eigenen Gruppe jeweils ausgelassen wurden. Mit Kontrolle der kategorialen Variablen (1) Gruppe im Workshop, (2) Befragungsgruppe (Bachelor-Studierende als Referenzkategorie) und (3) Chunk hatten die mithilfe des Topic Modelings erstellten Schlagwörter im multiplen linearen Regressionsmodell eine um 0,94 ($p < 0,001$) schlechtere Bewertung auf der Skala von 1-5 als die rein menschlichen Schlagwörter. Das heißt, die Schlagwörter des Topic Modeling wurden statistisch höchst signifikant als schlechter bewertet. Allerdings variiert der Performance-Unterschied je nach Chunk: Für den Textabschnitt mit der geringsten Differenz zwischen maschinengestützter und menschlicher Annotation beträgt der Unterschied lediglich 0,18 und ist statistisch nicht signifikant. Daran wird deutlich, dass je nach Kontext des Textes Topic Modeling sehr gute Ergebnisse der inhaltlichen Vorstrukturierung liefern kann.

Die nachgängige Befragung der Workshop-Teilnehmenden umfasste zudem eine Art Turing-Test, bei dem die Befragten bestimmen sollten, ob die ihnen präsentierten Schlagwörter jeweils durch manuelle Annotation oder über das Topic Modeling entstanden sind. Cramérs V zwischen tatsächlicher und zugeschriebener Annotationsform beträgt 0,28 und beschreibt damit einen moderaten Zusammenhang.

Selbst Personen, welche die unterschiedlichen Generierungsprozesse der Schlagwörter kennen, können bei separater Betrachtung also nicht mehr eindeutig auf den Generierungsprozess schließen. Die über das Topic Modeling generierten Schlagwörter besitzen für menschliche Betrachter folglich durchaus eine sinnhafte Qualität.

Als Fazit kann festgehalten werden, dass, wenig überraschend, epistemologische Unterschiede zwischen der computationellen und der manuellen Texterschließung bestehen. Gerade am Vergleich wurde jedoch sichtbar, und von den Teilnehmenden so diskutiert, dass in beiden Zugängen wir als Menschen es sind, die subjektiv

Sinn zuschreiben und neu ordnen. Wie kontrovers das Thema „Sinnzuschreibung“ ist, zeigte sich in einer Diskussion während des Abschlussplenums: Während einige die Art der statistischen Kondensierung, wie sie im Topic Modeling durchgeführt wird, bereits als einen Akt der Interpretation auffassten, widersprachen andere vehement, dass eine Interpretation Verstehen - konkreter: Sinnverstehen - voraussetze, was bei computationellen Auswertungen nicht der Fall sei. Ein Indiz, das letzteres Argument stützt, ist die Tatsache, dass letztlich den rein deskriptiven Topics analytische Kategorien zugeordnet werden, um die Texte zu annotieren. Immerhin konnte die Auswertung des nachgelagerten Surveys zeigen, dass sich die direkt aus dem Text und die aus den Wortlisten generierten Kategorien nicht grundlegend unterscheiden, wenngleich die analytische Tiefe teils signifikant abwich. Zielführend im Sinne einer Erschließung digitaler Großbestände qualitativer Daten erscheint daher ein Mixed-Method-Approach, der die komplementären analytischen Zugänge kombiniert und etablierte Forschungsprozesse ergänzt.

Gleichzeitig hat sich die hohe Relevanz systematischer Evaluation digitaler Methoden gezeigt. Im Setup des Workshops haben sich die verschiedenen Ansätze gut ergänzt und zu einer vertieften Reflexion der Vor- und Nachteile der Zugänge geführt: So können mit Hilfe der Grounded Theory - insbesondere im diskursiven Austausch innerhalb einer Gruppe - Sinnstiftungsprozesse minutiös und akkurat herausgearbeitet werden. Das Topic Modeling hingegen spielte seine Stärken in der rasanten Verschlagwortung ganzer Interviewkorpora aus. Auf der anderen Seite ist das Arbeiten nach den Regeln der Grounded Theory äußerst zeitintensiv und konsistente Topics können Sinnstiftung suggerieren, wo letztlich reine Statistik am Werk ist.

Die Schwächen des Experiments liegen am Ende vor allem in der notwendigerweise gesetzten Ausschnitthaftigkeit des Materials. Um in der begrenzten Zeit vergleichbare Ergebnisse zu erzeugen, wurden nur Interviewausschnitte durch die Gruppen gelesen und annotiert. Das ist in der Grounded Theory unüblich, da das Ausschnitthafte den Blick auf das Material verzerrt. Auch für Ansätze in den Digital Humanities wird üblicherweise das gesamte Korpus analysiert. Für Folgeprojekte sollte ein größeres Zeitbudget eingeplant werden, um komplette Interviewtranskripte analysieren zu können.

Fußnoten

1. <https://www.oral-history.digital/> [03.08.2022]
2. <http://www.deutsches-gedaechtnis.fernuni-hagen.de> [03.08.2022]
3. https://github.com/moebusd/mensch_und_maschine.

Bibliographie

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 Per-

sonal’“. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019), hg. von Patrick Sahle. 114–116. DOI: 10.5281/zenodo.4622122.

Andorfer, Peter. 2017. „Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich“. *Zeitschrift für digitale Geisteswissenschaften* 2. DOI: 10.17175/2017_002.

Andresen, Melanie, Vauth, Michael, Zinsmeister, Heike. 2020. „Modeling Ambiguity with Many Annotators and Self-Assessments“. *Proceedings of the 14th Linguistic Annotation Workshop*, 48–59. Barcelona. <https://aclanthology.org/2020.law-1.5/>.

Baumer, Eric P. S., Mimno, David, Guha, Shion, Quan, Emily, Gay, Geri K. 2017. „Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?“. *Journal of the Association for Information Science and Technology* 68 (6), 1397–1410. DOI: 10.1002/asi.23786.

Charmaz, Kathy. 2014. *Constructing Grounded Theory. Introducing Qualitative Methods*. Los Angeles u.a.: SAGE.

Dobson, James. 2021. „Interpretable Outputs. Criteria for Machine Learning in the Humanities.“ *Digital Humanities Quarterly* 15 (2). <http://www.digitalhumanities.org/dhq/vol/15/2/000555/000555.html>.

Fechner, Martin, Weiß, Andres. 2017. „Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts“. *Zeitschrift für digitale Geisteswissenschaften* 2. DOI: 10.17175/2017_005.

Franken, Lina. 2020. „Kulturwissenschaftliches digitales Arbeiten. Qualitative Forschung als digitale Handarbeit?“ *Berliner Blätter – Ethnographische und ethnologische Beiträge* 82: 107–118. DOI: 10.18452/22125.

Franken, Lina. 2022. „Digitale Daten und Methoden als Erweiterung qualitativer Forschungsprozesse. Herausforderungen und Potenziale aus den Digital Humanities und Computational Social Sciences“. *Forum Qualitative Sozialforschung* 23 (1). DOI: 10.17169/fqs-22.2.3818.

Franken, Lina, Koch, Gertraud, Zinsmeister, Heike. 2020. „Annotationen als Instrument der Strukturierung“. In *Annotations in Scholarly Editions and Research. Function, Differentiation, Systematization*, hg. von Julia Nantke und Frederik Schlupkothén, 89–108. Berlin/München/Boston. DOI: 10.1515/9783110689112-005.

Gius, Evelyn, Meister, Jan Christoph, Meister, Malte, Petris, Marco, Bruck, Christian, Jacke, Janina, Schumacher, Mareike, Gerstorfer, Dominik, Flüh, Marie, Horstmann, Jan. 2022. *CATMA 6 (Version 6.5)*. Zenodo. DOI: 10.5281/zenodo.1470118.

Glaser, Barney G., Strauss, Anselm L. 2010 [1967]. *Grounded Theory. Strategien qualitativer Forschung*. Bern: Huber.

Graham, Shawn, Milligan, Ian, Weingart, Scott. 2015. *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

Gref, Michael, Köhler, Joachim, Leh, Almut. 2017. „KA³. Weiterentwicklung von Sprachtechnologien im Kontext der Oral History.“ *BIOS* 30 (1+2). DOI: 10.3224/bios.v30i1-2.05.

Hodel, Tobias, Möbus, Dennis, Serif, Ina. 2022. „Topic Modeling im Vergleich: Aufbereitung, Umsetzung und Interpretation unterschiedlicher historischer Textkorpora“. In *Von Menschen und Maschinen - Mensch-Maschine-Interaktion in digitalen Kulturen*, hg. von Selin Gerlek, Sarah

Kissler, Thorben Mämecke und Dennis Möbus, 181-205. Hagen. DOI: 10.57813/20220623-153139-0.

Holton, Judith A. 2007. "The Coding Process and Its Challenges". In *The SAGE Handbook of Grounded Theory*, hg. von Antony Bryant und Kathy Charmaz, 265-289. Los Angeles. DOI: 10.4135/9781848607941.

Jacobs, Thomas, Tschötschel, Robin. 2019. "Topic Models Meet Discourse Analysis. A Quantitative Tool for a Qualitative Approach". *International Journal of Social Research Methodology* 22 (5), 469-485. DOI: 10.1080/13645579.2019.1576317.

Kitchin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts". *Big Data & Society* 1 (1), 1-12. DOI: 10.1177/2053951714528481.

Leh, Almut. 2015. "Vierzig Jahre Oral History in Deutschland. Beitrag zu einer Gegenwartsdiagnose von Zeitzeugenarchiven am Beispiel des Archivs 'Deutsches Gedächtnis'". *Westfälische Forschungen. Zeitschrift des LWL-Instituts für westfälische Regionalgeschichte* 65: 255-268.

Leh, Almut. 2018. "Zeitzeugenkonserven. Interviews für nachfolgende Forschergenerationen im Archiv 'Deutsches Gedächtnis'". *Archivar* 71 (2): 155-157.

Lemke, Matthias, Wiedemann, Gregor. 2016. *Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Wiesbaden: Springer VS.

Lempert, Lora Bex. 2007. "Asking Questions of the Data. Memo Writing in the Grounded Theory Tradition". In *The SAGE Handbook of Grounded Theory*, hg. von Antony Bryant und Katy Charmaz, 245-264. Los Angeles: SAGE.

Möbus, Dennis (2022): "Holleriths Vermächtnis – ein Beitrag zur Geschichte von Frauen in der EDV. Topic Modeling als Methode digitaler Sekundäranalyse lebensgeschichtlicher Interviews". *BIOS* 33 (1). DOI: 10.3224/bios.v33i2.01#.

Morse, Janice M. 2007. "Sampling in Grounded Theory". In *The SAGE Handbook of Grounded Theory*, hg. von Antony Bryant und Katy Charmaz, 229-244. Los Angeles: SAGE.

Nelson, Laura K., Burk, Derek, Knudsen, Marcel, McCall, Leslie. 2018. "The Future of Coding. A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods". *Sociological Methods & Research* 18 (4). DOI: 10.1177/0049124118769114.

Niethammer, Lutz (Hg.). 1983. *Die Jahre weiß man nicht, wo man die heute hinsetzen soll'. Faschismuserfahrungen im Ruhrgebiet. Lebensgeschichte und Sozialkultur im Ruhrgebiet 1930-1960, Bd. 1*. Berlin, Bonn: Dietz.

PhilroBERTa: Ein multilinguales Sprachmodell zur Beantwortung philosophiehistorischer Fragestellungen

Noichl, Maximilian

noichlmax@hotmail.co.uk
Universität Wien, Österreich; Universität Bamberg, Deutschland

Panzer, Lukas

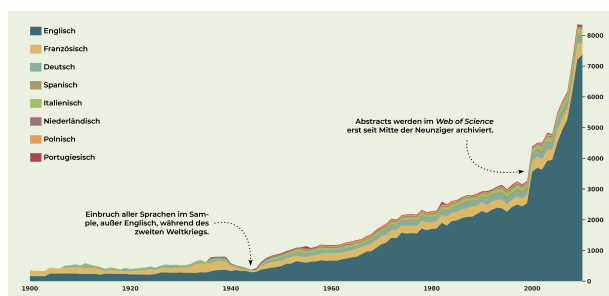
lukas.panzer@stud.uni-bamberg.de
Universität Bamberg, Deutschland

Einleitung

Die wohl bedeutendste Struktur der zeitgenössischen akademischen Philosophie ist die Trennung zwischen kontinentaler und analytischer Philosophie.

Obwohl wenig Einigkeit über die eigentliche Natur der Trennung besteht – ist es eine methodologische (vgl. Petrovich und Buonomo 2018), thematische, linguistische (vgl. Hobbs 2014), oder doch nur eine soziale? – spielt sie eine wichtige Rolle im philosophischen Berufsleben, und wird nicht nur informell verhandelt, sondern schlägt sich in den Aufnahmekriterien von Fachzeitschriften, in den Entscheidungen von Berufungskommissionen und im Aufbau zahlreicher professioneller Vereinigungen nieder.

Klarheit über die Topologie dieser Trennung zu erreichen, ist dementsprechend von großem Interesse. Dennoch besteht in der Literatur eine ausgesprochene Uneinigkeit über eine Reihe von basalen Fragen: Wann hat die Spaltung ihren Anfang genommen (man vgl. die Darstellungen von e.g. Glock 2008; Critchley 1997; 2001)? Handelt es sich überhaupt um eine Trennung, oder eher um zwei Extreme am Rande eines Kontinuums? Hat sie überhaupt noch Bestand, oder hat sich die Kluft im einundzwanzigsten Jahrhundert weitestgehend geschlossen (Bieri 2007; Beckermann 2003; Hoche 2009)?



Zusammensetzung des gereinigten Datensatz über hundertzehn Jahre. Wir beobachten eine sich massiv verstärkende Dominanz des englischen Materials über das Jahrhundert hinweg, die Teils den Datenquellen geschuldet, zu einem großen Teil aber auch aus den Bedingungen des modernen wissenschaftlichen Publizierens erwachsen ist.

Diese Uneinigkeit ist nicht überraschend. Da die Konkretisierung der Trennung in eine Zeit exponentieller Zunahme des wissenschaftlichen Outputs seit den 1950er Jahren fiel (vgl. Bornmann und Mutz 2015, siehe auch Abb. 1), und zugleich von geografischen und sprachlichen Grenzen modifiziert wurde, aber nicht mit diesen identifiziert werden kann, ist ihre Geschichtsschreibung mit außergewöhnlichen Herausforderungen konfrontiert, da sie sich nicht mit Hunderten, sondern eigentlich mit Hunderttausenden von heterogenen Quellen befassen muss, wenn sie Fragen nach der tatsächlichen disziplinären Meta-Struktur beantworten will, anstatt sich mit philosophischen Einzelschicksalen zu befassen.

In der vorliegenden Arbeit schlagen wir eine Methode zur Beantwortung solcher grenzüberschreitenden, globalen Fragestellungen vor. Mithilfe eines multilingualen Sprachmodells (PhiloBERTa), welches wir auf philosophischen Texten fein-tunen generieren wir Textvektoren von 288.546 philosophischen Texten aus den vergangenen hundert Jahren. In dem derart aufgespannten Vektorraum identifizieren wir die Achse welche der kontinental-/analytischen Unterscheidung entspricht. Indem wir die Positionierung der einzelnen Artikel auf dieser Achse erheben, können wir eine erste quantitative Einschätzungen der Topologie der analytisch/kontinentalen-Trennung vorschlagen.

Datensatz

Eine scharfe Umgrenzung des Gebietes der Philosophie, insbesondere eine, welche einem ganzen Jahrhundert und mehreren nationalen Philosophiekulturen, verpflichtet ist, ist ausgesprochen schwierig. Dementsprechend ist für die vorgelegte Arbeit der zugrundeliegende Datensatz so expansiv wie möglich gewählt worden. Der Datensatz enthält alle Texte welche in der JStor Datenbank unter der Rubrik 'Philosophy' archiviert worden sind (437.703 Rohtexte), sowie alle Abstracts aus dem *Web of Science* (WOS), welche in den Rubriken 'Philosophy' und 'History and Philosophy of Science' angesiedelt sind, oder aus in der *PhilPapers*-Journal-Liste verzeichneten Publikationen stammen (188.794 Einträge).

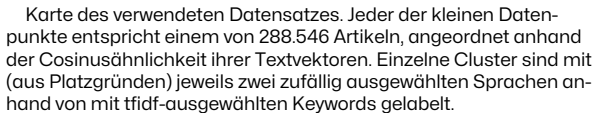
Da die Qualität der Rohdaten mäßig ist und die Fehlerquellen äußerst heterogen sind, verwenden wir einen Bulk-Labeling-Ansatz, bei dem alle Texte nach einem

BOW-Modell encodiert werden und mit UMAP (McInnes, Healy, und Melville 2018) kartographiert werden. Cluster verwendbarer Rohdaten werden in einem interaktiven Layout manuell selektiert. Dabei wurden Titeleien, publizierte Bibliographien, besonders ungenügendes OCR und nur partiell oder gar nicht erhaltene Artikel entfernt. JSTOR und WOS-Quellen wurden vereint und überschneidende Artikel angeglichen, was zu einem finalen Datensatz von 288.546 Artikeln führte.

Methode

Die Modellierung multilingualer Textcorpora stellt seit längerem ein Problem für zahlreiche Bereiche der Digital Humanities dar (Dombrowski 2020). Klassische Methoden, wie BOW-, Topic-, oder Wortvektormodelle stoßen hier an ihre Grenzen, da die von ihnen gelernten Repräsentationen hauptsächlich Unterschiede zwischen Sprachen als salienteste Muster erkennen, und das eigentliche übersprachliche Erkenntnisinteresse verdecken. Der aus diesen Problemen resultierende Fokus auf rein englischsprachiges Quellenmaterial ist unbefriedigend (Pitman und Taylor 2017; Galina Russell 2014). Multilinguale Sprachmodelle, wie z. B. xml-Roberta (Conneau u. a. 2020) sind zwar in der Lage, dieses Problem zu lösen, indem sie deckungsgleiche Vektorräume für verschiedene Sprachen bereitstellen. Ihre Anwendbarkeit auf spezifische Forschungskorpora ist allerdings begrenzt, da die notwendige Wissensrepräsentation über den spezifischen Textgehalt nicht gegeben ist. Das fine-tuning solcher Modelle auf den Forschungsdaten stellt hier allerdings eine Herausforderung dar, weil die dafür zur Verfügung stehenden Architekturen dazu tendieren, in multilingualen Trainingskorpora hauptsächlich Sprachunterschiede zu lernen und damit die Einbettungen 'auseinanderzuberechnen'.

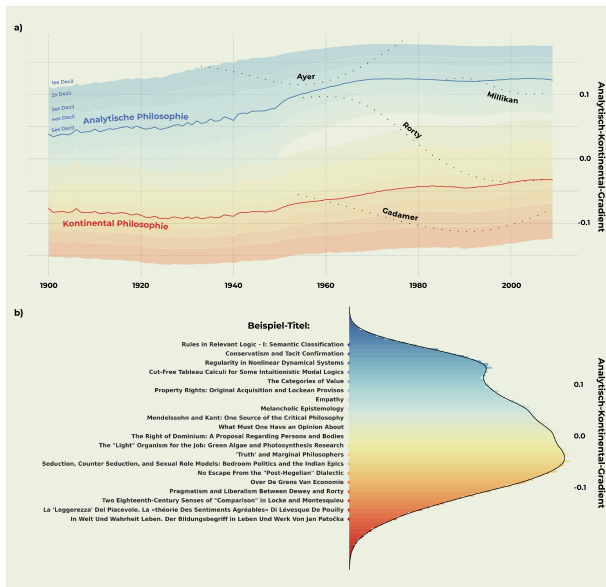
Ein kürzlich vorgeschlagener Lösungsansatz, nämlich die automatische Übersetzung des gesamten Textcorpus (vgl. Malaterre und Lareau 2022; siehe auch Böhm, Alexander u. a. 2022) ist vielversprechend für die Erforschung der thematischen Zusammensetzung von Korpora, aber ungeeignet für Anwendungen in denen die unterschiedlichen Konnotationen von Wörtern in unterschiedlichen Sprachen und Kontexten eine Rolle spielen. Weiterhin ist die automatisierte Überführung von anderen Sprachen in eine einzige Basissprache – Englisch – mit Blick auf den Wunsch nach einer Wertschätzung verschiedener Sprachkulturen nicht ideal.



Um von unserem Sprachmodell zu Antworten auf unsere Fragen nach der Struktur der analytisch-kontinentalen Kluft zu kommen, encodieren wir zuerst alle Texte in unserem Sample mit dem multilingualen Sprachmodell. Die analytisch-Kontinentale Trennung wird in philosophiehistorischen Werken häufig über die Angabe von paradigmatischen ReferenzautorInnen eingeführt (e.g. "Frege", "Russel", "Moore", Quine, Strawson, ... und "Hegel", "Husserl", "Heidegger", "Adorno"...). Wir sammeln solche Listen in der Literatur und wählen die am häufigsten genannten Autoren aus. Dann wählen wir zufällig 2000 Beispielartikel aus, welche Autoren aus der einen, aber nicht der anderen Gruppe zitieren, also tendenziell eher aus der analytischen, oder kontinentalen Ecke kommen. Das ist unsere 'seed'-Stichprobe. Einem von

Durch die Berechnung der Cosinus-Ähnlichkeit aller Artikelvektoren in dem Datensatz zu den kontinentalen und analytischen Artikeln in den ermittelten Artikelpaaren können wir so einen themenunabhängigen einzigen "Analytizität/Kontinentalität"-Score für jeden Artikel entlang der kontinentalen/analytischen Achse ableiten. Die Dichte-Verteilung aller Artikel auf diesem Score ist in Abb. 3.b dargestellt.

Um zu messen, wie sich die kontinentale/analytische Spaltung im Laufe der Zeit vergrößert/verkleinert hat, fitten wir eine Serie von verbundenen gaußschen Mischverteilungsmodellen auf den Datensatz. Unter der Annahme, dass der Datensatz tatsächlich durch die Wirkung zweier Prozesse, welche analytische und kontinentale Philosophie generieren, entstanden ist, geben uns diese Modelle die jeweilige zentrale Tendenz und Spannweite dieser Prozesse an.



Verteilung von Artikeln entlang des analytisch-Kontinentalen-Gradienten. (a) zeigt die Entwicklung des Gradienten im Verlauf der Zeit. Zentrale Tendenzen und Deizile sind einer Serie von verbundenen zwei-Komponenten Gaußschen Mischmodellen entnommen. Im ersten Drittel wäre allerdings ein ein-Komponenten-Modell vorzuziehen. Vier einzelne Philosophen-Karrieren sind anhand der fortschreitenden zentralen Tendenz der Werte ihrer Artikel auf dem Gradienten eingetragen. Man beachte insbesondere die Karriere Rortys von einem ursprünglich analytischen Philosophen, zu einem der wirkungsvollsten Proponenten kontinentaler Autoren im angloamerikanischen Raum. (b) Zufällig ausgewählte Beispiel-Titel entlang des analytisch-kontinentalen Gradienten, nebst Dichte-Verteilung über das gesamte Sample.

Vorläufige Ergebnisse

Die Ergebnisse dieser Modelle sind in Abb. 3.a wiedergegeben. Wir beobachten, dass sich die beiden Verteilungen bis in die späten 1940er Jahre nahezu parallel entwickeln – und in der Tat suggerieren die statistischen Kennzahlen der Modellwahl, dass eine einzige Gauß-Verteilung die Daten in diesem Bereich besser beschreiben würde. Von 1950 bis 1960 beobachten wir hingegen ein scharfes Ausschwenken der analytischen Verteilung, gemeinsam mit einer Verkleinerung der Breite der Verteilung – also eine Konzentration und Konsolidierung analytischer Tendenzen in unserem Korpus, verbunden mit einer asymmetrischen Polarisierung, die bis heute weitestgehend konstant zu bestehen scheint.

Diskussion

Diese Ergebnisse stehen im scharfen Kontrast zu früheren monolingualen Zitations-Studien, die eine Isoliertheit kontinentaler Philosophie identifiziert hatten (Noichl 2021). Die Erweiterung des Datensatzes um mehrere Sprachen und die damit einhergehende methodologische Komplexität stellt also in jedem Fall eine notwendige Grundlage für weitere Untersuchungen dar. Dabei hat der verwendete Datensatz noch nicht alle Möglichkeiten zu multilingualer Erweiterung ausgeschöpft: Die

reichhaltige spanischsprachige OpenAccess-Kultur hat in unserer Untersuchung zum Beispiel noch nicht ausreichend Eingang gefunden. Wenn für die gestellte Fragestellung auch nicht zwingend notwendig, wäre eine Erweiterung über den europäischen Sprachraum hinweg wünschenswert.

Bibliographie

Beckermann, Ansgar. 2003. „Muss die Philosophie noch analytischer werden? (Ist die Analytische Philosophie am Ende?)“. Universität Würzburg.

Bieri, Peter. 2007. „Was bleibt von der analytischen Philosophie?“ Deutsche Zeitschrift für Philosophie 55 (3). <https://doi.org/10.1524/dzph.2007.55.3.333>.

Böhm, Alexander, Reiners-Selbach, Stefan, Baedke, Jan, Fábregas Tejeda, Alejandro, und Nicholson, Daniel J. 2022. „What was Theoretical Biology? A Topic-Modelling Analysis of a Multilingual Corpus of Monographs and Journals, 1914-1945“. DHd2022: Kulturen des digitalen Gedächtnisses, März. <https://doi.org/10.5281/ZENODO.6328143>.

Bornmann, Lutz, und Rüdiger Mutz. 2015. „Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References“. Journal of the Association for Information Science and Technology 66 (11): 2215–22. <https://doi.org/10.1002/asi.23329>.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, und Veselin Stoyanov. 2020. „Unsupervised Cross-lingual Representation Learning at Scale“. arXiv. <http://arxiv.org/abs/1911.02116>.

Critchley, Simon. 1997. „What Is Continental Philosophy?“ International Journal of Philosophical Studies 5 (3): 347–63. <https://doi.org/10.1080/09672559708570862>.

—. 2001. Continental philosophy: A very short introduction. Oxford: Oxford University Press.

Dombrowski, uinn. 2020. „What's a 'Word': Multilingual DH and the English Default“. McGill, Oktober 15. <https://quinndombrowski.com/blog/2020/10/15/whats-word-multilingual-dh-and-english-default/undefined>.

Galina Russell, Isabel. 2014. „Geographical and linguistic diversity in the Digital Humanities“. Literary and Linguistic Computing 29 (3): 307–16. <https://doi.org/10.1093/lilc/fqu005>.

Glock, Hans-Johann. 2008. What is analytic philosophy? Cambridge: Cambridge University Press.

Hobbs, Valerie. 2014. „Accounting for the Great Divide: Features of Clarity in Analytic Philosophy Journal Articles“. Journal of English for Academic Purposes 15 (September): 27–36. <https://doi.org/10.1016/j.jeap.2014.05.001>.

Hoche, Hans-Ulrich. 2009. „Bieri über die Zukunft der analytischen Philosophie – Eine unerlässliche Entgegnung“. Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics 17: 415–44.

Kroß, Matthias. 2012. „Von Umgangskörpern, Vertikalspannungen, Responsivität und Musikphilosophie: Ludwig Wittgenstein im Spiegel neuerer Literatur“. Philosophische Rundschau 59 (3): 197–216.

Malaterre, Christophe, und Francis Lareau. 2022. „The Early Days of Contemporary Philosophy of Science: No-

vel Insights from Machine Translation and Topic-Modeling of Non-Parallel Multilingual Corpora". *Synthese* 200 (3): 242. <https://doi.org/10.1007/s11229-022-03722-x>.

McInnes, Leland, John Healy, und James Melville. 2018. „UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". *arXiv:1802.03426 [cs, stat]*, Februar. <http://arxiv.org/abs/1802.03426>.

Moulines, Ulises. 1989. „¿Hay una filosofía de la ciencia en el último Wittgenstein?" *Theoria: An International Journal for Theory, History and Foundations of Science* 4 (11): 327–42.

Noichl, Maximilian. 2021. „Modeling the Structure of Recent Philosophy". *Synthese* 198 (6): 5089–5100. <https://doi.org/10.1007/s11229-019-02390-8>.

Petrovich, Eugenio, und Valerio Buonomo. 2018. „Reconstructing Late Analytic Philosophy. A Quantitative Approach". *Philosophical Inquiries* 6 (1): 151–82. <https://doi.org/10.4454/philing.v6i1.184>.

Pitman, Thea, und Claire Taylor. 2017. „Where's the ML in DH? And Where's the DH in ML? The Relationship between Modern Languages and Digital Humanities, and an Argument for a Critical DHML". *DHQ: Digital Humanities Quarterly* 11 (1).

Reimers, Nils, und Iryna Gurevych. 2019. „Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". *arXiv*. <http://arxiv.org/abs/1908.10084>.

—. 2020. „Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". *arXiv*. <http://arxiv.org/abs/2004.09813>.

Waller, Isaac, und Ashton Anderson. 2021. „Quantifying Social Organization and Political Polarization in Online Platforms". *Nature* 600 (7888): 264–68. <https://doi.org/10.1038/s41586-021-04167-x>.

Wang, Kexin, Nils Reimers, und Iryna Gurevych. 2021. „TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning". In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 671–88. Punta Cana, Dominican Republic: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-emnlp.59>.

Wang, Kexin, Nandan Thakur, Nils Reimers, und Iryna Gurevych. 2022. „GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval". *arXiv*. <http://arxiv.org/abs/2112.07577>.

Provenienzforschung und ihre Quellenbestände. Aktuelle Nutzungsszenarien zwischen Open Access und Inaccessibility

Hopp, Meike

meike.hopp@tu-berlin.de
TU Berlin, Deutschland

von dem Bussche, Ruth

rbussche@fotostoria.de
Fotostoria

Die Frage nach einer spezifisch deutschen Verantwortung für historisches Unrecht und die Kritik an der Realisierbarkeit und Verhältnismäßigkeit von „Wiedergutmachung“ an Opfern der Verfolgung, Entrechtung und Enteignung durch den nationalsozialistischen Rechtsstaat in der Nachkriegszeit, beschäftigte in den vergangenen Jahrzehnten vor allem Forscher_innen aus den Bereichen der Zeitgeschichte, der Philosophie und der Rechtswissenschaften. In der kunsthistorischen Forschung haben die *Washington Principles on Nazi-Confiscated Art* von 1998 ein Umdenken herbeigeführt. In der Folge hat sich die Provenienzforschung zunächst mit der Koordinierungsstelle für Kulturgutverluste in Magdeburg und der 2008 eingerichteten Arbeitsstelle für Provenienzforschung/-forschung beim Institut für Museumsforschung, Stiftung Preussischer Kulturbesitz in Berlin – seit 2015 zusammengeschlossen zum Deutschen Zentrum für Kulturgutverluste (DZK) in Magdeburg – institutionalisiert. Die Provenienzforschung stellt inzwischen eines der virulentesten geisteswissenschaftlichen Forschungsfelder dar, welches von großem öffentlichem und medialem Interesse begleitet wird und inzwischen an einigen, v.a. kunsthistorischen Instituten zum Ausbildungs- bzw. zum Lehrangebot gehört. Seit 2015 wurden an fünf Universitäten in Bonn, Hamburg, München, Berlin und Lüneburg (Junior-)Professuren dieser Denomination eingerichtet, auch an anderen Universitäten werden Seminare oder spezialisierte Masterstudiengänge angeboten.

Dabei produziert die Provenienzforschung in erheblichem Maße objekt- und personenbezogene Daten: Die seit 2020 vom DZK betriebene Forschungsdatenbank PROVEANA bündelt Entitäten aus 467 von der Stiftung geförderten Projekten, d.h. Daten zu Akteur_innen, Objekten, Archivbeständen und Sekundärquellen, die den Forschenden aber ebenso betroffenen Nachfahren von Verfolgten oder aber Interessierten zur Verfügung zu stehen, darin enthalten auch die Such- und Fundmeldungen

der seit 2000 betriebenen Datenbank LostArt.¹ Parallel entstanden in den vergangenen zwei Jahrzehnten weitere unabhängige Datenbankprojekte, etwa die am Deutschen Historischen Museum in Berlin angesiedelte Datenbank zum Central Collecting Point (CCP) in München, welcher unter der Leitung amerikanischen Militärbehörden in der in der unmittelbaren Nachkriegszeit Wesentliches für die Rückführung des im Nationalsozialismus enteigneten Kulturguts geleistet hat², oder die Datenbank zum Kulturgutraub in Frankreich durch den Einsatzstabs Reichsleiter Rosenberg (ERR).³

Auch innerhalb der vom DZK geförderten Projekte werden Provenienzen dokumentiert, dies in vielfältiger Form, seien es Excelsheets, kleinere selbstgestrickte Datenbanklösungen oder auch zusätzliche Provenienzfelder, die in bestehende Museumsdatenbanken integriert wurden. Diese hausintern erzeugten Daten bleiben aber in der Regel nicht projekt- oder länderübergreifend nutzbar (s. Hopp 2018). Die Frage der Langzeitarchivierung dieser lokal gehaltenen Projektdaten ist vielfach noch gar nicht angegangen worden. Das mag auch mit daran liegen, dass wir es mit einer föderalen Arbeitsstruktur, mangelnden personellen Kontinuitäten (bedingt durch Drittmittelförderungen und befristete Verträge) sowie Heterogenitäten von Datenmodellierungen auf der räumlichen wie auf der zeitlichen Achse zu tun haben, was die Koordination des Datenrückflusses an Datenzentren zeitaufwendig und schwierig macht.

Ein Blick auf die Provenienzforschung in Museen zeigt, wie vielfältig Provenienzen zu einem Objekt sein können. Es handelt sich um Metadaten zu Erwerbsumständen (Zugangsdaten, Rückseitenbefunde, etc.) aber auch um solche, die aus externen Quellen und Überlieferungen in privaten oder öffentlichen Archiven (Beschlagnahme-, Wiedergutmachungsakten, etc.) stammen. Hinzu kommen Daten aus Forschungseinrichtungen, die sich mit den Mechanismen der Verlagerung von Kulturgütern und Akteursnetzwerken sowie dem Kunstmarkt befassen.⁴ Die Herausforderung besteht darin, diese Erkenntnisse und Informationen zu bündeln, wobei die erhobenen Daten immer auch eine Rückbindung an die jeweiligen archivalischen Belege benötigen, die ähnlich vielfältig wie die durch sie beschriebenen Objekte sind.

Wie ist es nun um die Zugänglichkeit zur Archivmaterialien bestellt? Die Archive selbst leisten seit Jahren wichtige Unterstützung für die Provenienzforschung. Die Zielvorgaben von Politik und Forschung waren dabei in der Regel bislang Quellen oder archivalische Bestände schnell und/oder unkompliziert als Digitalisate oder Datenbanken zur Verfügung zu stellen, um die direkte Zugänglichkeit für die Forschung zu gewährleisten, wobei auf spezielle bzw. individuelle Abfragemöglichkeiten weniger Wert gelegt wurde. Aufgrund der schier Menge und Fülle des vorhandenen Materials können die bisherigen Online-Angebote bzw. oben genannte Datenbankprojekte zwar einen wichtigen aber eben auch nur sehr kleinen Ausschnitt der Informationen zum NS-Kunst- und Kulturgutraub wiedergeben. Zudem scheinen bei den digitalen Zugriffsmöglichkeiten auf die für die Provenienzforschung relevanten Archivbestände, große Unterschiede auf.

Das Transparenzgebot der Washingtoner Prinzipien von 1998 steht dabei noch immer im Widerspruch zu den archivrechtlichen Bestimmungen einzelner Einrichtungen, ganz gleich ob auf bundes-, landes- oder kommunaler Ebene. Während die Ende 2016 in Kraft getretene DGSVO mit dem in ihr enthaltenen Erwägungsgrund 158 einen transparenten Umgang mit Daten empfiehlt, die „im Zusammenhang mit dem politischen Verhalten unter ehemaligen totalitären Regimen, Völkermord, Verbrechen gegen die Menschlichkeit, insbesondere dem Holocaust, und Kriegsverbrechen“ stehen,⁵ bleibt sie bezogen auf die praktische Umsetzung der Verarbeitung und Publikation der Daten weiterhin auslegbar, bzw. wird in der Praxis in Archiven und Verwaltungen sehr viel enger gefasst. Dabei fehlt häufig eine saubere Trennung von Daten die DSGVO-relevant sind, von solchen, die in keinem Bezug zu lebenden Personen stehen.⁶

Kleinere, regionale Forschungsverbünde arbeiten inzwischen mit digitalen Repositorien oder Portalen zur Bereitstellung von Quellenmaterial speziell für die Provenienzforschung, doch stehen diese in der Regel nur einem eingeschränkten Kreis an Nutzer_innen zur Verfügung, da neben der rechtlichen Lage auch viele moralisch-ethische Fragen im Umgang mit den erarbeiteten Informationen offen bleiben, so etwa privaten Informationen zu den Geschädigten (Familiendokumente, etc.). So erfüllen die existierenden Lösungsansätze bis heute kaum die Anforderungen an umfassende Transparenz im Umgang mit (Meta-)Daten zur Herkunft der in deutschen Einrichtungen verwahrten kulturellen Objekte, ein Problem, das auch auf die in jüngerer Zeit begründeten Netzwerke zum Umgang mit Objekten aus kolonialen Kontexten – hier sei exemplarisch auf das künftige CCC-Portal der Deutschen Digitalen Bibliothek verwiesen – übertragen werden kann.⁷

Allerdings gibt es aktuell verschiedene Ansätze zur Aufarbeitung von Archivbeständen stärker maschinell gestützten Verfahren einzusetzen, die über Volltexterschließung, Natural Language Processing (NLP) oder der Named Entity Recognition (NER) Dokumente zugänglich machen.⁸ Doch auch hier bleibt fraglich wie z.B. ein bundesweit angelegtes Projekt zur Digitalisierung der Wiedergutmachungsakten die zahlreichen archiv- und personenschutzrechtlichen Beschränkungen und Fristen umgehen wird bzw. wie transparent die Ergebnisse schließlich publiziert werden können, womit alle Digital-Projekte in diesem sensiblen Bereich in rechtlichen Grauzonen agieren. Bedarf es nicht sogar gerade in diesem Fall eines speziellen Schutzes von Daten zu Opfern der NS-Verfolgung? Die Akten, mit denen die Provenienzforschung arbeitet, sind Zeugnisse eines totalitären Unrechtsregimes. Neben den Akten der Finanzämter, die Auskunft über Vermögenswerte und fiskalische Verfolgung geben, erlauben Auflistungen des vor der Emigration in den Expeditionen deponierten Hausrats unmittelbare Einblicke in Hausstand und Familienleben. Schließlich finden sich in Entschädigungsverfahren nicht selten Zeugnisse, die Foltermethoden benennen, die die Geschädigten über sich ergehen lassen mussten, ebenso sowie die davongetragenen medizinischen Spätfolgen. Dürfen wir heute über dieses Wissen, diese Daten frei verfügen?

Doch gerade für Forschende ist der Ansatz z.B. über die maschinelle Erschließung von möglichst vielen Dokumententexten die für die Provenienzforschung wichtige objekt- oder personenbezogenen Daten schnell herauszufiltern natürlich essentiell, da auf Basis der unüberblickbaren Menge der europaweit verstreuten Quellen zum Kunst- und Kulturgutraub der Nationalsozialisten einzelfallbezogene Prüfungen und Sondierungen oft nicht effizient und nachhaltig bearbeitet werden können. Zwar ist es denkbar, maschinelle Verfahren der Texterschließung anzuwenden, aber es braucht zusätzliche Methoden, sensible Inhalte zu filtern und Teile von Beständen – wie bislang auch analog gehandhabt – nur auf begründeten Antrag zur Verfügung zu stellen.

Gleichzeitig steht die digitale Provenienzforschung vor der Herausforderung eine Basis für die Etablierung effizienter Modelle zur standardisierten Erfassung von eindeutigen bzw. uneindeutigen Provenienzen zu schaffen und Datenkompetenzen auszubilden. Denn neben der Recherche von Objektbiografien, Eigentumsübergängen und -verlusten,

hat Provenienzforschung auch das Ziel den heutigen Anspruchsberechtigten und der interessierten Öffentlichkeit gerecht zu werden – ergo zu dokumentieren, aufzuklären, zu informieren und damit auch Daten – weltweit – auffindbar zu machen. Bereits im Sommersemester 2020 sind Studierende am Fachgebiet Digitale Provenienzforschung der TU Berlin in einem Seminar der Frage nachgegangen, ob und inwiefern sich Nachfahren der im NS-Regime rassistisch oder anderweitig Verfolgten über laufende Projekte und bestehende Online-Datenangebote im Bereich der Provenienzforschung in Deutschland informieren können. Im Zentrum stand die Frage, ob die bereits bestehenden Datenbanken/-tools auch für ein nicht spezialisiertes Publikum zugänglich, ob sie auffindbar, transparent und verständlich sind, ob es Sprachhürden gibt und wie viel Vorwissen erforderlich ist, um die in ihnen enthaltenen Informationen richtig zu interpretieren. In Gesprächen mit heute international ansässigen Nachfahren, kristallisierte sich heraus, dass das Online-Angebot nicht nur unübersichtlich ist, sondern wesentlichen Adressat*innen bzw. Interessentengruppen weitestgehend verschlossen bleibt. Die aktuellen Angebote produzieren folglich nicht nur Expertenwissen, sondern auch Ausschlüsse der eigentlich Betroffenen.

Vor diesem Hintergrund haben wir uns seit Herbst 2020 intensiv mit Optimierungsprozessen im Umgang und bei der Erschließung von öffentlich verfügbaren Quellenbeständen zur Provenienzforschung befasst. Unsere Verfahren zielten dabei unter anderem auf den heute im Bundesarchiv Koblenz befindlichen, komplett digitalisierten aber bisher nicht effizient zu durchsuchenden Bestand B323 der in den Nachkriegsjahren agierenden ehem. Treuhandverwaltung von Kulturgut beim Oberfinanzpräsidium in München ab, der wesentliche Original-Quellen zum NS-Kunstraub sowie zu den alliierten Rückführungs-Bemühungen enthält und der in seiner sehr gemischten Zusammensetzung einen optimalen Testbestand lieferte. Ausgangspunkt des Pilotprojekts waren zunächst Metadaten aus der Archiverfassung,⁹ die wir als Graph aufbereitet haben (Bussche, Hopp 2022a und 2022b). Hierbei orientieren wir uns an bisherigen Bemühungen, den Nutzen von Linked-Data- und Se-

mantic-Web-Technologien für archivarische Sammlungen zu untersuchen (Ferris 2014, Gracy 2015, Niu 2016). Um diese zu einer verlässlichen Ressource für die Provenienzforschung zu entwickeln, wurden alle über die Rechercheplattform des Bundesarchivs zur Verfügung stehenden Digitalisate mittels OCR erfasst und die Daten in einer Suchmaschine zur Verfügung gestellt. Bereits auf dieser ersten Grundlage zeichnete sich ab, dass die Erschließung über Volltexte die Anforderungen der Provenienzforschung wesentlich besser abbilden kann, als das bisherige Online-Angebot mit Erschließungen zu Einzelakten über die Bestandsbeschreibungen, denn der Praxis geht es häufig darum die Dokumente mit Erwähnungen bestimmter Personen, Werke oder Institutionen erst einmal aufzufinden.

Die bisher in unserem experimentellen Projekt verwendeten Verfahren machine learning basierter Cloud-dienste, wie etwa *azure OCR* oder *Google NER* ermöglichten es uns auch mit geringen personellen Kapazitäten durchaus umfangreiche Aktenbestände zu verarbeiten, wenngleich die Nutzung der genannten Dienste in öffentlichen Einrichtungen problematisch wäre. Wesentliche Hindernisse für eine Zugänglichmachung von Provenienzen beginnen also nicht erst bei der Frage der Online-Stellung, sondern schon sehr viel früher bei der Verarbeitung der Daten. Die von verschiedenen Institutionen unterschiedlich strikt gehandhabten Regeln hierzu blieben ebenso uneinheitlich wie die Vorgaben zum Datenschutz selbst, wobei Empfehlungen für „rechtskonforme“ Ersatzprodukte fehlen.

Die Möglichkeiten der maschinellen Verarbeitung der bisher erzeugten Daten zum Bestand B323 gehen allerdings weiter (Stork 2021, Moss et al. 2018, Krenn 2019). Zum gegenwärtigen Zeitpunkt geht es uns darum, die Qualität der bislang eingesetzten Verfahren zu evaluieren und an einer künftigen verbesserten Verarbeitungspipeline zu arbeiten.

Welche qualitativen Vorteile bringt eine Vorverarbeitung der Scans (z.B. Ränder entfernen)?

Wo müssen Elemente eines Digitalisats segmentiert werden um die Auslesung der Daten bzw. Texte zu optimieren? Das betrifft vor allem kleine Notizen und Belege oder ausgeschnittene Bilder aus Mikrofilmen, die auf DinA4 Seiten montiert wurden.

Wo kann hingegen Layouterkennung eingesetzt werden (z.B. bei Listen, Karteikarten, Korrespondenzen)?

Welche Möglichkeiten haben wir Dokumente auch inhaltlich erkennen zu lassen, um z.B. eine Filterung nach Rechnungen oder Transportlisten vorzunehmen? Welche Algorithmen stehen für buchhalterische Dokumente wie Quittungen oder Rechnungen zur Verfügung, um deren Inhalte strukturiert erkennen zu lassen?

Wie können wir Texte über Entitäten erschließen?

Für die Erschließung von Dokumentenbeständen über Entitäten gibt es bereits Vorbilder: so wurden Dokumente aus der Zeit der deutschen Besatzung in den Niederlanden beispielsweise im Projekt Oorlogsbronnen aufbereitet (Borggräfe et al. 2020).¹⁰ Ein weiteres bemerkenswertes Beispiel aus dem Bereich der Provenienzforschung ist der Archivführer zur deutschen Kolonialgeschichte, der archivische Sammlungen katalogisiert und mit Wikidata-Elementen verknüpft und somit neben dem niedrigschwelligen Einstieg in die Recherchen auch die weitere Bearbeitung der Entitäten und den Einbezug von Exper-

t_innenwissen erlaubt, um die Qualität der Erkennung und Disambiguierung zu verbessern (Jung 2019).¹¹ Die beiden Beispiele liefern damit Entitäten, über die Texte mit der entsprechenden Qualität maschinell erschlossen werden können.

Das bedeutet im Folgeschluss, dass es immer wichtiger wird, dass strukturierte Daten auch als offene Daten zur Verfügung stehen. Plattformen wie PROVEANA haben die dazu notwendige Struktur und Qualität, bieten aber über Einzelrecherchen hinaus keinen Zugang zu den Rohdaten oder API an.

Die Anwendung von Semantic-Web-Technologien in der Provenienzforschung benötigt unserer Meinung nach: einen offenen Diskurs zum standardisierten und FAIRen Umgang mit den Digitalisaten und Forschungsdaten an allen Provenienzforschung betreibenden Einrichtungen, Verfahren zur effizienteren Aufbereitung von Dokumenten, um die nötige Qualität der Dokumentenverarbeitung (OCR, Layouterkennung usw.) sicherstellen und Methoden um Unterscheidungen zwischen den für die Forschung offenzulegenden Daten von den sensiblen Informationen vorzunehmen.

Fußnoten

1. Vgl. <https://www.proveana.de> (aufgerufen am 12.12.2022).
2. Vgl. https://www.dhm.de/datenbank/ccp/dhm_ccp.php?seite=9 (aufgerufen am 12.12.2022).
3. Vgl. <https://www.errproject.org> (aufgerufen am 12.12.2022).
4. Forschungsschwerpunkte zum Kunstmarkt gibt unter anderem am Zentralinstitut für Kunstgeschichte (ZI) München, hier sogar mit ausgewiesenem Schwerpunkt im Bereich der Provenienzforschung: <https://www.zik-g.eu/forschung/provenienzforschung-werte-von-kulturguetern>, am Zentralarchiv für deutsche und internationale Kunstmarktforschung an der Uni Köln: <https://khi.phil-fak.uni-koeln.de/fachgebiete/kunstmarkt/forschungsschwerpunkte> oder am Forum Kunst und Markt der TU Berlin https://www.kuk.tu-berlin.de/menue/forum_kunst_und_markt/ (alle aufgerufen am 12.12.2022).
5. Siehe: <https://dsgvo-gesetz.de/erwaegungsgruende/nr-158/> (aufgerufen am 12.12.2022).
6. Der rechtliche Rahmen findet sich skizziert in Schlagk 2019; Eisenberger et al. 2018.
7. Vgl. <https://ccc.deutsche-digitale-bibliothek.de/> (aufgerufen am 12.12.2022).
8. Wesentliche Projekte sind hier die Akten des Oberfinanzpräsidenten Brandenburg, s. <https://blha.brandenburg.de/index.php/projekte/ofp-projekt/> und das Projekt zu den Wiedergutmachungsakten, s. <https://www.bundesarchiv.de/DE/Content/Pressemitteilungen/publikation-portal-wiedergutmachung.html> und <https://www.fiz-karlsruhe.de/de/forschung/wiedergutmachung> (alle aufgerufen am 12.12.22).
9. Das Bundesarchiv stellt die Daten aus der Bestandserfassung als offene Daten zur Verfügung, vgl. <https://www.bundesarchiv.de/DE/Content/Artikel/Ueber-uns/Aus-unserer-Arbeit/open-data.html>.

10. Vgl. <https://www.oorlogsbronnen.nl> (aufgerufen am 12.12.2022).

11. Vgl. <https://archivfuehrer-kolonialzeit.de> (aufgerufen am 12.12.2022).

Bibliographie

Borggräfe, Henning, et al., editors. "Linking and Enriching Archival Collections in the Digital Age: The Dutch War Collections Network." Tracing and Documenting Nazi Victims Past and Present, De Gruyter Oldenbourg, 2020, pp. 315–38, <https://www.degruyter.com/document/doi/10.1515/9783110665376-018/html>.

"Digitale Provenienzforschung" Provenienz & Forschung, edited by Deutsches Zentrum Kulturgutverluste, vol. 1, 2020

"EAC-CPF." Encoded Archival Context for Corporate Bodies, Persons, and Families, <https://eac.staatsbibliothek-berlin.de>.

Eisenberg, Iris et al. "Zeitgeschichtliche Forschung im Spannungsfeld von Archiv- Datenschutz- und Urheberrecht." Wien, 2018.

Ferris, Virginia L. Beyond "Showing What We Have": Exploring Linked Data for Archival Description. University of North Carolina at Chapel Hill, 2014, <https://doi.org/10.17615/6n9k-q582>.

Gracy, Karen F. "Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges." Archival Science, vol. 15, 2015, pp. 239–94, <https://doi.org/10/gdck6>.

Greenberg, Jane. "The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives." The American Archivist, vol. 61, 1998, pp. 400–25, <https://meridian.allenpress.com/american-archivist/article/61/2/400/23942/The-Applicability-of-Natural-Language-Processing>.

Hopp, Meike. "Provenienzforschung und digitale Forschungsinfrastrukturen in Deutschland: Tendenzen, Devisen, Bedürfnisse" ... (k)ein Ende in Sicht. 20 Jahre Kunstrückgabegesetz in Österreich, edited by Eva Blimlinger and Heinz Schödl, Wien, 2018, pp. 37–61.

Jung, Uwe. "Archivführer zur deutschen Kolonialgeschichte" Archivar, no. 4, 2019, pp. 325–327.

Krenn, Brigitte. "Methoden der künstlichen Intelligenz und ihre Anwendung in der Erschließung von Textinhalten." Die Zukunft der Vergangenheit in der Gegenwart, Archive als Leuchtturm im Informationszeitalter, edited by Elisabeth Schöggel-Ernst, Thomas Stockinger and Jakob Wührer, Wien, 2019, pp. 169–184.

Marciano, R., et al. "Archival Records and Training in the Age of Big Data." Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education, vol. 44B, Emerald Publishing Limited, Bingley, 2018, pp. 179–99, <https://doi.org/10/gndpbh>.

Moss, Michael, et al. "The Reconfiguration of the Archive as Data to Be Mined." Archivaria, vol. 86, 2018, pp. 118–51, <https://archivaria.ca/index.php/archivaria/article/view/13646>.

Niu, Jinfang. "Linked Data for Archives" ARCHIVARIA, The Journal of the Association of Canadian Archivists, vol. 82, 2016, pp. 83–110, <https://archivaria.ca/index.php/archivaria/article/view/13582>.

Nadeau, David, and Satoshi Sekine. "Linked Data for Archives." *ARCHIVARIA*, The Journal of the Association of Canadian Archivists, vol. 82, 2016, pp. 83–0, <https://archivaria.ca/index.php/archivaria/article/view/13582>.

Schlagk, Patricia. „Die datenschutzrechtlichen Privilegien von im öffentlichen Interesse liegenden Archivzwecken.“ Bachelorarbeit, Fachhochschule Potsdam 2019, urn:nbn:de:kobv:525-24311.

Stork, L. Knowledge Extraction from Archives of Natural History Collections. Leiden, July 2021, https://scholarlypublications.universiteitleiden.nl/handle/1887/3192382?solr_nav%5Bid%5D=dc5e9a7ed703338e5b78&solr_nav%5Bpage%5D=0&solr_nav%5Boffset%5D=0

von dem Bussche, Ruth, and Meike Hopp. "Der ‚Bestand B323‘ als Knowledgegraph für die Provenienzforschung. Methodische Überlegungen zur Verarbeitung von Archivdaten als Linked Open Data." *Archivar*, no. 1, 2022, 60–63, https://www.archive.nrw.de/sites/default/files/media/files/Archivar_2022-1_Internet-NEU-28032022_Mod.pdf.

von dem Bussche, Ruth, and Meike Hopp. "The Archive as a Graph – Provenance Research on Bundesarchiv B323" *Graphs and Networks in the Humanities. Technologies, Models, Analyses, and Visualizations*, 6th International Conference, 2022, https://graphentechnologien.hypotheses.org/files/2022/01/The_Archive_as_a_Graph_Provenance_Research_on_etc-Hopp.pdf

#PublicDH oder doch nur #WissKomm?

Seltmann, Melanie Elisabeth-H.

melanie.seltmann@tu-darmstadt.de

Universitäts- und Landesbibliothek Darmstadt

Einleitung und Methode

Digitale Wissenschaftskommunikation und die Möglichkeiten der sozialen Medien werfen in den letzten Jahren vermehrt wieder Fragen der klassischen Forschung zu Wissenschaftskommunikation auf (vgl. Franzen 2019, 616). Der vorliegende Beitrag schließt hier für das Konzept der *Public Humanities* an. Damit ist auch klar, dass nicht jegliche Form der Wissenschaftskommunikation betrachtet werden kann. Im Laufe des Beitrages wird herausgearbeitet, dass es sich bei *Public Humanities* nicht um eine traditionelle Form der Wissenschaftskommunikation handelt, wie etwa die Veröffentlichung eines (populär-)wissenschaftlichen Beitrages oder das Halten eines (populär-)wissenschaftlichen Vortrages, sondern um eine moderne, digitale, in der durch moderne Medienformen die *One-to-Many-Communication* zu einer *Many-to-Many-Communication* wird und der Diskurs damit interaktiv und rekursiv wird (vgl. Bucher 2019, 64).

Damit einhergehend wird die Wissenschaftskommunikation von einem öffentlichen Bewusstsein für Wissenschaft zu einem bürgerlichen Engagement verschoben, die Kommunikation wird zum Dialog (vgl. Bucchi and Trench 2014, 4).

Nach einer Definition der Konzepte *Public Humanities* sowie *Public Digital Humanities* folgt eine Betrachtung verschiedener Arten und Dimensionen von Wissenschaftskommunikation. Hierbei fokussiere ich mich in diesem Beitrag auf den europäischen Raum. In den USA findet das Thema *Public Humanities* schon länger Betrachtung als in Europa, ebenso wie die Forschung zu Wissenschaftskommunikation allgemein ausgeprägter ist (vgl. Schäfer et al. 2019, 79), dieser Aspekt soll hier jedoch ausgeklammert werden und für die weitere Forschung aufgehoben werden. In einem dritten Schritt soll das Konzept der *Public Humanities* in die Wissenschaftskommunikationsmatrix nach Frick et al. (2021) eingebettet werden und damit herausgearbeitet werden, wer in den *Public Humanities* agiert und zu welcher Form von Wissenschaftskommunikation sie gehören. An diese theoretischen Betrachtungen schließt sich ein Kapitel zum Mehrwert von *Public Humanities* für verschiedene Zielgruppen und in Hinblick auf verschiedene Aktivitäten an. Daraufhin werden Handlungsempfehlungen für möglichst erfolgreiche *Public Humanities* gegeben. Der Beitrag schließt mit einem Resümee.

Definitionsversuche #PublicDH

Zu Beginn möchte ich den Terminus *#PublicDH* versuchen zu definieren. Wie bei so ziemlich jedem Terminus gibt es auch hier reichlich Möglichkeiten der Definition. Heinisch (2021) gibt einen umfassenden Überblick über den Unterschied zwischen *Public Humanities*, *Public Digital Humanities* und *Citizen Humanities*. Aus ihren Ausführungen zu den *Public Humanities* kann man zusammenfassen, dass ein wesentlicher Bestandteil ist, die Geisteswissenschaften aus ihrem universitären Elfenbeinturm herauszuheben und zum einen der Öffentlichkeit näher zu bringen, sie aber gleichermaßen auch in die Forschung einzubinden. Damit einher geht das Schaffen von neuen Formaten der Wissenschaftskommunikation sowie die Bidirektionalität selbiger. Es ist nicht mehr allein der:die Wissenschaftler:in, der:die kulturelle Artefakte untersucht, sondern sie werden in Gemeinsamkeit mit Bürger:innen reflektiert und interpretiert. Jacobson (2020) führt hierfür beispielsweise die Form des Storytellings an (vgl. Jacobson 2020, 165–169).

Anfangs gingen die *Public Humanities* beinahe missionarisch vor, wenn sie die Bevölkerung belehrten und ihre eigenen Aktivitäten rechtfertigten (vgl. Gibbs 2016, 1). Heutzutage steht jedoch viel mehr der Wunsch, durch Zusammenarbeit etwas zu verändern und in der Gesellschaft zu bewirken (vgl. Miller et al. 2017). Zudem werden Disziplinengrenzen überwunden und interdisziplinär etwas Größeres vermittelt und erarbeitet. Es handelt sich also um bidirektionale und disziplinübergreifende Wissenschaftskommunikation.

Im Unterschied dazu erweitern die *Public Digital Humanities* (oder kurz *#PublicDH*) die *Public Humanities* durch eine digitale Komponente: die vermittelte Wissenschaft wird mit Hilfe von digitalen Methoden durchge-

führt. Dies betrifft sowohl die originär von Wissenschaftler:innen durchgeführte Forschung als auch dialogisch mit Bürger:innen erzeugte Forschung. Zudem kann auch die gemeinsame (Weiter-)Entwicklung dieser digitalen Methoden Teil der *#PublicDH* sein (vgl. Heinisch 2021). Damit sind die *#PublicDH* die Art der *Public Humanities*, die eine spezielle Form der *#WissKomm* für die Disziplin(en) der *Digital Humanities* aufbereiten. An der Grenze der *#PublicDH* liegen Wissenschaftskommunikationsformate, in denen nicht-digitale Wissenschaft durch digitale Formate für die Vermittlung aufbereitet wird. Genau genommen handelt es sich bei dieser Form der *#WissKomm* eher um *Digital Public Humanities* denn um *#PublicDH*.

Arten und Dimensionen der Wissenschaftskommunikation

Wissenschaftskommunikation ist jedoch nicht gleich Wissenschaftskommunikation. Dafür möchte ich im Folgenden genauer betrachten, welche Formen von Wissenschaftskommunikation oder kurz *#WissKomm* es gibt.

Kommunikationsmatrix

Meistens wird hier vor allem zwischen interner und externer *#WissKomm* unterschieden. Die interne *#WissKomm*, auch *Scholarly Communication* genannt, richtet sich an die wissenschaftliche Community, häufig der eigenen Disziplin. Die externe *#WissKomm* adressiert dahingegen Nicht-Spezialist:innen (vgl. Könniker 2017, 454). Die Definition greift hier jedoch zu kurz. Für eine Spezifizierung verschiedener Wissenschaftskommunikationsformen sollte nicht nur der:die Adressat:in betrachtet werden, sondern auch der:die Sender:in. Folglich spannt sich hier eine Matrix auf (vgl. Figure 1).

Adressat:in Sender:in		
	Wissenschaft	Öffentlichkeit
Wissenschaft	Science-to-Science	Science-to-Public
Öffentlichkeit	Public-to-Science	Public-to-Public

Figure 1: Matrix Wissenschaftskommunikation (Verkürzung von Frick et al. 2021)

Das orange hinterlegte Feld (*Science-to-Science*) repräsentiert die interne, die blau hinterlegten Felder (*Science-to-Public*, *Public-to-Science*, *Public-to-Public*) die externe *#WissKomm*. Externe *#WissKomm* kann folglich viel mehr als Kommunikation, die nicht intern innerhalb einer Wissenschaftsdisziplin oder allgemein innerhalb der Wissenschaft erfolgt, verstanden werden. Frick et al. (2021) bestimmen externe *#WissKomm* genauer in drei Bereiche und setzen sie in den Kontext von traditionellen Veröffentlichungsformen und *Citizen Science*. *Science-to-Public-Communication* ist die „normale“ externe Wissenschaftskommunikation oder auch *Science Communication*, obgleich sie nicht nur in den STEM-Fächern (Naturwissenschaften, Technik, Ingenieurwissenschaften, Mathematik) vorhanden ist, sondern auch

in den SSH (Sozialwissenschaften, Geisteswissenschaften). Die Bereiche *Public-to-Science* und *Public-to-Public* betreffen den Bereich *Citizen Science*. In der *Public-to-Science-Communication* liefern Bürger:innen Erkenntnisse an die Wissenschaft. Zudem gehört in diesen Bereich auch das Einbringen von Forschungsfragen durch Bürger:innen. Hierdurch wird für die Allgemeinheit relevante Forschung besonders sichtbar. Den Bereich der *Public-to-Public-Communication* bezeichnen Frick et al. (2021) als *Citizen-Science-Communication*. Hier wird dem Rechnung getragen, dass Bürger:innen sich in *Citizen-Science*-Projekten nicht nur der Forschung beteiligen, sondern selbst als Multiplikator:innen fungieren und von ihrer Forschung und ihren Ergebnissen, seien sie positiv oder negativ, erzählen. Damit umfassen Frick et al. (2021) mit ihrem Begriff von *Citizen-Science-Communication* nur einen Teil der Kommunikationsformen, die anderswo inhärent wären. Häufig werden auch die *Science-to-Public-Communication* sowie die *Public-to-Science-Communication* mit in den Bereich *Citizen Science* gezählt.

Schumacher (2021) spricht sogar von mindestens fünf kommunikativen Schnittstellen in den digitalen Geisteswissenschaften: *Humanities-to-Public* (Wissenschafts-PR), *Humanities-to-Media* (Wissenschaftsjournalismus), *Public-to-Humanities* (*Citizen Science*), (*Digital*)-*Humanities-to-(Digital)-Humanities* (Kommunikation in der eigenen Community) sowie *Digital-Humanities-to-non-Digital-Humanities*. Wobei die Zuordnung zu Frick et al. (2021) nicht eins zu eins gelingt. Schumacher (2021) lässt die *Public-to-Public-Communication* außen vor, ergänzt dafür jedoch die *Humanities-(oder Science)-to-Media-Communication* sowie die *Digital-Humanities-to-non-Digital-Humanities-Communication*, also die disziplinenübergreifende *#WissKomm* bzw. die *#WissKomm* zwischen Wissenschaftler:innen verschiedener Methoden. Möchte man Schumachers kommunikative Schnittstellen in die Wissenschaftskommunikationsmatrix nach Frick et al. (2021) einordnen, so wäre die *Humanities-to-Media-Communication* eine Sonderform der *Science-to-Public-Communication*. Bucher (2019) führt aus, dass es bei der Wissenschaftskommunikation, die über Wege des Journalismus vermittelt wird, um die traditionelle Form der *Science-to-Public-Communication* handelt, die digitalen Medien jedoch Möglichkeiten für eine direkte *Science-to-Public-Communication* schaffen (vgl. Bucher 2019, 64). Bei der *Digital-Humanities-to-non-Digital-Humanities-Communication* handelt es sich um eine Sonderform der *Science-to-Science-Communication*.

Dimensionen

Alle Bereiche der *#WissKomm* haben ihre ganz eigenen Herausforderungen und je nach Form müssen fünf der sechs Dimensionen der Wissenschaftskommunikation (Frick et al. 2021) anders bedacht werden. Die erste Dimension (*Inhalt*) ist in allen Formen der *#WissKomm* gleich (für einen spezifischen Kommunikationsanlass). Im nächsten Schritt entscheidet man, welche *Zielgruppe* man ansprechen möchte (Dimension 2). Hier unterscheidet sich die folgende *#WissKomm* in die verschiedenen Formen. Die weiteren Dimensionen passen

sich an die getroffene Entscheidung an und stehen auch untereinander in enger Beziehung. Ändert sich eine der Dimensionen, so müssen auch die anderen angepasst werden. Mit der dritten Dimension wird der verwendete *Stil* betrachtet, in der vierten Dimension wird sich für ein adäquates *Format* entschieden. Wichtig ist auch zu betrachten, welche *Motivation* man bei der #WissKomm hat, also warum man etwas vermitteln möchte (Dimension 5) und schließlich sollte man sich vor dem Kommunikationsakt klar werden, welche *Rolle* man selbst spielen möchte (Dimension 6).

Auch das NaWik gibt mit ihrem NaWik-Pfeil (vgl. Köneker 2017, 468) Hilfestellung für die Planung von Wissenschaftskommunikation und rät dazu, fünf Dimension im Vorhinein zu beachten. Vier davon decken sich mit den eben beschriebenen sechs Dimensionen. Als erstes ist die Dimension *Thema* zu beachten, die der Dimension *Inhalt* entspricht. Daraufhin soll über den *Stil* (Dimension 2) nachgedacht werden, der bei Frick et al. (2021) erst im dritten Schritt betrachtet wird. Als dritte Dimension wird hier nun das *Medium* reflektiert, was dem *Format* entspricht. Erst im vierten Schritt wird bei NaWik über die *Zielgruppe* nachgedacht. Schließlich führt NaWik als fünfte Dimension das *Ziel* selbst an, was ähnlich wie die *Motivation* bei Frick et al. (2021) zu verorten ist. Der Unterschied ist also abgesehen von der Reihenfolge einiger Dimensionen, dass Frick et al. (2021) die Frage nach der Rolle, die die kommunizierende Person einnimmt, aufwirft.

Einbettung des Konzepts #PublicDH in die Kommunikationsmatrix

Doch wie lässt sich das zuvor definierte Konzept #PublicDH in die eben aufgestellte Kommunikationsmatrix einbetten? Da ich als einen zentralen Punkt die Einbindung der Öffentlichkeit definiert habe, gehören die Bereiche *Science-to-Public* und *Public-to-Science* zu den #PublicDH. Sowohl die verständliche Vermittlung der Forschung als auch der Dialog darüber sind wesentlicher Bestandteil der #PublicDH. Auch ist es relativ selbstverständlich, dass der Bereich *Science-to-Science* nicht eingeschlossen ist in die #PublicDH. Problematischer ist die Bestimmung, ob der Bereich *Public-to-Public-Communication* Teil der #PublicDH ist oder nicht. Zwar könnte man argumentieren, dass ebenso wie bei *Citizen-Science*-Projekten auch in den #PublicDH die Multiplikator:innenrolle der Bürger:innen wichtiger Bestandteil ist, jedoch plädiere ich dafür, den Bereich aus den #PublicDH auszuklammern. Nach der oben aufgestellten Definition sind #PublicDH die Kommunikation von Wissenschaft und Öffentlichkeit über Forschungsinhalte und -methoden der digitalen Geisteswissenschaften. Die Wissenschaft ist also den #PublicDH inhärent. Folglich kann die *Public-to-Public-Communication* nicht Teil der #PublicDH (zumindest im engen Sinne) sein (vgl. Figure 2). Hierin liegt auch die Unterscheidung zur externen #WissKomm. Die Wissenschaft muss ein Agens sein und bleiben, das ist bei der #WissKomm nicht der Fall.

Adressat:in	Sender:in	
	Wissenschaft	Öffentlichkeit
Wissenschaft	Science-to-Science interne #WissKomm	Science-to-Public externe #WissKomm #PublicDH
Öffentlichkeit	Public-to-Science externe #WissKomm #PublicDH	Public-to-Public externe #WissKomm

Figure 2: Matrix #WissKomm und #PublicDH (Erweiterung von Frick et al. 2021)

Mehrwert #PublicDH

Nachdem ich definiert habe, was #PublicDH sind und wie sie sich in die Wissenschaftskommunikationsmatrix einbetten lassen, betrachte ich in diesem Abschnitt, welchen Mehrwert #PublicDH eigentlich bringen. Betrachtet werden dafür unterschiedliche Gruppen: zum einen der Mehrwert für die Wissenschaftler:innen, zum anderen für die Öffentlichkeit.

Hierzu sei erst einmal allgemein die Frage nach dem Mehrwert von #WissKomm gestellt. Fröhlich (1994) führen drei potentielle Mehrwerte auf, die jedoch vor allem auf interne #WissKomm zutreffen: Zum einen fördert #WissKomm die Motivation und lässt neue Ideen aufkommen, Mehrfacherfindungen werden vermieden, aber dafür Synergieeffekte erzeugt und schließlich führt die #WissKomm zu einer Form der Qualitätskontrolle und damit auch zur Selektion von Forschung (vgl. Fröhlich 1994, 7). Zudem bezeichnet Fröhlich als Mehrwert von #WissKomm den „errungene[n] Kredit auf wissenschaftliche Glaubwürdigkeit“ (Fröhlich 1994, 8) von Individuen, Gruppen und Institutionen im Bourdieuschen Sinne.

Auch für die #PublicDH ist die Qualitätskontrolle sicherlich ein großer Mehrwert. Hinzu kommt die Anwendbarkeit der eigenen Forschung(ergebnisse) auch in alltäglichen Situationen. Durch die Kommunikation über Forschung, Methoden und Ergebnisse wird Wissenschaft nicht nur aus dem Elfenbeinturm gehoben und erfahrbar gemacht, sondern es kommt zu einer Resonanz, die die Wissenschaft wiederum bestärken und zu relevanter Forschung von Bedeutung lenken kann. Dadurch kann wiederum Motivation geschaffen und neue Ideen und Ansätze gefunden und ausprobiert werden.

So #PublicDH innerhalb der sozialen Medien betrieben werden, können zudem Mehrwerte aus dem speziellen Medium erzeugt werden. "Für Wissenschaftler stehen hier die Vernetzungsmöglichkeiten im Vordergrund. Wissenschaftsjournalisten nutzen sie primär als Rechercheinstrument, Laien zur Informationsbeschaffung." (Weitze und Heckl 2016, 191) Diese Verknüpfung verschiedener Benutzungsgründe eines Mediums können in Verbindung der kommunikativen Interaktionsmöglichkeiten hervorragend für #WissKomm und auch #PublicDH verwendet werden. Für den speziellen Fall von Twitter führen Geyer und Gottschling (2019) an, dass die Wissenschaftskommunikation hier „den fachinternen Dialog befördern und zugleich Erkenntnisse des Fachs der Öffentlichkeit zugänglich machen, über Methoden oder über aktuelle fachwissenschaftliche Diskus-

sionen und Projekte informieren und auf vielfältige Weise den unmittelbaren Austausch mit zumindest Teilen der Gesellschaft erleichtern [kann].“ (Geier und Gottschling, 2019, 284) An wenig anderen Orten als in den sozialen Medien kommen die verschiedenen Akteur:innen zusammen, um ihre jeweiligen Bedürfnisse der Informationsbeschaffung und des Informationsaustauschs zu stillen.

Zusätzlich kann als Mehrwert gesehen werden, dass die eigene Person bekannter wird und dass Wissenschaftskommunikation Spaß macht (vgl. Ziegler/Fischer 2020, 7).

All diese Mehrwerte beziehen sich jedoch auf die Wissenschaftler:innen selbst. Es stellt sich jedoch die Frage, ob es *#PublicDH* auch einen Mehrwert für die Öffentlichkeit erzeugt. Der gesellschaftliche Mehrwert von Wissenschaftskommunikation findet bisher wenig Beachtung in der Forschung (vgl. Siegel/Dunkel/ Terstriep (2021). Eine der raren Untersuchungen, die auch den gesellschaftlichen Mehrwert adressiert, ist die Umfrage von Ziegler/Fischer (2020) zu Zielen von Wissenschaftskommunikation. Als gesellschaftliche Mehrwerte werden eine Steigerung der Demokratiefähigkeit sowie die Stärkung der Wissensgesellschaft genannt (vgl. Ziegler/Fischer 2020, 7). Ziegler/Fischer (2020) führen zudem an, dass ihre Untersuchungen ergeben haben, „dass Politik und Öffentlichkeit in ihren Entscheidungen stärker auf wissenschaftliche Erkenntnisse zurückgreifen“ (Ziegler/Fischer 2020, 15). Sie zeigen, dass „Forschung klar als Triebkraft gesellschaftlicher Weiterentwicklung und Innovation gesehen [wird]“ (Ziegler/Fischer 2020, 16). Des weiteren wird eine Ausbildung der *scientific literacy* sowie die eigene Mündigkeit durch die Möglichkeit auf wissenschaftliches Wissen zuzugreifen als Mehrwert herausgearbeitet. Nichtsdestotrotz ist hier noch ein weites Feld für weitere Untersuchungen.

Konsequenzen und Handlungsempfehlungen

Was für Konsequenzen können nun aus den Mehrwerten der *#PublicDH* gezogen werden und welche Handlungsempfehlungen für digitale Geisteswissenschaftler:innen ergeben sich daraus? Zum einen sollte erkannt werden, welches Potential in den *#PublicDH* liegt, und überlegt werden, wie dieses Potential für die eigene Forschung genutzt werden kann.

Der Dialog mit anderen Wissenschaftler:innen, aber auch mit der Öffentlichkeit kann das Weiterkommen in der eigenen Arbeit bestärken. Durch Diskussionen und Fragen können sich über die relevanten Säulen ihrer Forschung bewusster werden und den Kern ihrer Forschung genau durchdenken. Insofern kann die Forschung von *#PublicDH* nur profitieren.

Wichtig ist jedoch, die verschiedenen Dimensionen (s. Kapitel 2.2) zu beachten und entsprechend anzuwenden, damit *#WissKomm* gelingen kann. Es ist keine Schande, sich für einzelne (oder auch alle) Dimensionen Hilfe zu holen und diesbezüglich weiterzubilden. Unterstützung kann hier beispielsweise der Referenzrahmen Wissenschaftskommunikation (Frick und Seltmann, in Vorbereitung) bieten. Es lohnt auch über den wissen-

schaftlichen Rahmen hinauszublicken und sich in den Bereichen Marketing und Journalismus Hilfestellungen zu holen. Zwar kann es auch mit gut ausgearbeiteten Dimensionen dazu kommen, dass der Dialog nicht funktioniert, aber dies sollte nur selten der Fall sein, insofern ein echtes Interesse des Austauschs besteht. Wichtig ist zudem zu beachten, dass *#WissKomm* im Allgemeinen und *#PublicDH* im Besonderen Zeit benötigen. Allerdings ist diese Zeit sinnvoll investiert.

Resümee

Der Titel des Beitrags stellt die Frage, ob *#PublicDH* mehr ist als „nur“ *#WissKomm*. Hierfür habe ich zuerst definiert, was *#PublicDH* ist (Kapitel 2). Daraufhin habe ich die Kommunikationsmatrix nach Frick et al. (2021) beschrieben und mit den kommunikativen Schnittstellen von Schumacher (2021) verglichen (Kapitel 3). Einen wesentlichen Beitrag zur Beantwortung der aufgestellten Frage liefert die Einbettung des Konzepts *#PublicDH* in die Fricksche Kommunikationsmatrix (Kapitel 4). Hierdurch wurde klar, wie der Zusammenhang zwischen *#WissKomm* und *#PublicDH* ist und dass beide nicht identisch sind. Das Konzept der *#PublicDH* ist präziser als nur *#WissKomm* für digitale Geisteswissenschaften, auch als externe *#WissKomm* für digitale Geisteswissenschaften. Denn es beinhaltet als wesentlichen Bestandteil den Diskurs zwischen Wissenschaft und Öffentlichkeit. Damit sind beide Seiten Agens und Patiens im Diskurs gleichermaßen. Nach dieser Feststellung habe ich den Mehrwert von *#PublicDH* beschrieben (Kapitel 5) und schließlich Handlungsempfehlungen daraus abgeleitet (Kapitel 6). Mit diesem Beitrag ist die Forschung zu den *#PublicDH* natürlich keinesfalls abgeschlossen. Zu betrachten ist einerseits, inwiefern die *#PublicDH* überhaupt anders funktionieren als Wissenschaftskommunikation anderer Disziplinen und was Abgrenzungskriterien sind. Des weiteren wäre interessant zu untersuchen, inwiefern in den *#PublicDH* spezifisches (methodologisches) Wissen der DH Anwendung findet. Viele weitere Fragestellungen können sich anschließen.

Bibliographie

- Bucchi, Massimo und Brian Trench. 2014. Science communication research: Themes and challenges. In: *Routledge handbook of public communication of science and technology*, hg. von Massimo Bucchi und Brian Trench, 1–14. New York: Routledge 10.4324/9780203483794.
- Bucher, Hans-Jürgen. 2019. The contribution of media studies to the understanding of science communication. In: *Science Communication*, hg. von Annette Leßmöllmann, Marcelo Dascal, und Thomas Gloning, 51–76. Berlin, Boston: De Gruyter Mouton 10.1515/9783110255522-003.
- Franzen, Martina. 2019. Reconfigurations of science communication research in the digital age. In: *Science Communication*, hg. von Annette Leßmöllmann, Marcelo Dascal und Thomas Gloning, 603–624. Berlin, Boston: De Gruyter Mouton 10.1515/9783110255522-028.

Frick, Claudia, Lambert Heller, Sabrina Ramünke und Florian Strauß. 2021. „Bibliotheken als Dienstleisterinnen und Labore der Wissenschaftskommunikation“. *Zenodo* 10.5281/zenodo.5752401.

Frick, Claudia und Melanie Seltmann. In Vorbereitung. Referenzrahmen selbständige digitale Wissenschaftskommunikation.

Fröhlich, Gerhard. 1994. „Der (Mehr-)Wert der Wissenschaftskommunikation“. In: *Mehrwert von Information - Professionalisierung der Informationsarbeit*, hg. von Wolf Rauch, 84–95. Universitätsverlag Konstanz. <http://eprints.rclis.org/15101/> (zugegriffen: 2. August 2022).

Geier, Andrea und Markus Gottschling. 2019. Wissenschaftskommunikation auf Twitter? Eine Chance für die Geisteswissenschaften! *Mitteilungen des Deutschen Germanistenverbandes* 66, Nr. 3: 282–291. 10.14220/mdge.2019.66.3.282.

Gibbs, Robert. 2016. „Meeting Our Publics: A Search for the Right Questions in Public Humanities“. *University of Toronto Quarterly* 85, Nr. 4: 1–5.

Heinisch, Barbara. 2021. „Ein Pfad durch den Begriffsdschungel der Public Humanities“. *Public Humanities*, hg. von Lisa Kolodzie, Mareike Schumacher, Melanie Seltmann und Daniel Brenn, <https://publicdh.hypotheses.org/136> (zugegriffen: 1. August 2022).

Jacobson, Matthew Frye. 2020. „Afterword: The “Doing” of Doing Public Humanities“. In: *Doing Public Humanities*, hg. von Susan Smulyan. New York: Routledge.

Könneker, Carsten. 2017. „Wissenschaftskommunikation in vernetzten Öffentlichkeiten“. In: *Forschungsfeld Wissenschaftskommunikation*, hg. von Heinz Bonfadelli, Birte Fähnrich, Corinna Lüthje, Jutta Milde, Markus Rhomberg, und Mike S. Schäfer, 453–476. Wiesbaden: Springer Fachmedien 10.1007/978-3-658-12898-2_24.

Miller, Elizabeth, Edward Little und Steven High. 2017. *Going Public: The Art of Participatory Practice*. Toronto: University of British Columbia Press. <https://press.uchicago.edu/ucp/books/book/distributed/G/bo69987452.html> (zugegriffen: 1. August 2022).

Schäfer, Mike S., Sabrina H. Kessler und Birte Fähnrich. 2019. Analyzing science communication through the lens of communication science: Reviewing the empirical evidence. In: *Science Communication*, hg. von Annette Leßmöllmann, Marcelo Dascal, und Thomas Gloning, 77–104. Berlin, Boston: De Gruyter Mouton 10.1515/9783110255522-004.

Schumacher, Mareike. 2021. „What’s in it for us? Public Humanities als Herausforderung und Chance für die Digital Humanities“. *Public Humanities*, hg. von Lisa Kolodzie, Mareike Schumacher, Melanie Seltmann und Daniel Brenn, <https://publicdh.hypotheses.org/314> (zugegriffen: 1. August 2022).

Siegel, Jessica, Kolja Dunkel und Judith Terstriep. 2021. Hochschulen als Kommunikatoren der Wissenschaft? Eine exemplarische Bestandsaufnahme. *Forschung Aktuell* 06/2021. Institut Arbeit und Technik (IAT), Gelsenkirchen. <http://hdl.handle.net/10419/235877>.

Weitze, Marc-Denis und Wolfgang M. Heckl. 2016. „Wissenschaftskommunikation in sozialen Netzwerken“. In: *Wissenschaftskommunikation - Schlüsselideen, Akteure, Fallbeispiele*, hg. von Marc-Denis Weitze und Wolfgang M. Heckl, 189–195. Berlin, Heidelberg: Springer 10.1007/978-3-662-47843-1_18.

Ziegler, Ricarda und Liliann Fischer. 2020. *Ziele von Wissenschaftskommunikation - Eine Analyse der strategischen Ziele relevanter Akteure für die institutionelle Wissenschaftskommunikation in Deutschland, 2014-2020*. Berlin: Wissenschaft im Dialog. https://www.wissenschaft-im-dialog.de/fileadmin/user_upload/Projekte/Impact_Unit/Dokumente/210701_Ergebnisbericht_Strategische_Ziele_der_Wissenschaftskommunikation.pdf (zugegriffen: 9. Dezember 2022).

Re: ARTigo. Neuentwurf eines Social-Tagging-Frameworks aus funktionalen Programmbausteinen

Schneider, Stefanie

stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland

Kristen, Maximilian

max.kristen@campus.lmu.de
Ludwig-Maximilians-Universität München, Deutschland

Vollmer, Ricarda

ricarda.vollmer@campus.lmu.de
Ludwig-Maximilians-Universität München, Deutschland

Das Spiel¹ ist mehr als ein selbstgenügsamer Zeitvertreib für Kinder und Kind Gebliebene (Huizinga 1956). Längst, und substanziell beeinflusst durch die Genese neuerer, digitaler Spielformen, etabliert es sich als Tool, um Wissen nahezu *en passant* zu gewinnen: So begreifen *Games with a Purpose* (GWAPs; von Ahn und Dabbish 2004) ihre Spieler:innen als unersetzliches, online Wissen generierendes Kollektivum (Crowd), das sich mäßig begeisternden Aufgaben ohne materielle Entlohnung widmet. Die offensichtliche Trivialität jener *Microtasks* wird mit spielähnlichen Mechanismen kaschiert (Deterding et al. 2011, Morschheuser et al. 2017) und durch pädagogische Elemente angereichert (Suttie et al. 2012). Im Folgenden fokussieren wir mit dem originär durch von Ahn und Dabbish (2004) konzipierten *Extra Sensory Perception (ESP) Game* eine wiederholt rezipierte GWAP-Unterkategorie, die *Image-Labeling* -Prozesse anstößt. Einer der bekanntesten, für die *Digital Humanities* relevanten Ableger ist die seit 2010 an der Ludwig-Maximilians-Universität München entwickelte Plattform ARTigo.² In ihr wird das ESP-Game zuvorderst auf kunsthistorische Bildbestände appliziert.

Spielprinzip

Das Spiel fußt auf dem sog. *Output-Agreement*-Prinzip: Zwei anonym bleibenden Spieler:innen wird über eine *Graphical User Interface* (GUI) zeitgleich dasselbe Bild präsentiert. Ziel ist es, dieses Bild innerhalb einer festgelegten Zeitspanne mit Schlagwörtern (*Tags*) derart zu annotieren, dass möglichst viele Punkte durch übereinstimmende Angaben (*Matches*) erzielt werden. Die Schlagwörter sind frei wählbar und nicht kontrolliert wie in einem Thesaurus, können je nach Spielmodus aber durch *Tabuwörter* eingeschränkt werden. Tabuwörter stellen bspw. häufig vergebene Schlagwörter dar, denen kein zusätzlicher Informationsgewinn attestiert wird; bei Franz Marcs *Die großen blauen Pferde* (1911) etwa die Begriffe „Landschaft“ und „Pferde“. Die Spieler:innen müssen nicht simultan operieren. Stattdessen können alte Spielrunden synthetisiert und die für ein Bild dort vergebenen Schlagwörter recycelt werden.

Bekanntlich erzeugen Output-Agreement-Spiele aufgrund jener Matching-Prozedur vor allem oberflächlich beschreibende *Surface Tags* (Bry und Wieser 2012, Bry et al. 2018); in ähnlicher Qualität mittlerweile ebenso von immer leistungsfähigeren *Neuronalen Netzwerken* produziert (Heinisch 2021, Milani und Fraternali 2021, Zhao et al. 2021). Bspw. durch Tabuwörter implementierte *Scripting*-Mechanismen, die zur Eingabe semantisch ‚tiefergehender‘ Schlagwörter auffordern, begünstigen die sukzessive Diversifikation eines Tag-Bestands zwar (von Ahn und Dabbish 2004, Wieser et al. 2013). Allerdings beobachten wir zwei Defizite:

1. Scripting wird oft nicht in Form modularer *Add-ons* implementiert. An ihre Stelle treten vermeintlich autonome Spiele, die in sog. „Ökosystemen“ komplementär wirken sollen (Wieser et al. 2013, Bry et al. 2018). Disparate, inflexible Codebasen – wie in ARTigo – sind die Folge.
2. 92,26 % der Nutzer:innen von ARTigo initialisieren ein Spiel mit Tabuwörtern, ohne im Anschluss Tags zu vergeben.³ Selbst aktiv Partizipierende verschlagworten jedoch durchschnittlich 4,29 Taggings weniger als in nicht-restringierten Spielrunden. Weder motiviert die GUI hinreichend noch adressiert sie potenziell am Spiel Interessierte und leitet sie klar, etwa mithilfe eines textuellen Dialogsystems, an (Scherz 2017, 214–229).

Im vorliegenden Beitrag schlagen wir einen Neuentwurf der ARTigo-Plattform vor, und zwar als plugin-basiertes Social-Tagging-Framework, konstituiert aus strikt funktionalen Programmbausteinen. In ihm wird Scripting mithilfe eines *proaktiven Dialogagenten* (Baudoin et al. 2005) über die GUI initiiert. Bei Vergabe eines *Surface Tags* (z. B. „Pferd“) empfiehlt der Agent initiativ bedeutsame Unterbegriffe („Zugpferd“, „Pony“) oder motiviert abhängig vom Spielmodus dazu, weniger deskriptiv zu annotieren (z. B. durch Fragen, etwa: „Welche Gefühle löst das Bild bei Dir aus?“). Je nach zugrunde liegendem Annotationsinteresse wird Domänenwissen gezielt provoziert, sodass bildwissenschaftlich hoch spezifische Aufgaben an die Crowd ausgelagert werden können. Dies befördert nicht nur die Akquise kunst- und kulturhistorischer Daten, son-

dern verortet auch das Spiel als propädeutisches Lehrmittel im Kontext der Bildwissenschaften, wie in zwei prototypischen Anwendungsszenarien deutlich werden soll.

Infrastruktur

Konzipiert ist das Framework als dreistufige Architektur mit einer Daten-, Anwendungs- und Präsentationsschicht, die jeweils in Docker-Containern gekapselt sind, um eine Vielzahl von Installationsszenarien zu erlauben (Merkel 2014). Mit Grafana⁴ überwachen wir die Auslastung und Interaktionsfrequenz der einzelnen Instanzen. Eine übergeordnete Konfigurationsdatei speist die Anwendungs- und die Präsentationsschicht. Der Quellcode ist frei auf GitHub unter einer *GNU General Public License* veröffentlicht.⁵ Alle Komponenten sind *Open Source*.

Datenschicht

Primär besteht die Datenschicht aus einem PostgreSQL-Datenbankmanagementsystem.⁶ Damit auch ressourcenintensive Suchanfragen effizient bearbeitet werden, ist OpenSearch implementiert, Amazons seit 2021 entwickelter Elasticsearch- *Fork*.⁷

Anwendungsschicht

Auf die Datenschicht greift die Anwendungsschicht nur zu, wenn spielrelevante Informationen erfasst oder sie der Präsentationsschicht über eine REST-API zur Verfügung gestellt werden. Die offen zugängliche Schnittstelle ist mit Swagger UI⁸ und ReDoc⁹ an zwei interaktive Dokumentationsumgebungen nach OpenAPI-3.0-Spezifikation¹⁰ gekoppelt, sodass externe Plattformen gleichermaßen in der Lage sind, valide Abfragen zu formulieren.

Kern der in Python mit Django¹¹ realisierten Anwendungsschicht ist der sieben Modulgruppen umfassende Spielkonfigurationsmechanismus. Jede Modulgruppe hat genau eine Funktion mit strikt festgelegtem *Output*:

1. *Ressourcenmodule* selektieren die zu annotierenden Bilder. Als Kriterien dienen etwa die Anzahl der bereits zugewiesenen Tags oder wann ein Bild zuletzt bespielt wurde. Der mit Memcached¹² integrierte Cache-Server sorgt für eine ressourcenschonende Auswahl.
2. (*optional*) Durch *Opponentmodule* werden Akteure generiert, die ihre Tags automatisch gemäß der bis dato für ein Bild hinterlegten Schlagwörter wählen. Ihre Potenz ist konfigurierbar.
3. (*optional*) *Tabumodule* limitieren vorsätzlich den möglichen Tagraum. Sie verhindern Annotationen ohne zusätzlichen Informationsgewinn – wie am häufigsten für ein Bild vergebene Tags. Manuell definierte Restriktionen sind möglich („Impressionismus“ darf z. B. nicht für eine Reihe bekannt impressionistischer Bilder verwendet werden).

4. (optional) Programmatisch ähneln *Input*- den zuvor eingeführten Tabumodulen. Jedoch restringieren sie nicht, sondern leiten je nach Spieltypus an, bspw. indem sie auf zu validierende Tags hinweisen.
5. (optional) Mit *Vorschlagsmodulen* werden über Scripting feingranularere Termini als die bislang eingegebenen motiviert. Dazu binden wir nachweislich variationsreiche Glossare der englisch- und deutschsprachigen Wiktionary ein (Meyer und Gurevych 2010).¹³ Ebenso zurückgegeben werden signifikant in ähnlichen Kontexten verortete Tags, sog. *Kookkurrenzen*.
6. Die jeweils hinterlegten Tags evaluieren *Filtermodule* auf ihre prinzipielle Validität. Nicht zugelassen können etwa orthografisch inkorrekte, als Tabuwörter definierte oder in derselben Spielrunde bereits annotierte Tags sein.
7. (optional) *Bewertungsmodule* messen die Güte der validen Schlagwörter. Mit Punkten belohnt werden u.
 - a. gematchte Tags oder solche, die auf Basis der Wiktionary explizit domänenrelevante Information tragen – bei kunsthistorischen Bildern zur Komposition, Ikonografie und Epoche. Auch durch veränderte Bewertungsmodi wird Scripting herbeigeführt (Jain und Parkes 2013).

Um individuellere Nutzungsszenarien zu berücksichtigen, kann das Framework erweitert werden, indem der jeweiligen Modulgruppe ein geeignetes Untermodul hinzugefügt und die übergeordnete Konfigurationsdatei aktualisiert wird.

Präsentationsschicht

Die Präsentationsschicht ist in HTML und JavaScript mit dem JavaScript-Framework Vue.js¹⁴ programmiert. Alle dynamischen Webseiteninhalte werden mithilfe der JavaScript-Bibliothek axios¹⁵ über die REST-API geladen. Die Anwendung ist responsiv und auch für mobile Endgeräte optimiert.



Abb. 1: Dem Spiel vorgeschaltet ist eine mit Typed.js animierte Homepage, die zugleich als Tutorial fungiert.

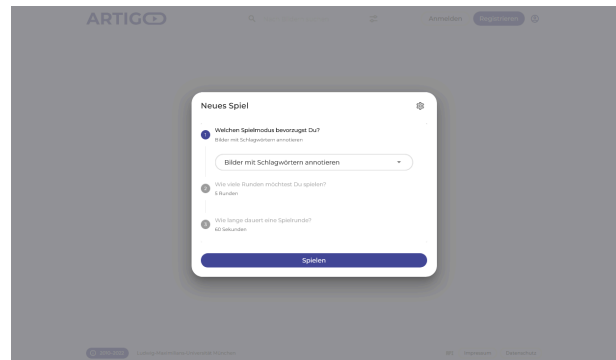


Abb. 2: Ein vertikaler Stepper assistiert den aus Fragen bestehenden Konfigurationsprozess.

Dem Spiel vorgeschaltet ist eine mit Typed.js¹⁶ animierte Homepage, die zugleich als Tutorial fungiert (Abb. 1, rechts). Zeitlich begrenzte *Quests*, die bspw. zum Tagging eines Künstlers reizen, sind in einem *Navigation Drawer*¹⁷ links gestellt (Abb. 1, links); sie werden zu meist randomisiert aus validen Modulkombinationen generiert. Ein vertikaler *Stepper*¹⁸ assistiert den folgenden, aus Fragen bestehenden Konfigurationsprozess (z. B. „Wie lange dauert eine Spielrunde?“ oder „Wie viele Runden möchtest Du spielen?“; Abb. 2). Er speist sich aus derselben Konfigurationsdatei wie die Anwendungsschicht. Alle Felder sind optional und mit Standardwerten vorbelegt, um in wenigen Klicks neue Spiele zu erstellen.

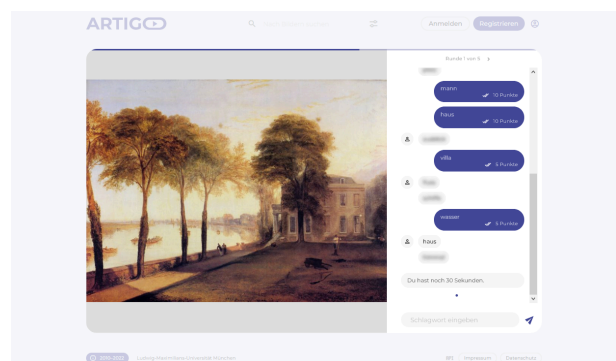


Abb. 3: Dominiert wird die Spiel-GUI von zwei Komponenten: dem zu annotierenden Bild, links, und der von *Instant-Messaging*-Diensten inspirierten Eingabemaske, rechts.

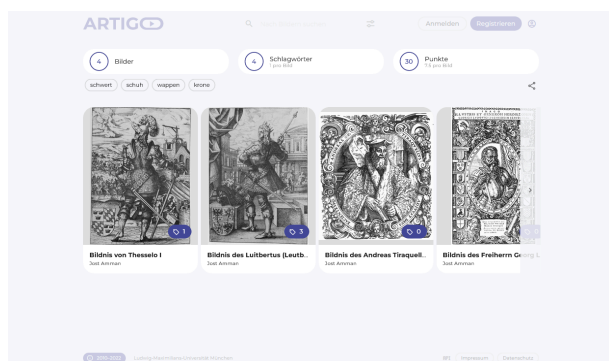


Abb. 4: Jede Spielsitzung endet mit einer nochmaligen Gegenüberstellung der Ergebnisse in aggregierter Form.

Dominiert wird die in Abb. 3 präsentierte Spiel-GUI von zwei Komponenten: dem zu annotierenden Bild, *links*, und der von *Instant-Messaging* -Diensten – etwa Metas Facebook Messenger¹⁹ – inspirierten Eingabemaske, *rechts*. Einen Chat(-bot) simulierend erscheinen dort sukzessive die Tags des zugeschalteten Akteurs; nicht-gematchte Eingaben sind zunächst unscharf. Damit jener trotz der synthetischen Anlage menschenähnlich – als tatsächlich Spielender – empfunden wird (Jenkins et al. 2007), verzögern wir seine Antworten minimal; derweil zeigt ein *Typing Awareness Indicator* den Eingabeprozess an (Auerbach 2014). Die GUI festigt so den *per se* dialogisch motivierten Aspekt der Verschlagwortung (Bredenkamp 2010). Scripting ist u. a. durch Steuerelemente wie Kombinations- und Optionsfelder eingebettet, sodass die Anzahl der auf dem Bildschirm präsenten Elemente reduziert wird. Ein Steuerelement ist demnach ausschließlich für die Zeit der Interaktion sichtbar. Zur Fokussierung auf den Annotationsprozess sind Kopf- und Fußzeile halbtransparent gesetzt. Der über beiden Komponenten – Bild und Eingabemaske – liegende, sie einschließende Fortschrittsbalken zeigt die Restlaufzeit der momentan aktiven Spielrunde an. Nach Ablauf dieser erscheint zunächst ein dreisekündiger Countdown, währenddessen die Spielsitzung pausiert werden kann, bevor das System zum nächsten Bild weiterschaltet. Jede Spielsitzung endet mit einer nochmaligen Gegenüberstellung der Ergebnisse in aggregierter Form (Abb. 4).

Um die Benutzerfreundlichkeit der GUI zu quantifizieren, beabsichtigen wir, Online-Fragebögen zu verschicken, die etablierte Metriken wie *Perceived Ease of Use*, *Perceived Usefulness* und *System Usability Scale* erfassen.

Use Cases

Nicht nur die kollaborative, durch die Crowd initiierte Akquise kunst- und kulturhistorischer Daten kann mit dem hier eingeführten Framework sinnvoll befördert werden,²⁰ sondern auch das Spiel als pädagogisches Instrument, etwa im musealen Raum. Nachweislich sind in vielen Fällen Inhalte spielerisch effektiver zu vermitteln als mit grundständigen Lernmethoden (Scherz 2017, 230). Dieses Potenzial schöpft das *digitale* Spiel aus, auch weil es ein Medium verwendet, das integraler Bestandteil der

Welt jüngerer Generationen ist; also derjenigen, auf die sich Bildungsbemühungen in erster Linie zu konzentrieren haben. Während das Spiel als propädeutisches Lehrmittel vermehrt in naturwissenschaftlich orientierten Domänen integriert wird,²¹ beabsichtigt das Framework, es auch im bildwissenschaftlichen Kontext zu verorten – etwa zum Studium der Grundlagen kunsthistorischer Argumentation.

Aufgrund seiner hochgradigen Modularisierung, und der damit einhergehenden Flexibilisierung der Parameter, sehen wir zwei prototypische Anwendungsszenarien, die im Folgenden exemplarisch beschrieben werden:

1. Im Fokus des *Attributionsspiels* steht wortwörtlich die Attribution eines Bildes zu einer näher definierten Oberkategorie: einem geografischen Kontext, einer Epoche, dem oder der Kuschtschaffenden selbst. Nicht zwingend entscheidend ist aber das exakte Matching. Stattdessen wird bereits eine regionale oder stilistische Annäherung an das interessierende Subjekt belohnt; etwa, wenn ein Kupferstich von Hans Baldung Grien fälschlicherweise Albrecht Dürer zugesprochen wurde. Der Schwierigkeitsgrad ließe sich erhöhen, indem das Bild als Explorationsfläche²² mit einer ‚Schablone‘ abgedeckt wird, die nur einen Teil des Bildes freigibt.²³ Durch sensorische Interaktion auf mobilen Geräten, etwa Berührungen oder Pusten, könnte der Ausschnitt bei gleichzeitigem Punktverlust modifiziert werden.
2. Ziel des *Ikongrafiespiels* ist die visuell motivierte Aneignung von historischen Sach- und Personenidentitäten. Jeweils vorgegeben ist eine Ikonografie, bspw. der heilige Hieronymus. Über die Eingabemaske werden die Ikonografie beschreibende Attribute verschlagwortet (z. B. „Löwe“, „Buch“, „Stundenglas“), während der textuelle Dialogagent sukzessive auf sie zutreffende Ikonografien empfiehlt. Durch die fortwährende Interaktion (Boiano et al. 2018) spüren die Partizipierenden der tatsächlich korrekten Ikonografie nach. Auf diese Weise vermittelt das Spiel zwar Fachwissen, schärft jedoch gleichermaßen implizit das kunsthistorische Sehen. Wie Scherz (2017, 226–227) nehmen wir an, dass u. a. die Dauer einer Spielrunde in ihrer Spezifik unterschiedliche Attributketten und daher Tagging-Muster begünstigt, die es weiter zu untersuchen gelte.

Immer treten die Spieler:innen in einen zweifachen pädagogischen Dialog: zum einen mit den jeweils angezeigten Bildern, zum anderen mit dem regelbasierten (Dialog-)System. Beide Spiele können dabei in kurzer Zeit verstanden und ihre Akzeptanz so wesentlich gesteigert werden (Kühn et al. 2009).

Zusammenfassung und Ausblick

Mit diesem Beitrag schlagen wir einen ganzheitlichen Neuentwurf der seit 2010 existierenden Social-Tagging-Plattform ARTigo vor, die kunsthistorische Image-Labeling-Prozesse anstößt. In dem nun aus funktionalen Programmbausteinen konstituierten Framework

wird Scripting – ein Mechanismus, der zur Eingabe tiefersemantischer Schlagwörter auffordert – mithilfe eines proaktiven, textuellen Dialogagenten motiviert. Wie in zwei prototypischen Anwendungsszenarien dargelegt, einem Attributions- und einem Ikonografiespiel, eignet sich das Framework für die Akquise kunst- und kulturhistorisch relevanter Daten ebenso wie als pädagogisches Instrument im Kontext der Bildwissenschaften.

Wir planen, die produzierten Taggings künftig automatisch in Wikidata²⁴ zu spiegeln. Um die Integration neuer Bilddatenbestände zu vereinfachen, soll des Weiteren eine Oberfläche zum Upload eigener Sammlungen bereitgestellt werden. Zur präziseren Datenerhebung entwickeln wir derzeit ein adaptierbares Werkzeug, mit dem komplexe Studiendesigns ermöglicht werden: So erlaubt es die freie Parametrisierung der Spielmodule, eine große Anzahl von Studienelementen zu quantifizieren und ausgewählten Testgruppen zuzuspielen. Neben grundsätzlich breiteren Forschungsmöglichkeiten kann durch jenen Prozess auch das jeweilige Spieldesign weiter optimiert werden. In diesem Kontext evaluieren wir auch jede Modulgruppe und loten ihre Eignung für unterschiedliche Forschungsszenarien aus.

Fußnoten

1. Im Deutschen wird nicht zwischen *Play* und *Game* unterschieden. Der vorliegende Beitrag stützt sich explizit auf das *Game* als wettstreitorientiertes, kompetitives und regelbasiertes Spiel (Walther 2003).
2. <https://www.artigo.org/>, wie alle folgenden letzter Zugriff 2. Dezember 2022.
3. Alle folgenden Statistiken basieren auf einem *Dump* der PostgreSQL-Datenbank vom 28. Juli 2022. Mit 56,65 % liegt der Wert für Spiele ohne Tabuwörter signifikant niedriger.
4. <https://github.com/grafana/grafana>.
5. <https://github.com/arthist-lmu/artigo>.
6. <https://github.com/postgres/postgres>.
7. <https://github.com/opensearch-project/OpenSearch>.
8. <https://github.com/swagger-api/swagger-ui>.
9. <https://github.com/Redocly/redoc>.
10. <https://github.com/OAI/OpenAPI-Specification>.
11. <https://github.com/django/django>.
12. <https://github.com/memcached/memcached>.
13. <https://de.wiktionary.org/wiki/>.
14. <https://github.com/vuejs/vue>.
15. <https://github.com/axios/axios>.
16. <https://github.com/mattboldt/typed.js/>.
17. <https://material.io/components/navigation-drawer>.
18. <https://material.io/archive/guidelines/components/steppers.html>.
19. <https://about.facebook.com/de/technologies/messenger/>.
20. Die grundlegende Eignung von ARTigo zur Datenakquise ist hinreichend belegt: Seit 2010 wurden 10.679.711 Taggings in der deutsch-, englisch- und französischsprachigen Version annotiert.
21. Ausnahmen bilden im deutschsprachigen Raum z. B. das Multimedia-Spiel „Ricardas Geheimnis“ des Wallraf-Richartz-Museums Köln (Vielhaber 2021). Im

angelsächsischen Raum ist das sieben Online-Spiele umfassende Projekt „Discover Ancient Egypt“ der National Museums Scotland zu nennen (<https://www.nms.ac.uk/explore-our-collections/games/discover-ancient-egypt/>).

22. So bereits kooperativ mit ARTigo in dem Top Citizen Science-Projekt „Stadt-Land-Bild. Eine soziale Bildanalyse zeitgenössischer Sehnsuchtserscheinungen“ zwischen 2018 und 2019 geschehen (<https://www.akbild.ac.at/de/forschung/projekte/forschungsprojekte/2019/stadt-land-bild.-eine-soziale-bildanalyse-zeitgeno308ssischer-sehnsuchtserscheinungen>).

23. Siehe Iose von Ahn et al. (2006).

24. <https://www.wikidata.org/wiki/>.

Bibliographie

- von Ahn, Luis und Laura Dabbish . 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.
- von Ahn, Luis, Ruoran Liu und Manuel Blum . 2006. „Peekaboom. A Game for Locating Objects in Images.“ In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 55–64 10.1145/1124772.1124782.
- Auerbach, David . 2014. „I Built That ‚So-and-So Is Typing‘ Feature in Chat. And I’m Not Sorry.“ In *Slate*, 14.02.2014.
- Baudoin, Frédéric, Philippe Bretier und Vincent Corruble . 2005. „A Dialogue Agent with Adaptive and Proactive Capabilities.“ In *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, 293–296 10.1109/IAT.2005.8.
- Boiano, Stefania, Ann Borda, Guiliano Gaia, Stefania Rossi und Pietro Cuomo . 2018. „Chatbots and New Audience Opportunities for Museums and Heritage Organizations.“ In *Electronic Visualisation and the Arts*, 164–171.
- Bredenkamp, Horst. 2010. *Theorie des Bildakts. Frankfurt Adorno-Vorlesungen 2007*. Frankfurt am Main: Suhrkamp.
- Bry, François, Clemens Schefels und Corina Schema-inda . 2018. „Eine qualitative Analyse der ARTigo-Annotationen.“ In Piotr Kuroczyński, Peter Bell und Lisa Dieckmann (Hrsg.). *Computing Art Reader. Einführung in die digitale Kunstgeschichte*. Heidelberg: arthistoricum.net, 96–114 10.11588/arthistoricum.413.c5771.
- Bry, François und Christoph Wieser . 2012. „Squaring and Scripting the ESP Game. Trimming a GWAP to Deep Semantics.“ In *Serious Games Development and Applications. SGDA 2012. Lecture Notes in Computer Science* 7528: 183–192 10.1007/978-3-642-33687-4_16.
- Deterding, Sebastian, Dan Dixon, Rilla Khaled und Lennart Nacke . 2011. „From Game Design Elements to Gamefulness. Defining ‚Gamification‘.“ In *Proceedings of the 15th International Academic MindTrek Conference. Envisioning Future Media Environments*, 9–15 10.1145/2181037.2181040.
- Heinisch, Barbara, Kristin Oswald, Maik Weißpflug, Sally Shuttleworth und Geoffrey Belknap . 2021. „Citizen Humanities.“ In Katrin Vohland, Anne Land-Zandt, Luigi Ceccaroni, Rob Lemmens, Josep Perelló, Ma-

risa Ponti, Roeland Samson und Katherin Wagenknecht (Hrsg.). *The Science of Citizen Science*. Berlin: Springer, 97–118.

Huizinga, Johan . 1956. *Homo Ludens. Vom Ursprung der Kultur im Spiel*. Reinbek: Rowohlt.

Jain, Shaili und David C. Parkes . 2013. „A Game-Theoretic Analysis of the ESP Game.“ In *ACM Transactions on Economics and Computation* 1.1 10.1145/2399187.2399190 .

Jenkins, Marie-Claire, Richard Churchill, Stephen Cox und Dan Smith . 2007. „Analysis of User Interaction with Service Oriented Chatbot System.“ In *International Conference on Human-Computer Interaction. HCI 2007. Lecture Notes in Computer Science* 4552: 76–83 10.1007/978-3-540-73110-8_9 .

Kühn, Eileen, Jens Reinhardt und Jürgen Sieck . 2009. „Spielen im Museum.“ In Jürgen Sieck und Michael A. Herzog (Hrsg.). *Kultur und Informatik. Serious Games*. Boizenburg: Verlag Werner Hülsbusch, 201–222.

Merkel, Dirk . 2014. „Docker. Lightweight Linux Containers for Consistent Development and Deployment.“ In *Linux Journal* 239: 2.

Meyer, Christian M. und Iryna Gurevych . 2010. „Worth its Weight in Gold or Yet Another Resource. A Comparative Study of Wiktionary, OpenThesaurus and GermaNet.“ In Alexander Gelbukh (Hrsg.). *Computational Linguistics and Intelligent Text Processing. 11th International Conference. Lecture Notes in Computer Science* 6008: 38–49.

Milani, Federico und Piero Fraternali . 2021. „A Dataset and a Convolutional Model for Iconography Classification in Paintings.“ In *Journal on Computing and Cultural Heritage* 14.4 10.1145/3458885 .

Morschheuser, Benedikt, Juho Hamari, Jonna Koivisto und Alexander Maedche . 2017. „Gamified Crowdsourcing. Conceptualization, Literature Review, and Future Agenda.“ In *International Journal of Human-Computer Studies* 106: 26–43 10.1016/j.ijhcs.2017.04.005 .

Scherz, Sabine . 2017. *Kunstgeschichte berechnet. Interdisziplinäre Bilddatenanalyse crowdgesourcter Annotationen*. Dissertation, Ludwig-Maximilians-Universität München.

Suttie, Neil, Sandy Louchart, Theodore Lim, Andrew Macvean, Wim Westera, Damian Brown und Damien Djaouti . 2012. „Introducing the ‚Serious Games Mechanics‘. A Theoretical Framework to Analyse Relationships Between ‚Game‘ and ‚Pedagogical Aspects‘ of Serious Games.“ In *Procedia Computer Science* 15: 314–315 10.1016/j.procs.2012.10.091 .

Vielhaber, Christiane . 2021. „Computerspiel für junge Besucher des Wallraf. Jagd nach Reliquien im Kölner Museum.“ In *Kölner Stadt-Anzeiger*, 02.07.2021.

Walther, Bo Kampmann . 2003. „Playing and Gaming. Reflections and Classifications.“ In *International Journal of Computer Game Research* 3.1.

Wieser, Christoph, François Bry, Alexandre Bérard und Richard Lagrange . 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*, 15–20.

Zhao, Wentao, Dalin Zhou, Xinguo Qiu und Wei Jiang . 2021. „Compare the Performance of the Models in Art Classification.“ In *PLOS ONE* 16.3 10.1371/journal.pone.0248414 .

Sammlungsdaten mit Wikidata anreichern und für die Vernetzung öffnen. Konzepte und praktische Erprobungen

Schelbert, Georg

georg.schelbert@hu-berlin.de

Humboldt-Universität zu Berlin, Deutschland

Müller, Michael

m.mueller@culture-to-go.de

Humboldt-Universität zu Berlin, Deutschland

Gegenstandsbereich und Ausgangslage

Die Aufgabe, kunst- und kulturhistorisch relevante Objekte digital zu dokumentieren, sie inhaltlich zu erschließen und über das Internet zugänglich und recherchierbar zu machen, stellt sich zunehmend nicht nur den „klassischen“ Gedächtnisorganisationen wie Museen, Bibliotheken und Archiven. Auch kleinere Sammlungen, etwa an Universitäten (vgl. Koordinierungsstelle für wissenschaftliche Universitätssammlungen in Deutschland, Empfehlungen 2016) oder themenspezifische Forschungs- und Erschließungsprojekte (vgl. Kieven, Schelbert 2014) stehen vor der Aufgabe, Daten zu Kulturgütern strukturiert zu erfassen, sie nachhaltig zu speichern und sie dann – eben im Sinne von *Open Humanities*, *Open Culture* – der Scientific Community und der interessierten Öffentlichkeit zur Nachnutzung zur Verfügung zu stellen (vgl. Schöch 2017).

Doch gerade ein solcher, wissenschaftsorientierter und zugleich Disziplinen übergreifender Ansatz ist mit zwei entgegengesetzten Problemen konfrontiert: Zum einen sind die zur Verfügung stehenden Ressourcen für die Beschaffung, Anpassung und Entwicklung von Infrastrukturen, für die inhaltliche Arbeit an Datenmodellen oder die Aufbereitung von Daten und Medien typischerweise sehr knapp bemessen. Zum anderen soll die digitale Erfassung hier oft deutlich mehr leisten als die klassische Objektdokumentation, da Wissensrepräsentation und Kontextualisierung eine größere Rolle spielen. Die traditionelle Vorstellung der Sammlungsdatenbank als eines abgeschlossenen, auf Vollständigkeit und Verbindlichkeit abzielenden digitalen Katalogs ist vor einem solchen Hintergrund zugleich unreichbar und unzureichend.

I. Konzeptionelles. Offene Referenzen als Schlüssel zu einer

zugleich mächtigen und offenen Datenstruktur

Interessante Perspektiven eröffnen in diesem Zusammenhang die Konzepte des Semantic Web bzw. der Linked Open Data: Durch die semantisch explizit definierte Verknüpfung öffentlich zugänglicher und über persistente Identifier (URI) adressierbarer Datenobjekte ergibt sich ein umfassender, offener Knowledge Graph, der sich aus den unterschiedlichsten epistemologischen Perspektiven befragen lässt. Allerdings ist dieses Konzept, aufs Ganze der digital verfügbaren Ressourcen von Museen, Archiven und Forschungseinrichtungen gesehen, bislang bestenfalls ansatzweise realisiert. Ein Beispiel sei genannt: Der Knowledge Graph des Konsortium NFDI4Culture ist als eines der ehrgeizigsten Projekte auf diesem Feld zunächst darauf beschränkt, die Ressourcen des Konsortiums als Linked Open Data auf der Basis einer Kultur-Ontologie zu beschreiben. (<https://nfdi4culture.de/resources/knowledge-graph.html>). Er bewegt sich also noch auf der Metaebene, die Einbeziehung der Domänen selbst, von konkreten Kulturgütern bis hin zu allgemeinen Sachverhalten wäre der nächste Schritt.

Es stellt sich also die Frage: Wie kann die eigene Sammlung Teil des universellen Knowledge Graph werden? Die Antworten erscheinen zunächst trivial: technisch durch die Publikation der Daten über eine Schnittstelle, konzeptionell durch die semantisch adäquate Datenmodellierung, bestenfalls nach einem zertifizierten Standard (etwa CIDOC CRM, ISO 21127:2014; LIDO oder EAD).

Nicht trivial sind allerdings, gerade im Hinblick auf die oben geschilderte Konstellation – forschungsbezogene Sammlungsdokumentation jenseits der klassischen Gedächtnisorganisationen mit sehr begrenzten Ressourcen –, die konkret und praktisch damit verbundenen Herausforderungen, insbesondere im Bereich der Datenmodellierung. Jedenfalls muss man zur Kenntnis nehmen, dass das Konzept der offenen und vernetzten Daten im Bereich der Sammlungsdokumentation noch zu einem sehr geringen Grad umgesetzt ist, obwohl nach unserer Einschätzung die Zielsetzung, Daten zu vernetzen, inzwischen bei vielen Verantwortlichen Zustimmung findet und zunehmend als Notwendigkeit verstanden wird. Auch ist oft die Bereitschaft zu erkennen, im Hinblick auf die eigene Arbeit umzudenken. Es fehlen aber die Perspektiven für eine Implementierung unter gegebenen infrastrukturellen Voraussetzungen.

Zwar stehen für Datenbank-Lösungen, die über klassische, katalogisierende Objektdokumentation hinausreichen, die Kontextinformation und umfassende semantische Zusammenhänge abbilden und damit vernetzbar sind, heute etwa mit WissKI, ResearchSpace oder Wikibase (vgl. Müller 2022) ausgereifte Open-Source-Systeme zur Verfügung, um die erwähnten Konzepte zu implementieren. Insbesondere kleineren Sammlungen und Forschungsprojekten, denen unser besonderes Interesse gilt, erscheint die notwendige Kombination aus Systemwechsel, Datenmigration und semantisch anspruchsvoller und normgerechter Datenmodellierung jedoch oft als Überforderung.

Dessen ungeachtet wäre prinzipiell zu fragen, ob die Ausstattung einzelner Institutionen oder Projekte mit hochkomplexen Datenbank-Management-Systemen ein uneingeschränkt erstrebenswertes Ziel ist. Man kann darin auch eine Sackgasse sehen: Die Planung, Implementierung, Wartung und Aufrechterhaltung dieser Großinfrastrukturen binden in hohem Maße Ressourcen und Aufmerksamkeit, während der reale Mehrwert von Nachnutzbarkeit und Vernetzung erfahrungsgemäß gerne etwas aus dem Blick gerät. Eher modular strukturierte, Agilität befördernde Ansätze, die wir zunächst und vor allem in Betracht ziehen, weil sie sich mit geringerem Ressourceneinsatz realisieren lassen, bieten hier aus unserer Sicht eine überlegenswerte Alternative oder zumindest Ergänzung. Der Fokus könnte sich von einer Optimierung der Vernetzbarkeit hin zu einer Vernetzung bereits in der Datenerfassung und -erschließung verschieben.

Ein weiterer Punkt: Die gerade im kulturhistorischen Bereich relevante und zumeist angestrebte Vernetzung mit globalen Wissensbeständen konnte damit noch nicht erreicht werden.

Wir setzen da an, wo auch traditionelle Datenbanken seit langem punktuell vernetzen. Bekanntlich bilden Normdaten den ersten Schritt des Verweisens über den eigenen Katalog hinaus. Normdaten wurden im Bibliothekswesen entwickelt, um die einheitliche Ansetzung von Elementen des Katalogs zu garantieren, die in mehr als einem Eintrag und nicht nur im einzelnen (Bibliothek-s-)Katalog vorkommen (Woitas 2013). Klassische Anwendungsfelder sind die Ansetzung von Personen- und Ortsnamen. Zunächst waren Normdaten damit primär Schreibregeln. Das Prinzip wurde erweitert auf die Normierung von Begriffen und deren Einordnung in klassifikatorische Systeme: Als Klassifikationen oder Thesauri – etwa bei der Schlagwortnormdatei (SWD) – erfüllen Normdaten erweiterte semantische Funktionen bei der inhaltlichen Erschließung.

Indem die Normansetzungen zentral gehalten werden und über einen persistenten Identifier stabil referenzierbar sind, werden sie zu Referenzdaten. Sie erzielen neben der Sicherung von Konsistenz und Qualität der Daten einen klassischen Normalisierungseffekt (im Sinne des Entity Relationship Model von E. F. Codd), nur dass die „normalisierenden“ Entity-Relationen über die eigene Datenbank hinausgreifen: Die Daten zur referenzierten Entität müssen (im Prinzip) nicht mehr selbst angelegt und gepflegt werden, es reicht die ID der Normansetzung, in der Praxis ergänzt durch importierte oder begleitende Kerndaten (vgl. zur Normdatennutzung im Kulturbereich Kailus, Stein 2018, Kett et al. 2019).

Damit beschränkt sich die Funktion des Verweisens nicht mehr auf die Funktion der Normierung, sondern eröffnet zusätzliche Aspekte. Nach Ansicht der Autoren sollte auch daher eher von Referenzdaten gesprochen werden. Wir identifizieren mindestens die folgenden Referenzdatentypen:

- kontrollierte Vokabulare: Ziel ist die Vereinheitlichung der Fachterminologie und die Sicherstellung semantischer Interoperabilität, d.h. dass in unterschiedlichen Ressourcen vom Gleichen die Rede ist, wenn derselbe *Begriff* verwendet wird.

- Identifikationsfunktion, Disambiguierung: Anspruch, bei Kulturgütern, Monumenten und Orten unverwechselbar deutlich zu machen, von welcher *Entität* die Rede ist.
- Taxonomien, Thesauri und Ontologien: Einrichtung einer begrifflich-inhaltlichen Ordnung der Gegenstandsbereiche, auch mit dem Anspruch, das weitere Konzept des Gegenstands zu erfassen.
- Lexikalische Anreicherung: zum Beispiel erweiterte biographische Angaben zu referenzierten Personen; historische Kontextualisierung von Gegenständen und Sachverhalten.
- Lokalisierung, Georeferenzierung als Sonderfall einer quantitativ definierten Verortung.

Derart konzeptualisiert, differenzieren sich die Funktionen, die Referenzdaten leisten können: Neben Normierung und Normalisierung spielen auch Identifikation, Erschließung und Anreicherung eine Rolle – und all dies in einer potentiell unendlich und für alle Anwendungsperspektiven skalierbaren Struktur, die daher auch offen sein muss für die möglichst umfangreiche Beteiligung der wissenschaftlichen Community.

Für die gestellte Aufgabe der Anreicherung und Öffnung einfacher Objektkataloge eröffnen die Referenzdaten in diesem erweiterten Verständnis interessante Perspektiven. Denn sowohl ihre strukturierte und stabile semantische Bestimmung als auch die Vernetzung findet (etwas pauschalisiert und idealisiert gesagt) auf Seiten der referenzierten Repositorien statt und muss nicht (nur) in der eigenen Datenhaltung geleistet werden. Die Realisierung eines semantischen Mehrwerts – Explizität, Präzision, Eindeutigkeit, und damit: Eignung für die Vernetzung der eigenen Daten – geht also theoretisch sogar einher mit einem Effizienzgewinn in der eigenen Datenhaltung. Die Umsetzung stellt jedoch neue Herausforderungen: Die Implementierung der Referenzierung von Daten in unzähligen Repositorien muss bewältigt werden. Vorrangiges Ziel muss es also sein, die Komplexität der vielfältigen Referenzierungen zu reduzieren.

Die Nutzung von Wikidata (<https://www.wikidata.org>, vgl. Vrandečić, Krötzsch 2014, Müller-Birn et al. 2015), der unter einer freien Lizenz verfügbaren Wissensdatenbank der Wikimedia-Familie, erscheint uns hier als der vielversprechendste Ansatz. Wikidata erfüllt als Infrastruktur und hinsichtlich des Datenbestandes einen erheblichen Teil der Anforderungen, die in unserem Bereich an Referenzdaten gestellt werden, und ist ein bedeutender Knoten im Linked Data Web (vgl. Erxleben et al. 2014).

Entscheidend ist zunächst, dass Wikidata nicht nur Funktionen traditioneller Normdaten erfüllen kann (vgl. Voß et al. 2014): Wikidata bietet mit seinem aus Items und Properties aufgebauten Datenmodell eine logische Struktur, mit den darin erfassten Daten einen der größten globalen Wissensbestände (vgl. die DHd-Beiträge Schelbert 2017 und Müller-Birn et al. 2018; zu den allgemeinen Potentialen von Wikidata für den Kulturerbesektor vgl. Krötzsch 2016, Poulter 2017, Schmidt et al. 2022). Darüber hinaus steht mit der zugrundeliegenden Software-Suite Wikibase auch eine konkrete Arbeitsplattform frei zur Verfügung, die in einer eigenen Implementierung genutzt werden könnte, was im Zusammenhang dieses Projekts jedoch nicht vorgesehen ist (anders als

etwa in den Projekten Rhizome und ArtBase, vgl. Rossetto 2021).

Unser Lösungsansatz sieht vor, von der eigenen Datenhaltung zunächst lediglich auf Wikidata zu referenzieren. Die Nutzung der Daten aus Wikidata kann von dieser Basis aus flexibel erfolgen: Das Konzept unseres Projekts verlangt keine Festlegung auf bestimmte Daten. Eine Option ist, Wikidata-Referenzen nur zu verwenden, um die Durchsuchbarkeit zu verbessern, eine naheliegende andere ist der indirekte Zugriff auf klassische Normdaten (Statements der Rubrik „Identifiers“), möglich ist etwa auch die Kategorisierung von Objekten (bspw. nach übergeordneten Kategorien, die in Wikidata festgehalten sind, bspw. Ort → Land).

Der Aufwand ist selbstverständlich skalierbar. Aus der geschilderten Ausgangslage (reiche Sammlungen, wenige Ressourcen) richtet sich unser Augenmerk aber (zunächst) auf möglichst einfache Lösungen mit dennoch beachtlichem Ertrag. Die Beurteilung der Erschließungsleistung misst sich bei unserem Ansatz nach dem Verhältnis von Aufwand und erzieltm Mehrwert für Erschließung und Vernetzbarkeit.

Der notorisch schwierigen und schwierig einzuschätzenden „Datenlage“ in Wikidata begegnen wir, indem wir uns auf Daten konzentrieren, die mit einer gewissen Zuverlässigkeit vorhanden sind, sowie durch die Einbeziehung von Daten aus den sekundär referenzieren „klassischen“ Normdatensätzen (GND etc.).

Auch wenn wir aus dem an sich viel reicheren Daten in Wikidata nur relativ basale Teilmengen verwenden, ergeben sich für die Erschließung des Materials erhebliche Verbesserungen, wenn man etwa für die auf einem historischen Lehr-Dia dargestellte Kirche auch Benennungsvarianten, die landessprachliche Bezeichnung, die Konfession, das Bistum oder die Geokoordinaten des Ortes für die Erschließung (aus Sicht der Nutzer*innen: für das Suchen, Filtern, Sortieren) zur Verfügung hat.

Die Komplexität der Daten in Wikidata (z.B. Unschärfe der Datierung), die nicht im eigenen System abgebildet werden kann oder soll bzw. nicht vollständig in der Suche wirksam gemacht werden kann, stellt eine Herausforderung dar, der zunächst pragmatisch mit Hinweisen (Disclaimer) begegnet wird. Die darüber hinaus mögliche Implementierung entsprechend komplexerer Lösungen für Datenhaltung, -suche und -visualisierung ist nicht Gegenstand unseres Vortrags.

Ein weiterer Aspekt ist die Frage nach der Form der Anbindung und damit der Aktualität der verwendeten Referenzdaten aus Wikidata. Hierbei soll nach folgenden Prinzipien vorgegangen werden, die wir im Fallbeispiel im Einzelnen ausführen.

1. Eine Aktualisierung, also der wiederholte Abgleich mit den Referenzdaten, soll grundsätzlich stattfinden (können).
2. Es muss steuerbar sein, welche Daten aktualisiert werden sollen, ob die selbst gepflegten Daten Priorität vor Daten haben, die über die Referenzierung gewonnen wurden.

Dem wichtigen Aspekt der Transparenz der Datenprovenienz wird durch die Kennzeichnung der aus Wikidata bezogenen Daten im Frontend Rechnung getragen.

II. Fallbeispiel. Die kunsthistorische Lehrbildsammlung Berlin

Die Lehrbildsammlung des Instituts für Kunst- und Bildgeschichte der Humboldt-Universität eignet sich in zweifacher Hinsicht, um diese Konzeption prototypisch zu erproben. Zum einen sind die skizzierten Limitierungen hier gegeben: Die Daten werden im Medienrepositorium der Universität verwaltet, das als Digital-Asset-Management-System (auf der Basis von ResourceSpace, <https://www.resourcespace.com/>) mit einer relativ flachen Datenstruktur an sich wenig geeignet ist, um einen Linked-OpenData-Ansatz zu realisieren. Zum anderen ist bei der Erschließung von kunsthistorischen Bildobjekten die Vernetzung mit weiteren Daten zum Kulturerbe inhaltlich besonders vielversprechend, da diese bekannte, singuläre Artefakte abbilden (vgl. Schelbert 2022).

Die kunsthistorische Lehrsammlung entstand mit der Gründung des Fachs seit den 1870er Jahren und stellt trotz erheblicher, teils kriegsbedingter Verluste eine der größten und reichhaltigsten ihrer Art dar (Haffner 2007, Schelbert 2018). Die Fotografien verschiedener Formate und Techniken (Papierabzüge, Glas- und Planfilmnegative, Glas- und Filmdiapositive) verweisen auf Kulturgut und Sachverhalte von großer zeitlicher und räumlicher Bandbreite. Sie interessieren außerdem in hohem Maß hinsichtlich fach-, institutions- und sammlungsgeschichtlicher Bezüge, wodurch sich die Herausforderungen der Referenzierung in besonderer Weise stellen. Es ist eine hohe Zahl an Digitalisaten vorhanden, denen die inhaltliche Erschließung fehlt, und es kann aufgrund der infrastrukturellen Bedingungen nur mit einfachen Datenbanksystemen und einfachen Datenmodellen gearbeitet werden.

Wikidata wird an dieser Stelle bereits bisher als Schlüssel sowohl zur Anreicherung als auch für die Anschlussfähigkeit der Digitalisate eingesetzt. Anstelle der Anlage komplexer Stammdaten werden die grundlegenden Entitäten (Personen, Geographica, Werke, Körperschaften) mit Wikidata-ID versehen und so mit den entsprechenden Wikidata-Items verknüpft. Potentiell kann das referenzierte Wikidata-Item die oben genannten Funktionen der Referenzierung erfüllen: Es identifiziert, disambiguiert, verweist auf andere Referenz-Repositorien, enthält Geodaten und in seinen Statements zahlreiche relevante Daten.

Allerdings sieht das Datenmodell im verwendeten DBMS ResourceSpace keine Einbindung von Referenzierungen vor. Ersatzweise sind die Wikidata-IDs deshalb lediglich in den Text der jeweiligen Metadatenfelder geschrieben. Da sie in ihrer Struktur (Q + Nummer) leicht zu identifizieren sind, sind sie prinzipiell maschinell verwendbar, jedoch nicht bzw. nur sehr mühsam in den von ResourceSpace vorgesehenen Bedienelementen.

Es stellt sich also die Frage, wie die in Wikidata enthaltenen Informationen auch in der konkreten Anwendung praktisch nutzbar gemacht werden können. Hier bedarf es einiger infrastruktureller Schritte, die wir in einem Pilotprojekt u.a. im Rahmen des Projekts Digitales Netzwerk Sammlungen der Berlin University Alliance ([https://www.ub.hu-berlin.de/de/ueber-uns/pro-](https://www.ub.hu-berlin.de/de/ueber-uns/projekte/digitales-netzwerk-sammlungen/projekt-digitales-netzwerk-sammlungen)

[jekte/digitales-netzwerk-sammlungen/projekt-digitales-netzwerk-sammlungen](https://www.ub.hu-berlin.de/de/ueber-uns/projekte/digitales-netzwerk-sammlungen/projekt-digitales-netzwerk-sammlungen)) erprobt haben.

Die technische Konzeption, an der wir uns orientieren, basiert auf den Prinzipien einer Separation of Concerns und der zeitlichen Entkopplung von Nutzung und Aggregation der (Referenz-)Daten. Gerade im Hinblick auf die genannten infrastrukturellen Limitierungen kann es nicht darum gehen, Sammlungsdatenbanken neu zu konzipieren oder in ihrer Struktur an die anspruchsvolle Referenzdatenvernetzung anzupassen. Wir ergänzen vielmehr die eingesetzten Systeme "von außen", also jenseits der Schnittstelle (im Idealfall ein dynamisch abfragbarer Endpoint, notfalls ein händischer Datenexport) mit einer modular aufgebauten Middleware (Data processing und data aggregator tools auf der Basis von Node.js; sekundärer Datenspeicher mit API), über die die semantisch reiche Verknüpfung geleistet wird. Schon auf der Ebene von Wikidata als Referenzdaten-Hub sind diese Operationen komplex und umfangreich, die Einbeziehung der dort referenzierten primären Repositorien (z.B. GND, Getty AAT, Iconclass ...) steigert die technischen Anforderungen, die sich aus den Verknüpfungsoperationen ergeben, weiter.

Auch hier suchen wir die Lösung nicht in der Skalierung, sondern in der Modularisierung. Module, die bei der Nutzung, etwa bei der Suche nach relevanten Objekten, beim Filtern und Sortieren, angesprochen werden, arbeiten mit bereits hochgradig aggregierten Daten, die praktisch instantan zur Verfügung stehen. Der Prozess der Aggregation läuft dagegen periodisch bzw. durch Triggerung aus der Datenhaltung, also zeitlich entkoppelt von der Nutzung ab. Mit anderen Worten: Der Umsetzungsvorschlag sieht keine dauerhafte Übernahme von Daten aus Wikidata vor. Vielmehr handelt es sich um eine temporäre, asynchron-dynamische Übernahme von Daten aus Wikidata zum Zweck der Suche bzw. zum Zweck der Anzeige angereicherter Daten in einem entkoppelten Frontend (hier realisiert in Vue.js).

Die Middleware, in die periodisch Datenabzüge aus Wikidata und der Quelldatenbank eingespielt werden, wird lediglich aus Performanzgründen eingesetzt, da eine reine on-the-fly-Lösung aufgrund der notwendigen kaskadierenden Abfragen in Wikidata nicht praktikabel wäre. Im Fallbeispiel der kunsthistorischen Bilddatenbank werden aus Wikidata Daten für abgebildete Werke in die Middleware übernommen, um auf dieser Basis schnelle Abfragen und Darstellungen von zusätzlichen Daten zum jeweiligen Bildobjekt zu erzeugen. Das Ausspielen der eigenen Daten nach Wikidata zur Korrektur und/oder Ergänzung des dortigen Datenbestandes wäre ein reizvoller weiterer Schritt im Sinne einer echten Vernetzung, den wir aber noch nicht gegangen sind.

Eine Vorstellung vom Aufbau einer nach diesen Prinzipien konzipierten Architektur vermittelt das folgende Schema:

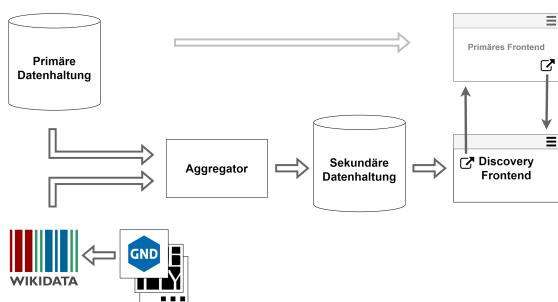


Abb. 1: Wikidata-basierte Anreicherung und Erschließung von Sammlungsdaten - Systemarchitektur des Pilotprojekts (Grafik M. Müller)

Bibliographie

Erxleben, Fredo. 2014. Michael Günther, Markus Krötzsch, Julian Mendez, Denny Vrandečić. "Introducing Wikidata to the Linked Data Web." in: Proceedings of the 13th International Semantic Web Conference. New York. 50-65

Haffner, Dorothee. 2007. "Die Kunstgeschichte ist ein technisches Fach." Bilder an der Wand, auf dem Schirm und im Netz." In *Bild/Geschichte. Festschrift für Horst Bredekamp*, hrsg. von Philine Helas, Maren Polte, Claudia Rückert, Bettina Uppenkamp, Berlin: Walter de Gruyter. 119-129

Kailus, Angela und Regine Stein. 2018. "Besser vernetzt: Über den Mehrwert von Standards und Normdaten zur Bilderschließung." In *Computing Art Reader*, hrsg. v. Piotr Kuroczyński, Peter Bell und Lisa Dieckmann, Heidelberg: arthistoricum. 119-139. <https://doi.org/10.11588/arthistoricum.413.c5772> (letzter Zugriff 15.12.2022)

Kett, Jürgen, Detlev Balzer, Barbara K. Fischer, Susanne Laux, Jens Lill, Jutta Lindenthal, Mathias Manecke, Martha I. Rosenkötter und Axel Vitzthum. 2019. "Das Projekt 'GND für Kulturdaten' (GND4C)." *o/bib*, Bd. 6 Nr. 4: 59-97. <https://doi.org/10.5282/o-bib/2019H4S59-97> (letzter Zugriff 15.12.2022)

Kieven, Elisabeth und Georg Schelbert. 2014. "Architekturzeichnung, Architektur und digitale Repräsentation. Das Projekt LINEAMENTA." *kunsttexte.de [Sektion Architektur Stadt Raum, Titel der themenspezifischen Ausgabe "Architecture on display"]* 4/2014: 1-7. <https://doi.org/10.48633/ksttx.2014.4.88343> (letzter Zugriff 15.12.2022)

Koordinierungsstelle für wissenschaftliche Universitätsbibliotheken in Deutschland. 2016. *Sammlungen an Universitäten*, herausgegeben vom wissenschaftlichen Beirat der Koordinierungsstelle. https://wissenschaftliche-sammlungen.de/download_file/view/1331/ (letzter Zugriff 15.12.2022)

Krötzsch, Markus. 2016. Wikidata as a Cultural Heritage Information Hub (Invited talk at the Europeana Network Association AGM 2016). <https://iccl.inf.tu-dresden.de/web/Misc3015> (letzter Zugriff 15.12.2022)

Lill, Jens M. 2019. "Gemeinsam neu definiert: das Projekt GND für Kulturdaten (GND4C)." *AKMB-news* 25, 1: 18-23. <https://doi.org/10.11588/akmb.2019.1.72474> (letzter Zugriff 15.12.2022)

Müller, Michael. 2022: *Wikibase – taugt die Wikidata-Software zur Sammlungsdokumentation?* (Teil 1), Blog Digitales Netzwerk Sammlungen https://dns.hypotheses.org/546?thumbnail_id=621 (letzter Zugriff 15.12.2022)

Müller-Birn, Claudia, Benjamin Karran, Janette Lehmann, Markus Luczak-Rösch. 2015: "Peer-production system or collaborative ontology development effort: what is Wikidata?" in Proceedings of OpenSym 2015 Conference on Open Collaboration. San Francisco. <https://doi.org/10.1145/2788993.2789836> (letzter Zugriff 15.12.2022)

Müller-Birn, Claudia, Georg Schelbert, Martin Raspe, Thorsten Wübbena. 2018. "Wikidata. Nutzungsmöglichkeiten und Anwendungsbeispiele für den Bereich Digital Cultural Heritage." In *DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum"*. Köln. 63-67. <https://doi.org/10.5281/zenodo.4622497> (letzter Zugriff 15.12.2022)

Poulter, Martin. 2017. Wikidata – the new hub for cultural heritage. Wikimedia UK-Blog, 20. Jan. 2017. <https://blog.wikimedia.org.uk/2017/01/wikidata-the-new-hub-for-cultural-heritage/> (letzter Zugriff 15.12.2022)

Rossenova, Lozana. 2021. Model-Database-Interface: A study of the redesign of the ArtBase, and the role of user agency in born-digital archives, PhD diss., London South Bank University. DOI: <https://doi.org/10.18744/lb-bu.8wz7x> (letzter Zugriff 15.12.2022)

Schelbert, Georg. 2017. "Warum nicht gleich Wikidata...?!" In *DHd 2017 - Digitale Nachhaltigkeit. 4. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum"*. Bern. 287-288. <https://doi.org/10.5281/zenodo.4622713> (letzter Zugriff 15.12.2022)

Schelbert, Georg. 2018. Bildgeschichte digital greifbar. Die Glasdiasammlung des Instituts für Kunst- und Bildgeschichte der Humboldt-Universität zu Berlin. Bericht von einem work in progress. Berlin. <https://edoc.hu-berlin.de/handle/18452/20233> (letzter Zugriff 15.12.2022)

Schelbert, Georg. 2022. "Die kunsthistorische Bilddatenbank zwischen digitalisierter Diathek und visuellem Wissensraum." In *Lehrmedien der Kunstgeschichte*, hrsg. v. Hubert Locher und Maria Männig. Berlin: Walter de Gruyter. 354-373

Schmidt, Sophie C., Florian Thiery und Martina Trognitz. 2022: "Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata." *Digital* 2(3): 333-364. <https://doi.org/10.3390/digital2030019> (letzter Zugriff 15.12.2022)

Schöch, Christoph. 2017. "Aufbau von Datensammlungen." In *Digital Humanities: Eine Einführung*, hrsg. v. Fotis Jannidis, Hubertus Kohle, und Malte Rehbein. Stuttgart: J.B. Metzler. 223-233. https://doi.org/10.1007/978-3-476-05446-3_16 (letzter Zugriff 15.12.2022)

Voß, Jacob, Susanna Bausch, Julian Schmitt, Jasmin Bogner, Viktoria Berkelmann, Franziska Ludemann, Oliver Löffel, Janna Kitroschat, Maiia Bartoshevskaja und Katharina Seljuzki. 2014. *Normdaten in Wikidata - Handbuch*, Version 1.0, 22.5.2014. <https://serviss.bib.hs-hannover.de/frontdoor/deliver/index/docId/438/file/normdaten-in-wikidata.pdf> (letzter Zugriff 15.12.2022)

Vrande#i#, Denny und Markus Krötzsch. 2014. " Wikidata: a free collaborative knowledge base ." Commun. ACM 57(10): 78–85. <https://doi.org/10.1145/2629489> (letzter Zugriff 15.12.2022)

Woitas, Kathi. 2013. Bibliografische Daten, Normdaten und Metadaten im Semantic Web – Konzepte der Bibliografischen Kontrolle im Wandel (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft). Berlin. <https://doi.org/10.18452/2086> (letzter Zugriff 15.12.2022)

Schrifttradition digital: Rituell reine Torarollen in der jüdischen Diaspora

Frank, Laura

laura.frank@kit.edu

Karlsruher Institut für Technologie, Deutschland

Eichhorst, Dana

dana.eichhorst@fu-berlin.de

Freie Universität Berlin, Deutschland

Ullrich, Rebecca

rebecca.ullrich@fu-berlin.de

Freie Universität Berlin, Deutschland

Hadassah Wendl, Katharina

katharina.wendl@fu-berlin.de

Freie Universität Berlin, Deutschland

Martini, Annett

annett.martini@fu-berlin.de

Freie Universität Berlin, Deutschland

Tonne, Danah

danah.tonne@kit.edu

Karlsruher Institut für Technologie, Deutschland

Einleitung

Rituell reine Torarollen sind ein außerordentliches kodikologisches, theologisches und soziales Phänomen der jüdischen Schrifttradition. Zum ersten Mal sollen diese besonderen Objekte im Zuge unseres Forschungsprojekts umfassend mit digitalen Mitteln erforscht werden.

Die Abschrift der heiligen Schriftrollen ist seit der Antike in ein dichtes Geflecht religionsgesetzlicher Regulierungen eingebunden. Die stark idealisierten Vorstellungen und Theorien der jüdischen Tradition zum Verhältnis

von Material, Reinheit und Heiligkeit im Schreibkontext werden durch eine reiche Kommentarliteratur ergänzt. Diese behandelt aus ethisch-philosophischer, mystischer oder magischer Perspektive die symbolische Bedeutung der materialen Elemente einer Torarolle, den rituellen Schreibprozess oder die außergewöhnlichen Charakteristika des Schreibers (Martini 2022).

Trotz der immensen Bedeutung, die den Torarollen als Mediatoren zwischen dem Heiligen und Profanen, der Vergangenheit und der Zukunft, aber auch in der jüdischen und nichtjüdischen Gesellschaft beigemessen wurde, beschränkte sich die bisherige Forschung weitestgehend auf die Untersuchung der überlieferten materialen Artefakte selbst. Dabei standen insbesondere die Schriftrollen vom Toten Meer im Fokus diverser Studien, die Textvarianten sowie die Beschaffenheit der Schreibhäute, der Tinten und bestimmte Merkmale des Layouts thematisieren. In der letzten Dekade führten Forscher*innen wie Judith Olszowy-Schlanger (Olszowy-Schlanger 2019), Mauro Perani (Perani 2019), Jordan Penkower (Penkower 2019, 2014), Josef M. Oesch (Oesch 2005, 2003) und Franz D. Hubmann (Hubmann und Oesch 2012) diesen Ansatz an ausgewählten Zeugnissen der mittelalterlichen Schrifttradition fort und bereicherten damit das Wissen um regionale Schrifttraditionen enorm – jedoch ohne das metaphysische Potential der Thematik auszuschöpfen.

Dieses Forschungsdesiderat wird im Rahmen des Verbundprojekts „Materialisierte Heiligkeit. Torarollen als kodikologisches theologisches und soziologisches Phänomen der jüdischen Schriftkultur in der Diaspora“ eingelöst. Gegenstand des Forschungsprojektes sind mittelalterliche Torarollen und Torarollenfragmente vornehmlich europäischer Provenienz sowie die umfangreiche Schreiberliteratur zur Herstellung rituell reiner Torarollen. Das Schriftbild und seine Besonderheiten bilden dabei einen eigenen Forschungsschwerpunkt, wobei erstmals auch Texte in den Fokus gerückt werden, die sich explizit mit den dekorativen Elementen der Buchstaben, den ‚Krönchen‘ (*Tagin*) oder den sogenannten ‚besonderen Buchstaben‘ (*Otiijot meshunnot*), beschäftigen.

Ziel ist es, den bedeutenden Schatz an bislang vernachlässigten Texten zu heben und erstmalig im Zusammenhang mit den überlieferten Torarollen und deren Geschichte zu lesen. Hierfür sollen die Torarollen als auch die relevanten Texte mit digitalen Mitteln aufbereitet und so deren Verständnis und Verknüpfung darstellbar gemacht werden.

Der Forschungsschwerpunkt: *Tagin* und *Otiijot meshunnot*

Laut traditionell-jüdischer Auffassung ist der Text der Tora heilig und jedes einzelne Element von semantischer und paläografischer Bedeutung – auch die dekorativen Elemente der einzelnen Buchstaben, deren Form, Ausgestaltung und Anordnung. Der unveränderliche heilige Text bildet das Zentrum der Kommentarliteratur.

In den Torarollen findet man neben den regulären Buchstaben der hebräischen Quadratschrift auch Buchstaben mit Verzierungen und besonderen Schreibwei-

sen. Die Darstellung dieser Krönchen und besonderen Buchstaben weicht jedoch in den überlieferten Handschriften stark voneinander ab und durchläuft erst im Laufe der Zeit einen Standardisierungsprozess. Im Mittelalter gibt es hingegen einen regelrechten „Wildwuchs“ der Buchstabenformen und *Tagin*, der sich in den Torarollen und in der Kommentarliteratur widerspiegelt. Die Quellen zeigen, dass es keine einheitliche und standardisierte Schreibweise der *Tagin* und besonderen Buchstaben in den Torarollen gab, was durch die unterschiedlichen Auslegungen und Diskurse in der Kommentarliteratur bestätigt wird.

Allein der kleine Midrasch Rabbi Aqiva al ha-Tagin (Midrasch Rabbi Aqiva über die Krönchen) bietet in jeder Version des in den Handschriften überlieferten Textes unterschiedliche Buchstabengestaltungen (beispielhaft in den Abbildungen 1-3).

unterschiedliche Buchstabengestaltungen (beispielhaft in den Abbildungen 1-3).



Abb. 1: rechts: Aleph ohne *Tagin*, MS Bodleian Libraries, Can. Or. 1 (1303-1304). Abb. 2: Mitte: Darstellung des Buchstaben Aleph mit drei *Tagin* oben, zwei auf dem linken Plateau und einen auf dem rechten Plateau, MS BL Harley 5510 (14.-15. Jahrhundert). Abb. 3: links: Darstellung des Buchstaben Aleph mit drei *Tagin* oben rechts und zwei unten links, MS Parma 2295 (13. bis 14. Jahrhundert).

Problematik

Die Quellen

Eine grundsätzliche Herausforderung stellt der heterogene Charakter und die Intertextualität des von uns bearbeiteten Textkorpus dar – sowohl im Hinblick auf die Textebene als auch auf die Komplexität der Handschriftenlage (die im Projekt bearbeiteten Texte liegen größtenteils nur in Handschriften vor). Der Midrasch Rabbi Aqiva al ha-Tagin beispielsweise ist in mind. 24 Handschriften und dabei in unterschiedlichen Versionen überliefert. Die Unterschiede in den Versionen des Textes betreffen teils nur marginal einzelne Wörter, teils recht gravierend die zentralen Inhalte des Textes. Ein umfassender Vergleich der Diskrepanzen innerhalb der Texte verlangt nach einer digitalen synoptischen Edition mit zentraler Darstellung.

Intertextualität und Bezug zwischen Texttradition und Torarollen

Welchen Rückschluss erlauben die Texte der jüdischen Tradition auf die erhaltenen Torarollen und wie lassen

sich Bezüge darstellen? Um die Vielzahl an relevanten Texten mit dem unveränderlichen Text der Tora zu analysieren und zu verknüpfen, bedarf es einer innovativen Herangehensweise. Die systematische Erfassung, Analyse und Interpretation der Texte sowie deren kodikologischer Besonderheiten ist erstmalig in diesem Umfang zu bewältigen. Diese komplexen Anforderungen zeigen deutlich, dass eine neue, digitale Forschungsmethodik unabdingbar ist, um die Texttraditionen zu kontextualisieren und intertextuelle Beziehungen hervorzuheben.

Verknüpfung von Form und Inhalt

Wie gezeigt, ist die Ausgestaltung der Buchstaben höchst variabel und lässt auf Entwicklungsschichten in der Ausprägung der besonderen Buchstaben und der *Tagin* schließen. Eine Schwierigkeit stellt die Beschreibung und Darstellung der *Tagin* und besonderen Buchstaben dar, die mit derzeitigen Möglichkeiten der (digitalen) Textverarbeitung unzureichend ist. Gleichzeitig führt die Erfassung der Gestaltung und Bedeutung der *Tagin* und *Otijot meshunnat* zu noch unbekannten Zusammenhängen und neuen Erkenntnissen. Die quellenübergreifende und strukturierte Analyse der Buchstabenverzerrungen im Kontext der Torarollen und Sekundärliteratur bringt das Verständnis der *Tagin* und *Otijot meshunnat* auf eine neue inhaltlich-philologische Ebene. Auch hier zeigt sich, dass eine umfassende Erforschung mit altbekannten Mitteln unzureichend ist und es neuer digitaler Zugänge bedarf.

Für alle Schwerpunkte des Projektes sind digitale Methoden und standardisierte Richtlinien unabdingbar. Diese werden in unserem Forschungskonzept vereint und stellen die Basis unseres Projektvorhabens dar. Die Konzeption und Verknüpfung der Methodiken sollen in diesem Vortrag erläutert werden.

Methodik

Das Editions-konzept

Insbesondere die digitale Edition, Kommentierung und englische Übersetzung der Schreiberliteratur, in der über kulturelle Grenzen hinweg in einem Zeitraum von etwa 1700 Jahren die Herstellung von rituell reinen Torarollen diskutiert wird, ist ein zentraler Baustein des Vorhabens. Ein entwickeltes digitales Editions-konzept sowie ein zugrundeliegendes Schema gemäß den Richtlinien des TEI-Konsortiums (TEI Consortium 2022) stellen die Basis einer einheitlichen historisch-kritischen Edition und inhaltlich-semantischen Kodierung für rabbinische, narrative und mystische Schreiberliteratur dar. Die eigens entwickelten Editionsrichtlinien ermöglichen es, die *Tagin* in den Torarollen und der Kommentarliteratur in einer digitalen Edition systematisch zu identifizieren und zu beschreiben.

Nachhaltigkeit und Vernetzung

Um die gesamtheitliche Erforschung dieser gesamten Daten und Metadaten nicht nur innerhalb dieses Projekts, sondern auch in Zukunft zu ermöglichen, wird ein digitales Forschungsdatenrepositorium gemäß der FAIR Prinzipien (Wilkinson 2016) zur zentralen Speicherung verwendet. In diesem werden sowohl die ausgewählten Torarollen als auch die diskutierten Handschriften und Sekundärliteratur in Form von strukturierten, digitalen Objekten inklusive eindeutiger Identifier und standardisierter Metadaten angelegt, um die Wiederverwendbarkeit der Daten und eine Vergleichbarkeit mit anderen Datensätzen zu gewährleisten. Mit Hilfe von standardisierten Schnittstellen können die digitalen Objekte sowohl maschinenlesbar als auch durch die Forschenden abgefragt werden. Dies ermöglicht nicht nur eine nachhaltige (Nach-) Nutzung der Daten, sondern auch eine gezielte Verknüpfung untereinander und eine kontinuierliche Erweiterung des Wissensspeichers durch neue Quellen.

Wie schon bei den Editions- und Erfassungsrichtlinien sowie dem Repositorium betont, legt unser Forschungskonzept Wert auf Standardisierung und eine leistungsfähige Forschungsdateninfrastruktur, weshalb zur Nomenklatur von spezifischen Charakteristika der Torarollen der Vokabulareditor EVOKS (Ernst 2022) verwendet wird. Das zugrundeliegende Konzept des Vokabulars beruht dabei auf dem von W3C empfohlenen Datenmodell SKOS (Simple Knowledge Organization System) (Miles 2009). Durch festgelegte Begriffe der schriftlichen Besonderheiten entsteht eine kontrollierte kodikologische Sammlung, welche dynamisch im Laufe der Forschung ergänzt und editiert werden kann. Jedoch sollen die handschriftlichen Merkmale nicht nur erfasst und thesauriert werden, sondern auch in den digitalen Versionen der Handschriften ausgezeichnet werden. Dazu steht ein Annotationsdienst zur Verfügung, der mit dem digitalen Repositorium in Verbindung steht (Tonne et al. 2019). Die heterogene und aussagekräftige Auszeichnung sowie Validierung der Annotationen folgt nach W3C-Empfehlung den Richtlinien des Web Annotation Data Model (Young et al. 2017).

Virtuelle Torarolle

Das Zusammenspiel aller genannten Komponenten wird in einer Virtuellen Torarolle gebündelt. Die Virtuelle Torarolle dient zum einen der Visualisierung und Präsentation der editierten Handschriften und der erfassten Torarollen, zum anderen bildet sie das Verbindungsstück von Forschungsdatenrepositorium und methodischen Werkzeugen. Die Anbindung des Vokabulareditors und des Annotationstools an die Virtuelle Torarolle bietet so die Möglichkeit, paläographische Details der Schrift und Besonderheiten des Schriftbildes der überlieferten mittelalterlichen Artefakte aufzunehmen und im Verhältnis zueinander und zu den Vorgaben der mittelalterlichen Regelwerke zu verknüpfen. Die qualitative und quantitative Analyse und Präsentation der Forschungsergebnisse in der Virtuellen Torarolle hat das Ziel, den Ursprung, den Wissens- und Praxiswandel und schließlich die Bedeutung der schriftlichen Besonderheiten im kulturellen Gedächtnis des Diasporajudentums freizulegen.

Anhand eines Fallbeispiels möchten wir zeigen, wie die verschiedenen Komponenten in unserem Forschungsablauf ineinandergreifen und vielfältige Potentiale und Forschungsmöglichkeiten der Handschriften und Torarollen bieten.

Fallbeispiel

Präsentiert wird der bereits oben erwähnte Midrasch Rabbi Aqiva al ha-Tagin, ein kurzer exegetisch-mystischer Text, der die Gründe für die Anzahl der Krönchen auf den Buchstaben diskutiert. Der ins Forschungsdatenrepositorium eingespeiste Text der Handschrift liegt in TEI-konformer Repräsentation als digitales Objekt vor, wobei nicht nur die Daten, sondern auch angereicherte strukturierte Metadaten gespeichert sind. Im Fokus des Fallbeispiels stehen die *Tagin* und *Otiyyot meshunnot*, deren erfasste Varianten im kontrollierten Vokabular eingesehen werden können, wobei jede Variante eine Relation zum übergeordneten unverzierten Buchstaben erhält. Des Weiteren zeigen wir die Visualisierung des Midrasch-Textes in der Virtuellen Torarolle und gehen auf ihre Konzeption und Implementierung ein. Anhand eines ausgewählten Buchstabens wird erläutert, wie dieser in der Virtuellen Torarolle annotiert ist und welche Analyse- und Visualisierungsschritte auf diese Weise möglich werden.

Das beschriebene Fallbeispiel zeigt nicht nur eine spannende kodikologische Anwendung des Forschungskonzepts, sondern illustriert das enorme Anknüpfungspotential des Vorhabens. Denn auch Erkenntnisse aus anderen Disziplinen wie der Material- und Sozialwissenschaft können in den Wissensspeicher einfließen und bisher unerforschte Verknüpfungen ermöglichen. Die Virtuelle Torarolle und der umfangreiche Werkzeugkasten eröffnen folglich neue methodische Zugänge zur Erforschung von Torarollen aus kodikologischer, geschichtlicher, material- und religionswissenschaftlicher Perspektive. Solch eine disziplinübergreifende Interpretation derart komplexer und heterogener Datenbestände trifft dabei auf den Kern der aktuellen Forschung und bietet immenses Forschungs- und Innovationspotential im Bereich der Digital Humanities.

Bibliographie

- Ernst, Felix. 2022. "Towards a Similarity Algorithm for Controlled Vocabularies Within the Digital Humanities." In *The Semantic Web. ESWC 2022 Satellite Events*, hg. von Paul Groth u.a., 179-188. <https://doi.org/10.1007/978-3-031-11609-4> (zugriffen: 26. Juli 2022).
- Hubmann, Franz und Josef M.Oesch. 2012. "Betrachtungen zu den Torarollen der Erfurter Handschriften-Sammlung. Untersuchungen zu Gliederung und Sonderzeichen." In *Erfurter Schriften zur jüdischen Geschichte, Bd. 1: Die jüdische Gemeinde von Erfurt und die SchUM-Gemeinden. Kulturelles Erbe und Vernetzung*, hg. von Landeshauptstadt Erfurt, 96-116. Jena: Bussett & Stadel.
- Lehmann, Manfred M. 1985). "Al Pe-in Lefufin." *Beit Mikra* 30,4: 449-455.

Martini, Annett. 2022. *„Arbeit des Himmels“: Jüdische Konzeptionen rituellen Schreibens in der europäischen Kultur des Mittelalters*. Berlin, Boston: De Gruyter.

Michaels, Marc. 2020. *Sefer Tagin-Fragments from the Cairo Geniza. A Critical Edition, Commentary and Reconstruction*. Leiden: Brill.

Miles, Alistair und Sean Bechhofer. 2009. *SKOS Simple Knowledge Organization System Reference. W3C Recommendation*. <https://www.w3.org/TR/2009/REC-skos-reference-20090818/> (zugegriffen: 26. Juli 2022).

Oesch, Josef M. 2003. „Skizze einer formalen Gliederungshermeneutik der Sifre Tora.“ In *Unit Delimitation in Biblical Hebrew and Northwest Semitic Literature*, hg. von Marjo C.A. Korpel und Josef M. Oesch, 162-203. Assen: Koninklijke Van Gorcum.

Oesch, Josef M. 2005. „Metatextelemente in hebräischen Torarollen.“ In *Von Sumer bis Homer. Festschrift für Manfred Schretter zum 60. Geburtstag am 25. Februar 2004*, hg. von Robert Rollinger, 521-533. Münster: Ugarit-verlag.

Olszowy-Schlanger, Judith. 2019. „The Making of the Bologna Scroll. Paleography and Scribal Traditions.“ In *The Ancient Sefer Torah of Bologna. Features and History*, hg. von Mauro Perani, 107-134. Leiden, Boston: Brill.

Penkower, Jordan S. 2014. „The Aschkenazi Pentateuch Tradition as Reflected in the Erfurt Hebrew Bible Codices and Torah Scrolls.“ In *Erfurter Schriften zur jüdischen Geschichte, Bd. 3: Zu Bild und Text im jüdisch-christlichen Kontext im Mittelalter*, hg. von Frank Bussert, 118-141. Jena u.a.: Bussert & Stadelers.

Penkower, Jordan S. 2019. „The 12th-13th Century Torah Scroll in Bologna. How It Differs from Contemporary Scrolls.“ In *The Ancient Sefer Torah of Bologna. Features and History*, hg. von Mauro Perani, 135-166. Leiden, Boston: Brill.

Perani, Mauro. 2019a. „Textual and Para-textual Devices of the Ancient Proto-Sefardic Bologna Torah Scroll.“ In *The Ancient Sefer Torah of Bologna. Features and History*, hg. von Mauro Perani, 53-106. Leiden, Boston: Brill.

Sanderson, Robert. 2017. *Web Annotation Protocol. W3C Recommendation*. <https://www.w3.org/TR/2017/REC-annotation-model-20170223/> (zugegriffen: 26. Juli 2022).

TEI Consortium (Hg.). 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 4.4.0.]*. <http://www.tei-c.org/Guidelines/P5/> (zugegriffen: 26. Juli 2022).

Tonne, Danah, Germaine Götzelmann, Philipp Hegel, Michael Krewet, Julia Hübner, Sibylle Söring, Andreas Löffler, Michael Hitzker, Markus Höfler, Timo Schmidt. 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> (zugegriffen: 26. Juli 2022).

Wilkinson, Marc, Michel Dumontier, u.a. 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship.“ *Sci Data* 3, 160018. <https://doi.org/10.1038/sdata.2016.18> (zugegriffen: 26. Juli 2022).

Young, Benjamin, Paolo Ciccarese, Robert Sanderson. 2017. *Web Annotation Data Model. W3C Recommendation*. <https://www.w3.org/TR/2017/REC-annotation-model-20170223/> (zugegriffen: 29. Juli 2022).

Zucker, Shlomo. 1977. „Ha-Otillot ha-Meshunnot. Ke-gon Lefufot we-ha-‘Aqummot.“ *Al Sefarim we-Anashim* 12: 5-12.

Selbstoptimierung vs. Selbstliebe? Eine vergleichende Inhaltsanalyse von Fitspiration- und Bodypositivity- Bildern auf Instagram mit Methoden der automatischen Bildklassifikation

Glas, Julia

Julia.Glas@psy.lmu.de

Ludwig Maximilians Universität München, Deutschland

Wolff, Christian

Christian.Wolff@informatik.uni-regensburg.de

Universität Regensburg, Deutschland

Ludwig, Bernd

Bernd.Ludwig@ur.de

Universität Regensburg, Deutschland

Achmann, Michael

Michael.Achmann@informatik.uni-regensburg.de

Universität Regensburg, Deutschland

Einleitung

Negative Auswirkungen schlankkeitsidealisierender Medieninhalte auf Körperbild und -zufriedenheit sind hinreichend belegt (z.B. Frederick et al. 2019, 193, 195; Grabe, Ward und Hyde 2008, 469-470). Inhalte, die Untergewicht oder Essstörungen glorifizieren (z.B. #thinspiration), sind mittlerweile auf Instagram verboten. Die *Fitspiration*-Bewegung („fitness“ + „inspiration“) stellt sich selbst als Alternative dar, die Unterstützung auf dem Weg zu einem gesünderen Lebensstil bieten will (Boepple et al. 2016, 133; Holland und Tiggemann 2017, 76). Empirische Untersuchungen belegen zwar ihr Potential, Follower:innen zu mehr Bewegung und gesunder Ernährung zu motivieren

(Santarossa et al. 2019, 381; Tiggemann und Zaccardo 2015, 66), weisen aber zugleich negative Auswirkungen auf Körperbild, Selbstzufriedenheit und Stimmung nach (Prichard et al. 2020, 4; Tiggemann und Zaccardo 2015, 64-65). Inhaltsanalysen von #fitspiration-Posts auf Instagram stellen fest, dass diese den Körpertyp „dünn und muskulös“ idealisieren und durch bestimmte Posen oder Kleidung bewusst in Szene setzen – die beworbene Sport- und Ernährungsweise dient also nicht der Förderung von Gesundheit und Wohlbefinden, sondern der „Optimierung“ des eigenen Körpers (Pilgrim und Bohnet-Joschko 2022, 117; Santarossa et al. 2019, 380; Tiggemann und Zaccardo 2018, 1007).

Bewusst als Gegenströmung bezeichnet sich die *Bodypositivity*-Bewegung und stellt es als ihr Ziel dar, gegen unrealistische Gewichts- und Körperideale einzutreten und so insbesondere Frauen zu mehr Körperakzeptanz und Selbstliebe zu verhelfen (Cwynar-Horta 2016, 40). Studien belegen teilweise positive Auswirkungen von Bodypositivty-Inhalten auf die Körperwertschätzung und -zufriedenheit (Cohen et al. 2019a, 1554-1556; Fioravanti et al. 2021, 13-14), stellen aber auch eine erhöhte Selbst-Objektifizierung fest (Cohen et al. 2019a, 1556-1557). Inhaltsanalysen belegen, dass #bodypositivity-Posts zwar Frauen unterschiedlichen Körperbaus repräsentieren, gleichzeitig aber häufig objektifizierende Merkmale (z.B. freizügige Kleidung, aufreizende Posen) enthalten (Cohen et al. 2019b, 51-52; Lazuka et al. 2020, 89). Außerdem mangle es Bodypositivity in anderen Dimensionen jenseits des Körpergewichts (z.B. Geschlecht, Ethnie, klassische Schönheitsideale) an Variabilität und Vielfalt (Cohen et al. 2019b, 51-52; Cwynar-Horta 2016, 40; Lazuka et al. 2020, 87).

Es ist daher die Frage zu stellen, ob die Selbstinszenierung von Bodypositivity als Gegenströmung zu Fitspiration (im Sinne von „Selbstliebe“ statt „Selbstoptimierung“) gerechtfertigt ist oder ob beide Bewegungen doch mehr Gemeinsamkeiten als Unterschiede aufweisen. Bisherige Inhaltsanalysen zu Instagram-Bildern betrachten immer entweder nur Fitspiration- oder nur Bodypositivity-Posts, was einen inhaltlichen Vergleich beider Bewegungen erschwert. Ansatz des hier vorgestellten Projekts ist es daher, Bilder beider Hashtags hinsichtlich derselben inhaltlichen Kategorien zu analysieren. Während in früheren Studien Bilder ausschließlich per Hand kodiert wurden, finden zudem erstmals Methoden der automatischen Bildklassifikation Anwendung, um statistische Gegenüberstellungen von Fitspiration und Bodypositivity an einem großen Bilddatensatz vornehmen zu können.

Korpus und händische Annotation

Das Korpus umfasst 10000 Bilder, die im Zeitraum März bis April 2022 mit einem der Hashtags #fitspiration oder #bodypositivity auf Instagram gepostet wurden. Eine zufällig gezogene Stichprobe von je 500 Bildern pro Hashtag wurde manuell hinsichtlich verschiedener für einen Vergleich der beiden Bewegungen relevanter inhaltlicher Kategorien kodiert. Betrachtet wurden allgemeine Kategorien zur Art des Beitrags (Bild, Text) und

des Abbildungsgegenstands (Person, Lebensmittel) sowie spezifischere Kategorien zur Analyse abgebildeter Personen aus den Bereichen Demographie, körperbezogene Eigenschaften, Kleidung, Freizügigkeit, Art der Pose und Anzeichen einer Objektifizierung. Diese sind in Tabelle 1 mit den Häufigkeiten der einzelnen Ausprägungen unter den Bilddaten aufgeführt. Zur Absicherung der Objektivität der Kodierung wurde ein Teil der Bilder von zwei Personen unabhängig kodiert und die Inter-Koder-Reliabilität bestimmt. Für die einzelnen inhaltlichen Kategorien wurden zwischen 200 und 800 Bilder doppelt kodiert, sodass Cohen's Kappa Werte über 0.7 für alle Kategorien erzielt werden konnten.

Automatische Bildklassifikation

Um den Vergleich von Fitspiration- und Bodypositivity-Bildern auf Instagram auf einer möglichst großen Datenbasis durchführen zu können, wurde versucht, die verbleibenden 9000 Bilder mit an den händisch kodierten Daten trainierten Modellen automatisch zu klassifizieren. Der manuell kodierte Bilddatensatz ($N = 1000$) wurde dazu im Verhältnis 4:1 in Trainings- und Testdaten aufgeteilt. Implementiert wurden neben Modellen aus dem Bereich des traditionellen *Machine Learning* (*Support Vector Machine SVM*, *Decision Tree DT*) als aktuelles Beispiel für *Deep Learning*-Verfahren auch *Convolutional Neural Networks* (CNN), die relevante Bildmerkmale selbst extrahieren können (Guo et al. 2017, 721).

Da der Trainingsprozess von CNN-Modellen eigentlich einen weitaus größeren Trainingsdatensatz erfordert (Sheykhmousa et al. 2020, 6309), wurde versucht, die Performanz der Modelle durch Transferlernen zu verbessern. Dabei werden an großen Datensätzen vortrainierte Modelle an wenige annotierte Daten einer neuen Domäne angepasst (*Fine Tuning*, vgl. Boumaraf et al. 2021, 3). Im vorliegenden Projekt wurde das am umfassenden Bilddatensatz *ImageNet* trainierte Modell VGG16 (Swason, Tjandrasa und Fathicah 2019, 178) als Grundlage verwendet, da auf diesem Modell basierende Transferlernprozesse in der Vergangenheit in verschiedenen Domänen die Klassifikationsergebnisse verbessern konnten (Boumaraf et al. 2021, 23; Dubey und Jain 2020, 5; Swason, Tjandrasa und Fathicah 2019, 179-181). Für das hier vorgestellte Projekt wurden die letzten vier *Dense Layers* aus der Netzwerkarchitektur entfernt und durch ein weiteres *Pooling Layer*, ein *Dense Layer*, ein *Dropout Layer* (Dropout-Rate 50%) und ein abschließendes *Dense Layer* zur finalen Klassenvorhersage (Aktivierungsfunktion Sigmoid) ersetzt. Nur die Gewichte dieser äußeren Schichten wurden während des *Fine Tunings* (Adaption an den Instagram-Bilddatensatz) neu trainiert, die vortrainierten Gewichte der tieferen Schichten werden beibehalten.

Da für einige inhaltliche Kategorien ein starkes Klassenungleichgewicht im Trainingsdatensatz bestand (z.B. bei der Kategorie „nackte Haut“: Nur 20% der untersuchten Bilder zeigten keine nackten Hautstellen), wurde mit *Upsampling* gearbeitet: Mithilfe von SMOTE wurden zusätzliche Daten der unterrepräsentierten Klasse(n) syn-

thetisch erzeugt, um ein Klassengleichgewicht herzustellen (Chawla et al. 2002, 328).

Tabelle 1 zeigt für die verschiedenen Inhaltskategorien, welche Anteile der Daten pro Klasse durch das für diese Kategorie am besten funktionierende Modell korrekt klassifiziert werden konnten. Teilweise erzielten implementierte SVM-Modelle die besten Ergebnisse, die meisten Kategorien ließen sich aber mit einem der CNN-Modelle am besten klassifizieren. Die auf Transferlernen basierenden CNN-Modelle zeigten in allen Fällen bessere Resultate als die CNNs ohne Transferlernen. Auch das Upsampling der unterrepräsentierte(n) Klasse(n) wirkte sich in einigen Fällen positiv auf die Performanz aus.

Wie Tabelle 1 zu entnehmen ist, gelang dennoch nicht für alle Kategorien mit einem der implementierten Modelle die Klassifikation mit zufriedenstellenden Ergebnissen. Probleme zeigten sich insbesondere bei multinominalen Klassifikationsproblemen, also Kategorien, die mehr als zwei Ausprägungen unterscheiden. In solchen Fällen, wie etwa dem Körperbau, der beim händischen Kodieren auf einer 9-stufigen Skala (Pulvers et al. 2014, 1643) eingeordnet worden war, wurden für die automatische Klassifikation mehrere Klassen zusammengelegt; dennoch wurden nur unzureichende Klassifikationsergebnisse erzielt. Auch die automatische Klassifikation von Kategorien mit starkem Klassenungleichgewicht führte zu Schwierigkeiten, etwa bei der Kategorie „Entsprechung klassischer Schönheitsideale“: Bilder von Personen, die klassischen Schönheitsidealen überhaupt nicht entsprechen, waren im händisch annotierten Datensatz selten und wurden von dem implementierten Klassifikator nur unzuverlässig erkannt, wohingegen Bilder der häufigeren Klasse „normale/extreme Entsprechung“ meistens korrekt klassifiziert wurden.

Für die Kategorien, die im Testdatensatz mit einem der implementierten Modelle gut klassifiziert werden konnten, wurde das jeweilige Modell auf den Gesamtdatensatz angewandt. Das war insgesamt nur für acht Kategorien der Fall, fünfmal wurde dabei ein CNN-Modell mit Transferlernen eingesetzt, dreimal wurde ein traditionelles *Machine Learning*-Modell angewandt. Die statistischen Vergleiche zwischen #fitspiration und #bodypositivity konnten in diesen Fällen anhand des automatisch klassifizierten Gesamtdatensatzes (N = 10000) erfolgen, für die übrigen Kategorien wurden die statistischen Tests nur mit den 1000 händisch annotierten Daten durchgeführt.

Tabelle 1: Übersicht der inhaltlichen Kategorien mit jeweils am besten klassifizierenden Modell sowie der Häufigkeitsverteilungen in händisch annotiertem und Gesamtdatensatz

Inhaltliche Kategorie	Modell mit den besten Klassifikationsergebnissen	Anteil richtig klassifizierter Daten pro Klasse (Testdatensatz) ¹	Klassenverteilung händisch annotierter Datensatz (N = 1000)	Klassenverteilung Gesamtdatensatz (N = 10000)
Beitrag enthält ein Bild	CNN mit Transferlernen, mit Upsampling der Klasse „Nein, kein Bild“	Ja: Bild enthalten: 97% Nein, kein Bild enthalten, Beitrag ist rein textbasiert: 93%	Fitspiration: Ja: 93,0% Nein: 7,0% Bodypositivity: Ja: 92,2% Nein: 7,8%	Fitspiration: Ja: 83,8% Nein: 6,2% Bodypositivity: Ja: 81,9% Nein: 8,1%
Abbildung einer Person	CNN mit Transferlernen, ohne Upsampling	Ja: 84% Nein: 54%	Fitspiration: Ja: 60,0% Nein: 40,0% Bodypositivity: Ja: 65,4% Nein: 34,6%	Fitspiration: Ja: 63,5% Nein: 36,5% Bodypositivity: Ja: 63,7% Nein: 36,3%
Abbildung eines Lebensmittels	CNN mit Transferlernen, ohne Upsampling	Ja: 55% Nein: 100%	Fitspiration: Ja: 4,9% Nein: 95,1% Bodypositivity: Ja: 6,9% Nein: 93,1%	Fitspiration: Ja: 2,1% Nein: 97,9% Bodypositivity: Ja: 2,2% Nein: 97,8%
Geschlecht	SVM mit Upsampling der Klasse „männlich“	Männlich: 74% Weiblich: 71%	Fitspiration: Männlich: 48,0% Weiblich: 53,0% Bodypositivity: Männlich: 19,0% Weiblich: 81,0%	Fitspiration: Männlich: 46,9% Weiblich: 53,1% Bodypositivity: Männlich: 38,1% Weiblich: 61,9%
Ethnie ²	SVM ohne Upsampling	Weiß: 79% Andere: 29%	Fitspiration: Weiß: 81,0% Andere: 19,0% Bodypositivity: Weiß: 82,9% Andere: 17,1%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Körperbau	SVM ohne Upsampling	Nicht beurteilbar: 42% Skalawert 1-3: 23% Skalawert 4: 31% Skalawert 5: 15% Skalawert 6-9: 35%	Fitspiration: 1-3: 23,9% 4: 39% 5: 19,7% 6-9: 6,7% Nicht beurteilbar: 10,7% Bodypositivity: 1-3: 12,5% 4: 20,2% 5: 20,8% 6-9: 24,1% Nicht beurteilbar: 22,3%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Muskulosität	SVM mit Upsampling der Klasse „nicht muskulös“	Nicht muskulös: 0% Sichtbar muskulös: 38% Sehr/extrem muskulös: 28% Nicht beurteilbar: 54%	Fitspiration: Nicht muskulös: 1,3% Sichtbar muskulös: 44,3% Sehr/extrem muskulös: 39,3% Nicht beurteilbar: 17,0% Bodypositivity: Nicht muskulös: 13,7% Sichtbar muskulös: 30,6% Sehr/extrem muskulös: 15,5% Nicht beurteilbar: 40,1%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Entsprechung klassischer Schönheitsideale	CNN mit Transferlernen, mit Upsampling der Klasse „überhaupt nicht“	Überhaupt nicht: 16% Normal / extrem: 81% Nicht beurteilbar: 44%	Fitspiration: Gar nicht: 1,7% Normal: 52,7% Extrem: 15,7% Nicht beurteilbar: 30,0% Bodypositivity: Gar nicht: 5,2% Normal: 47,1% Extrem: 23,5% Nicht beurteilbar: 23,5%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Art der Kleidung	CNN mit Transferlernen, mit Upsampling der Klasse „Sportkleidung“	Sportkleidung: 54% Sonstige Kleidung: 78%	Fitspiration: Unterwäsche: 11,3% Sportkleidung: 61,5% eng: 11,7% normal: 10,1% weit: 4,1% Bodypositivity: Unterwäsche: 23,2% Sportkleidung: 20,5% eng: 20,2% normal: 24,2% weit: 6,4%	Fitspiration: Sportkleidung: 52,7% Sonstige Kleidung: 47,3% Bodypositivity: Sportkleidung: 30,7% Sonstige Kleidung: 69,3%
Abbildung nackter Haut	SVM mit Upsampling der Klasse „Nein“	Ja: 79% Nein: 50%	Fitspiration: Ja: 85,7% Nein: 14,3% Bodypositivity: Ja: 74,6% Nein: 25,4%	Fitspiration: Ja: 63,3% Nein: 36,7% Bodypositivity: Ja: 64,3% Nein: 35,7%

Abbildung nackter Bauch	DT mit Upsampling der Klasse „Ja“	Ja: 58% Nein: 63%	Fitspiration: Ja: 44,4% Nein: 55,6% Bodypositivity: Ja: 41,0% Nein: 59,0%	Fitspiration: Ja: 38,8% Nein: 61,2% Bodypositivity: Ja: 38,3% Nein: 61,7%
Person in Bewegung	CNN mit Transferlernen, mit Upsampling der Klasse „Ja“	Ja: 36% Nein: 91%	Fitspiration: Ja: 31,3% Nein: 68,7% Bodypositivity: Ja: 13,1% Nein: 86,9%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Objektifizierung: Fokus auf bestimmtes Körperteil	SVM mit Upsampling der Klasse „Ja“	Ja: 44% Nein: 82%	Fitspiration: Ja: 24,3% Nein: 75,7% Bodypositivity: Ja: 25,7% Nein: 74,3%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Objektifizierung: Aufreizende Pose	CNN mit Transferlernen, mit Upsampling der Klasse „Ja“	Ja: 23% Nein: 95%	Fitspiration: Ja: 9,0% Nein: 91,0% Bodypositivity: Ja: 25,4% Nein: 74,6%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Objektifizierung: Gesicht nicht erkenntlich	SVM mit Upsampling der Klasse „Ja“	Ja: 41% Nein: 72%	Fitspiration: Ja: 32,7% Nein: 67,3% Bodypositivity: Ja: 30,0% Nein: 70,0%	/ (keine Modell-anwendung auf Gesamtdatensatz)
Vorher-Nachher-Vergleich	CNN mit Transferlernen, ohne Upsampling	Ja: 90% Nein: 99%	Fitspiration: Ja: 8,1% Nein: 91,9% Bodypositivity: Ja: 4,0% Nein: 96,0%	Fitspiration: Ja: 2,1% Nein: 97,9% Bodypositivity: Ja: 1,1% Nein: 98,9%

Befunde zum Vergleich der Fitspiration- und Bodypositivity-Bilder

Statistische Vergleiche der Fitspiration- und Bodypositivity-Bilder bezüglich der untersuchten Kategorien konnten sowohl Unterschiede als auch Gemeinsamkeiten in der Häufigkeit bestimmter Merkmale nachweisen (siehe auch Tabelle 1). Bilder beider Hashtags zeigen zumeist mindestens eine nackte Körperstelle: Die abgebildeten Personen tragen häufig sehr freizügige Kleidung (bei Fitspiration meist knappe Sportkleidung, bei Bodypositivity eher Unterwäsche/Badebekleidung oder enge Alltagskleidung). Auch weitere Anzeichen einer Objektifizierung (Bildfokus auf ein Körperteil gerichtet, anzügliches Posieren, Gesicht nicht erkenntlich) und die Tatsache, dass Personen hauptsächlich in statischer Pose statt in Bewegung abgebildet sind, belegen für beide Hashtags die eindeutige Fokuslegung auf das äußere Erscheinungsbild. Unterschiede zwischen den Hashtags zeigen sich hinsichtlich körperbezogener Merkmale: Frauen haben auf Fitspiration-Bildern einen signifikant schlankeren Körperbau und mehr Muskelmasse als auf Bodypositivity-Bildern, wo mehr unterschiedliche Körperformen repräsentiert sind. Die Betonung klassischer Schönheitsideale (z.B. reine Haut, glänzende Haare) sowie die Überrepräsentation weißer Personen lässt sich hingegen beiden Bewegungen vorwerfen: Personen mit äußeren „Makeln“ wie Cellulite oder Hautunreinheiten sind unter beiden Hashtags deutlich unterrepräsentiert, ebenso wie nicht-weiße Personen.

Zusammenfassend lässt sich festhalten, dass unter dem Hashtag #fitspiration hauptsächlich schlanke und muskulöse Frauen und Männer in Sportkleidung so posieren, dass ihr durchtrainierter Körper möglichst gut zur Geltung kommt. Bodypositivity hingegen repräsentiert zwar mehr unterschiedliche Körpertypen – ist dafür

aber größtenteils auf Bilder von Frauen beschränkt, die klassischen Schönheitsidealen entsprechen. Beide Bewegungen setzen durch freizügige Kleidung, Nacktheit und Objektifizierungen einen starken Fokus auf das äußere Erscheinungsbild.

Fazit und Ausblick

Die durchgeführte Analyse von #fitspiration- und #bodypositivity-Bildern auf Instagram mit einheitlichen inhaltlichen Kategorien ermöglichte erstmals eine konkrete Gegenüberstellung beider Bewegungen hinsichtlich relevanter inhaltlicher Bildmerkmale. Durch die implementierten Modelle zur automatischen Bildklassifikation konnten die Analysen sich zudem zumindest teilweise auf einen großen Bilddatensatz (N = 10000) stützen. Zwar zeigen sich hinsichtlich Körperbau und Muskulosität die erwarteten Unterschiede zwischen den beiden Hashtags, in den meisten anderen Dimensionen unterschieden sich Bilder beider Bewegungen aber nicht signifikant. Die Selbstinszenierung der Bodypositivity-Bewegung als Gegenströmung zu Fitspiration lässt sich daher nicht belegen: Statt für Selbstliebe unabhängig von Äußerlichkeiten einzutreten, ist an Bodypositivity ebenso wie an Fitspiration die Unterrepräsentation bestimmter Personengruppen, die Objektifizierung von insbesondere Frauen, sowie der starke Fokus auf ein makelloses äußeres Erscheinungsbild im Zuge einer (äußerlichen) Selbstoptimierung kritisch hervorzuheben.

Die automatische Bildklassifikation gelang nicht für alle betrachteten inhaltlichen Kategorien, weitere Forschung scheint hier notwendig. Besonderes Augenmerk sollte auf den Umgang mit multinominalen Klassifikationsproblemen sowie unbalancierten Klassen gelegt werden. Der erfolgsversprechenden Methodik des Transferlernens mit CNN-Modellen sollte in Zukunft weiter nachgegangen werden, dabei könnte als Grundlage für den Transferlernprozess spezifischer Modelle, die bereits auf konkrete Klassifikationsaufgaben wie das Erkennen von Gesichtern oder Körperteilen hin vortrainiert sind, hilfreich sein. Ein anderer Ansatz könnte sein, neben den Bildern selbst auch weitere Daten andere Modalitäten miteinzubeziehen, um die beiden Hashtags noch umfassender vergleichen zu können. In Frage kämen hier etwa die sprachlichen Informationen aus Beitragstexten, Kommentare anderer User:innen unter dem Beitrag oder Metadaten zur:in Verfasser:in (z.B. Follower:innenzahl).

Fußnoten

1. Wie viel Prozent der Daten aller Klassen der inhaltlichen Kategorie jeweils korrekt klassifiziert wurden. Z.B. für die Kategorie „Beitrag enthält ein Bild“: Von den Beiträgen, die ein Bild zeigen, wurden 97% korrekterweise dieser Klasse zugeordnet (3% wurden fälschlicherweise als kein Bild enthaltend klassifiziert). Von den rein textbasierten Beiträgen wurden 93% korrekterweise dieser Klasse zugeordnet, 7% wurden falsch klassifiziert.
2. Die automatische Klassifikation von Ethnien ist grundsätzlich kritisch zu betrachten: Sie birgt verschiedene Risiken, etwa die Diskriminierung von Personen

oder im medizinischen Bereich das Stellen einer falschen Diagnose. Da im vorliegenden Projekt aber nur Instagram-Bilder ohne die Möglichkeit der Zuordnung zu realen Personen betrachtet wurden und daher auch keinerlei negative Konsequenzen für Personen zu befürchten sind, wurde dennoch eine automatische Klassifikation der Ethnie vorgenommen.

Bibliographie

- Boepple, Leah, Rheanna N. Ata, Ruba Rum und J. Kevin Thompson. 2016. „Strong Is the New Skinny: A Content Analysis of Fitspiration Websites.“ *Body image* 17: 132–135. <https://doi.org/10.1016/j.bodyim.2016.03.001> (Zugegriffen: 16 November 2022).
- Boumaraf, Said, Xiabi Liu, Yuchai Wan, Zhongshu Zheng, Chokri Ferkous, Xiaohong Ma, Zhuo Li und Dalal Bardou. 2021. „Conventional Machine Learning Versus Deep Learning for Magnification Dependent Histopathological Breast Cancer Image Classification: A Comparative Study with Visual Explanation.“ *Diagnostics* 11 (3): 528. <https://doi.org/10.3390/diagnostics11030528> (Zugegriffen: 16 November 2022).
- Chawla, N. V., K. W. Bowyer, L. O. Hall und W. P. Kegelmeyer. 2002. „SMOTE: Synthetic Minority Over-sampling Technique.“ *jair* 16: 321–357. <https://doi.org/10.1613/jair.953> (Zugegriffen: 16 November 2022).
- Cohen, Rachel, Jasmine Fardouly, Toby Newton-John und Amy Slater. 2019a. „#BoPo on Instagram: An experimental investigation of the effects of viewing body positive content on young women's mood and body image.“ *New Media & Society* 21 (7): 1546–1564. <https://doi.org/10.1177/1461444819826530> (Zugegriffen: 16 November 2022).
- Cohen, Rachel, L. Irwin, Toby Newton-John und Amy Slater. 2019b. „# bodypositivity: A content analysis of body positive accounts on Instagram.“ *Body image* 29: 47–57. <https://www.sciencedirect.com/science/article/pii/S1740144518304595> (Zugegriffen: 16 November 2022).
- Cwynar-Horta, Jessica. 2016. „The Commodification of the Body Positive Movement on Instagram.“ *Stream* 8 (2): 36–56. <https://doi.org/10.21810/strm.v8i2.203> (Zugegriffen: 16 November 2022).
- Dubey, Arun Kumar und Vanita Jain. 2020. „Automatic facial recognition using VGG16 based transfer learning model.“ *Journal of Information and Optimization Sciences* 41 (7): 1589–1596. <https://doi.org/10.1080/02522667.2020.1809126> (Zugegriffen: 16 November 2022).
- Fioravanti, Giulia, Andrea Svicher, Giulia Ceragioli, Viola Bruni und Silvia Casale. 2021. „Examining the impact of daily exposure to body-positive and fitspiration Instagram content on young women's mood and body image: An intensive longitudinal study.“ *New Media & Society*: 1–23. <https://doi.org/10.1177/14614448211038904> (Zugegriffen: 16 November 2022).
- Frederick, David A., Elizabeth A. Daniels, Morgan E. Bates und Tracy L. Tylka. 2017. „Exposure to Thin-Ideal Media Affect Most, but Not All, Women: Results from the Perceived Effects of Media Exposure Scale and Open-Ended Responses.“ *Body image* 23: 188–205. <https://doi.org/10.1016/j.bodyim.2017.03.001>

doi.org/10.1016/j.bodyim.2017.10.006 (Zugegriffen: 16 November 2022).

Grabe, Shelly, L. Monique Ward und Janet Shibley Hyde. 2008. „The Role of the Media in Body Image Concerns Among Women: A Meta-Analysis of Experimental and Correlational Studies.“ *Psychological bulletin* 134 (3): 460–476. <https://doi.org/10.1037/0033-2909.134.3.460> (Zugegriffen: 16 November 2022).

Guo, Tianmei, Jiwen Dong, Henjian Li und Yunxing Gao. 2017. „Simple convolutional neural network on image classification.“ In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*: 721–724. (Zugegriffen: 16 November 2022).

Holland, Grace und Marika Tiggemann. 2017. „'Strong Beats Skinny Every Time': Disordered Eating and Compulsive Exercise in Women Who Post Fitspiration on Instagram.“ *The International journal of eating disorders* 50 (1): 76–79. <https://doi.org/10.1002/eat.22559> (Zugegriffen: 16 November 2022).

Lazuka, Rebecca F., Madeline R. Wick, Pamela K. Keel und Jennifer A. Harriger. 2020. „Are We There yet? Progress in Depicting Diverse Images of Beauty in Instagram's Body Positivity Movement.“ *Body image* 34: 85–93. <https://doi.org/10.1016/j.bodyim.2020.05.001> (Zugegriffen: 16 November 2022).

Pilgrim, Katharina und Sabine Bohnet-Joschko. 2022. „Influencer und das Problem mit dem Sixpack.“ *Prävention Gesundheitsforschung* 17 (1): 113–118. <https://doi.org/10.1007/s11553-021-00845-w> (Zugegriffen: 16 November 2022).

Prichard, Ivanka, Eliza Kavanagh, Kate E. Mulgrew, Megan S. C. Lim und Marika Tiggemann. 2020. „The Effect of Instagram #fitspiration Images on Young Women's Mood, Body Image, and Exercise Behaviour.“ *Body image* 33: 1–6. <https://doi.org/10.1016/j.bodyim.2020.02.002> (Zugegriffen: 16 November 2022).

Pulvers, Kim M., Rebecca E. Lee, Harsohena Kaur, Matthew S. Mayo, Marian L. Fitzgibbon, Shawn K. Jeffries, James Butler, Qingjiang Hou und Jasjit S. Ahluwalia. 2004. „Development of a Culturally Relevant Body Image Instrument Among Urban African Americans.“ *Obesity research* 12 (10): 1641–1651. <https://doi.org/10.1038/oby.2004.204> (Zugegriffen: 16 November 2022).

Santarossa, S., P. Coyne, C. Lisinski und S. J. Woodruff. 2019. „#fitspo on Instagram: A mixed-methods approach using Netlytic and photo analysis, uncovering the online discussion and author/image characteristics.“ *Journal of health psychology* 24 (3): 376–385. <https://doi.org/10.1177/1359105316676334> (Zugegriffen: 16 November 2022).

Sheykhoumousa, Mohammadreza, Masoud Mahdianpari, Hamid Ghanbari, Fariba Mohammadimanesh, Pedram Ghamisi und Saeid Homayouni. 2020. „Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review.“ *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 6308–6325. <https://doi.org/10.1109/jstars.2020.3026724> (Zugegriffen: 16 November 2022).

Swasono, Dwiretno Istiyadi, Handayani Tjandrasa und Chastine Fathicah. 2019. „Classification of Tobacco Leaf Pests Using VGG16 Transfer Learning.“ In *2019 12th International Conference on Information & Communication*

Technology and System (ICTS): 176–181. (Zugegriffen: 16 November 2022).

Tiggemann, Marika und Mia Zaccardo. 2015. „'Exercise to Be Fit, Not Skinny': The Effect of Fitspiration Imagery on Women's Body Image.“ *Body image* 15: 61–67. <https://doi.org/10.1016/j.bodyim.2015.06.003> (Zugegriffen: 16 November 2022).

Tiggemann, Marika und Mia Zaccardo. 2018. „'Strong Is the New Skinny': A Content Analysis of #fitspiration Images on Instagram.“ *Journal of health psychology* 23 (8): 1003–1011. <https://doi.org/10.1177/1359105316639436> (Zugegriffen: 16 November 2022).

Skalierungspraktiken in der computergestützten Analyse von literarischen Texten

Krautter, Benjamin

Benjamin.Krautter@uni-koeln.de
Universität zu Köln, Deutschland

Einleitung

In den vergangenen Jahren haben Publikationen aus dem Bereich der digitalen Literaturwissenschaft vermehrt auf das durch den Altphilologen und Anglisten Martin Mueller geprägte Konzept *scalable reading* hingewiesen, es diskutiert oder sogar zum Vorbild für das eigene methodische Vorgehen auserkoren (vgl. exemplarisch Arnold, Fiechter 2022, 162–165; Horstmann, Kleyermann 2019, insb. Kap. 1 und 5; Schruhl 2018, Kap. 5; Weitin 2017, 1–6; Willand, Reiter 2017, 178). Ein Vorzug des „integrative[n] Konzept[s]“ – so wird immer wieder betont – sei dessen Anlage, die eine Verbindung von „qualitativ-hermeneutische[n] und quantitativ-statistische[n] Methoden“ erlaube (Weitin 2015, 2).¹ Dementsprechend ist der von Mueller als „happy synthesis of ‚close‘ and ‚distant‘ reading“ (Mueller 2012, o.S.) angelegte Begriff des *scalable reading* verbreitet als *mixed-methods*-Ansatz wahrgenommen worden (vgl. etwa Herrmann 2018, § 5–7; Kleymann 2022, § 23–25) und wird in der Hauptsache als methodologisches Schlagwort verwendet, das die Verknüpfung qualitativer und quantitativer Methoden anzeigen soll (vgl. etwa Viehhauser 2017, Kap. 4 oder Krause und Pethes 2017, 108). Die unterschiedlichen Skalierungspraktiken, die dem Konzept anhaften, und das häufig betonte brückenbildende Potential von *scalable reading* scheinen mir bislang aber noch nicht ausreichend reflektiert worden zu sein.²

In einem ersten Schritt meines Beitrags werde ich die verschiedenen Dimensionen, auf die sich Muellers Konzeption von *scalable reading* erstreckt, ausdifferen-

zieren und erläutern. Daran anschließend werde ich in einem zweiten Schritt am Beispiel von literarischen Netzwerkanalysen dramatischer Texte exemplarisch darlegen, wie sich etablierte computergestützte Methoden (vgl. Jannidis 2017, 147–161; Trilcke 2013, 201–247) zu diesen Dimensionen verhalten. Um das Konzept des *scalable reading* für die analytische Praxis fruchtbar zu machen, scheint es mir grundlegend, die dafür angeordneten computergestützten Methoden auf ihre Skalierbarkeit hin zu prüfen. Denn während Franco Moretti und Matthew Jockers mit ihren Begriffen *distant reading* und *macroanalysis* die gewohnte literaturwissenschaftliche „Beobachtungshaltung, die ihre Gegenstände auf einer ‚mittleren Skala‘ situier[t]“ (Spoerhase 2020, 7), als ungeeignet für eine umfassende Literaturgeschichtsschreibung kritisieren (vgl. etwa Moretti 2000a, 207–209),³ versucht Mueller mit *scalable reading* Mikro-, Meso- und Makro-Skalen zusammenzudenken.

Die Dimensionen von *scalable reading*

In einem programmatisch ausgerichteten Blogbeitrag hob Mueller 2012 hervor, wie ihn „[t]he charms of Google Earth“ (Mueller 2012, o.S.) zu *scalable reading* als methodischer Metapher für die Betrachtung literarischer Texte geführt habe. Seine Überlegungen konzentrieren sich hauptsächlich auf die Operation des *Zoomens* (vgl. dazu Krautter, Willand 2020, 77–79). Durch Herein- und Herauszoomen würden in Google Earth unterschiedliche Repräsentationsformen entstehen, die je verschiedene Informationen tragen: „different properties of phenomena are revealed by looking at them from different distances“ (Mueller 2012, o.S.).⁴ Ein ähnliches Verfahren imaginiert Mueller nun auch für die Analyse literarischer Texte. Er zielt darauf ab, Verbindungslinien zwischen Einzeltextbetrachtungen und der Untersuchung größerer Zusammenhänge in Textsammlungen ziehen zu können und dabei die aufwändige Aufarbeitung von Kontexten, das Lesen sehr vieler Texte und letztlich die Identifikation von aufschlussreichen Mustern zu beschleunigen und zu vereinfachen.⁵ Grundlage dafür sind die verschiedenen Skalen, die sich hinter Muellers *reading*- bzw. Analysebegriff verbergen. *Scalable reading* erstreckt sich nach meiner Ansicht auf mindestens vier skalierbar gedachte analytische Dimensionen.

Die Skalenpluralität beginnt erstens bei der Textgrundlage: Literarische Texte liegen in „einer weiten ‚Scale‘ von Surrogaten“ (Weitin 2015, 10) vor, die nebeneinander koexistieren: „Our typical encounter with a text is through a surrogate“ (Mueller 2013, o.S.). Mueller spricht an dieser Stelle von Surrogaten, da immer schon mit unterschiedlich gearteten Repräsentationen des Originals gearbeitet wurde und wird: Das können beispielsweise Faksimiles, Text- und Werkausgaben, Digitalisate oder auch speziell kodierte Textsammlungen sein (vgl. dazu Mueller 2014, § 4–20). Surrogate können darüber hinaus in stark transformierter oder abstrahierter Form auftreten, beispielsweise in Gestalt von Häufigkeitswortlisten. Auch die Netzwerkanalyse fußt demnach auf Surrogaten. Peer Trilcke und Frank Fischer sprechen von einem „Zwischenfor-

mat“, das in ihrem Fall nur noch diejenigen Strukturinformationen der Dramen vorhalte, die zur Netzwerkerstellung herangezogen werden (Trilcke, Fischer 2018, Kap. 3). Der Dramentext selbst ist nicht mehr Teil des Zwischenformats.

An diese unterschiedlichen Repräsentationsformen von Literatur ist zweitens die Frage des Umfangs geknüpft: Wie groß ist der Untersuchungsgegenstand? Handelt es sich nur um einen einzelnen Text, vielleicht sogar nur um einen Ausschnitt des Textes, oder aber um eine größere Sammlung von Texten? Wie umfangreich ist diese Sammlung? Nicht nur die Zahl der zu betrachtenden Texte, auch die Textsorte kann hier Teil der Skalierungsfrage sein: Sollen kurze Novellen oder 1000-seitige Langromane untersucht werden, ein kurzer Einakter oder Karl Kraus' monumentales Lesedrama *Die letzten Tage der Menschheit*, in dem in 220 Szenen fast 1000 sprechende Figuren auftreten (vgl. Fischer u.a. 2020, 279).

Drittens stellt sich die Frage nach der Größe der Analyseeinheiten. Morettis *distant reading* grenzt sich, wie Carlos Spoerhase herausgearbeitet hat, von der üblichen Meso-Skala literaturwissenschaftlicher Untersuchungen ab, bei der das Verständnis eines oder einiger weniger literarischer Texte im Fokus stehe (vgl. Spoerhase 2020, 7). Für Moretti ist dagegen alles interessant, was abseits dieser mittleren Skala liegt, das sind „units that are much smaller or much larger than the text“ (Moretti 2000b, 57). Moretti geht es also nicht mehr um die ganzheitliche Interpretation von Texten, sondern um Mikro- und Makro-Eigenschaften von Textsammlungen, wie die Verteilung einzelner Wortformen oder die diachrone Entwicklung von Gattungen. Anders als *close reading*, das an die Meso-Skala gebunden sei, würden *distant reading* oder *macroanalysis* hinsichtlich der Analyseeinheiten sowohl ein „zooming in“ als auch ein „zooming out“ ermöglichen (Jockers 2013, 23). Auch Mueller betont, dass quantitative Methoden gleichermaßen ein Heraus- wie ein Hereinzoomen erlauben würden. Die Metapher des Zoomens ist bei ihm aber nicht an ein bestimmtes methodisches Instrumentarium gebunden (vgl. Mueller 2014, § 31).

Mueller denkt die methodische Bezugsgröße viertens vielmehr selbst auf einer Art Skala. Wie Weitin gemeinsam mit Thomas Gilli und Nico Kunkel (2016, 115) herausstellt, umfasse *scalable reading* bei Mueller nämlich „prinzipiell alle Akte des Lesens und Analysierens von Texten“. Unterschiedliche qualitative und quantitative Methoden würden dann gleichberechtigt nebeneinanderstehen und könnten den analytischen Anforderungen der Fragestellung und der gewählten Textsammlung gemäß kombinatorisch zusammengedacht werden. Anders als es das Begriffspaar *close* und *distant reading* nahelegt, ist der Einsatzzweck verschiedener Formen der Analyse bei Mueller nicht im Vorhinein determiniert. Relevant ist für ihn stattdessen, wie sich qualitative und quantitative Methoden für eine bestimmte Fragestellung so kombinieren lassen, dass ein analytischer Mehrwert entsteht.

Praktische Überlegungen zum scalable reading

Im folgenden Abschnitt möchte ich die mit einer Praxis des *scalable reading* verbundenen Herausforderungen genauer beleuchten. Zur Veranschaulichung greife ich dabei auf literarische Netzwerkanalysen zurück. Abbildung 1 zeigt ein Kopräsenznetzwerk von Friedrich Schillers *Die Räuber* (1781). Jeder Knoten im Netzwerk repräsentiert eine Figur des Dramas, die Kanten zwischen zwei Knoten zeigen an, dass die beiden verbundenen Figuren innerhalb eines bestimmten Textsegments interagieren. Im vorliegenden Fall bedeutet Interaktion, dass die beiden Figuren in der gleichen Szene sprechen (vgl. Trilcke u.a. 2015, 1). Solche Netzwerke können automatisiert erstellt werden, wenn die digitalisierten Dramen entsprechend kodiert vorliegen, wie es etwa beim *Drama Corpora Project* der Fall ist (siehe Fischer u.a. 2019).⁶ Dadurch lässt sich eine große Zahl an Netzwerken nicht nur visuell, sondern vor allem mit Blick auf mathematische Netzwerkmetriken vergleichen.

Die Automatisierung führt jedoch zu einigen Einschränkungen. So ist die oben dargelegte Formalisierung von Figureninteraktionen zwar ähnlich, aber nicht deckungsgleich mit dem von Solomon Marcus (1973, 358) vorgeschlagenen und zum kodifizierten Handbuchwissen (vgl. etwa Pfister 2011, 235–240) gewordenen Begriff der Konfiguration. Die Figurenkonfiguration eines Dramas ändert sich immer dann, wenn eine Figur die Bühne betritt oder verlässt, also das am Bühnengeschehen beteiligte Personal zumindest in Teilen wechselt. Dramen, die Prinzipien des französischen Klassizismus folgen, sind durch die im Nebentext markierten Auf- und Abtritte strukturiert. Konfiguration und Szenengrenze fallen dann – zumindest in der Theorie – zusammen. Anders ist das bei Stücken, die sich an Shakespeares Poetik orientieren. Hier sind die Szenengrenzen zumeist an einen Ortswechsel gebunden. Daher können Figuren auf- oder abtreten, ohne dass zwangsläufig eine neue Szene konstituiert wird. Da Auf- und Abtritte von Figuren im *Drama Corpora Project* (noch) nicht kodiert sind, ist die automatisierte Erstellung von Kopräsenznetzwerken auf die Szenengrenzen als Segmentierung angewiesen.⁷ Das kann Begleiterscheinungen zur Folge haben, die es in der Untersuchung zu reflektieren gilt. Schillers Stück *Die Räuber* ist dafür ein gutes Beispiel. So hat schon Marcus (1973, 326–333) darauf hingewiesen, dass sich zwischen der Anzahl an Szenen und der Anzahl an Konfigurationen eine große Diskrepanz auftut. Im Verlauf der 15 Szenen von *Die Räuber* zählt Marcus ganze 78 Konfigurationen.

Nun stellt sich die Frage, wie sich solche Kopräsenznetzwerke sinnvollerweise in die etablierte Dramenanalyse integrieren lassen. Einen prominenten Versuch, Netzwerkanalysen in den Verstehensprozess literarischer Texte zu integrieren, unternimmt Moretti in seinem Essay *Network Theory, Plot Analysis* (2011). Entgegen seiner Rhetorik der *large scale* erprobt er die Methode nur an einzelnen literarischen Texten, insbesondere an Shakespeares *Hamlet* (1609). Sein Vorgehen lässt sich als *close reading* von Netzwerkvisualisierungen beschreiben. Moretti zerlegt die Visualisierungen in verschiedene Teile und versucht seine netzwerkanalytischen Beobach-

tungen anschließend mit generellen textanalytischen Erkenntnissen zu verbinden (vgl. Moretti 2011, 3–7).

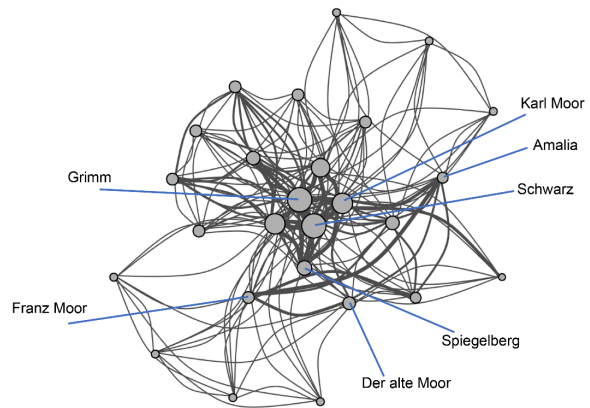


Abbildung 1: Kopräsenznetzwerk von Friedrich Schillers *Die Räuber* (GEM force directed layout algorithm). Die Knotengröße repräsentiert den Grad.

Wie verhält sich Morettis Studie aber zu Muellers *scalable reading* und wie lässt sie sich hinsichtlich der vier von mir herausgearbeiteten Dimensionen (Textgrundlage, Größe des Untersuchungsgegenstandes, Analyseeinheiten und Methoden) einordnen? Verglichen mit einer herkömmlichen Interpretation von *Hamlet* ist die auffälligste Veränderung mit Sicherheit die fundamentale Modifizierung der Textgrundlage. Moretti interpretiert nicht den Dramentext, sondern ein auf Strukturdaten basierendes abstraktes Netzwerk in Form von Knoten und Kanten. Die Größe des Untersuchungsgegenstandes bleibt zwar auf einen literarischen Text beschränkt, das spezifische Surrogat beziehungsweise Textmodell gibt aber die im Zentrum der Untersuchung stehenden Analyseeinheiten vor: Untersucht wird die zu einem Figurennetzwerk subsumierte Kopräsenz von Dramenfiguren. Aufschlussreich ist Morettis methodisches Vorgehen. Grundlage des Netzwerks sind quantitative Strukturdaten. Erkenntnisse gewinnt er aber vor allem im Modus der Interpretation und nur äußerst begrenzt aus statistischen Auswertungen der mathematischen Netzwerkmetriken. Moretti beschreibt das als „using networks to gain intuitive knowledge of plot structures“ (Moretti 2011, 12). Entspricht das nun Muellers Vorstellung von *scalable reading*? Insbesondere Morettis Schlussfolgerung ist dahingehend bezeichnend. Seine Analysen, so urteilt er, würden nämlich nach einer „radical reconceptualization of characters and their hierarchy“ in der Literaturwissenschaft verlangen (Moretti 2011, 5). Etablierte Konzepte – etwa das Protagonistenkonzept – ließen sich aus seiner Perspektive kaum produktiv mit seinen netzwerkanalytischen Ergebnissen verbinden. Denn die explanative Funktion abstrakter Modelle sei mit „concepts of ‚consciousness‘ and ‚interiority‘“ nicht kompatibel (Moretti 2011, 4). Für Moretti ist es demnach nicht einmal dann erstrebenswert, die Netzwerkanalyse in die typische Meso-Skala literaturwissenschaftlicher Interpretationen zu integrieren, wenn nur ein einzelner Text untersucht wird. Denn schon hier zielt er auf textübergreifende Konzepte ab. Beim Zoomen zwi-

schen Mikro- und Makro-Ebene, so ließe sich schließen, wird die Meso-Skala übersprungen.

Abbildung 1 verdeutlicht zudem, dass nicht alle Kopräsenznetzwerke vom „intermediate‘ status of visualization“ profitieren, den Moretti in seinem Essay als so wichtig erachtet (Moretti 2011, 11). Das Netzwerk von Schillers *Die Räuber* kann nämlich auch in die Irre führen. Die getrennten Sphären von Familie und Räubern, die die Struktur des Stücks in den ersten drei Akten stark prägen (vgl. Krautter, Willand 2021, 115–118), lassen sich in der Abbildung nicht identifizieren. Ganz im Gegenteil: Die Netzwerkvisualisierung stellt die Räuberbande, insbesondere Schwarz, Grimm und Karl Moor ins Zentrum des Dramas, während die Familie um Franz, den alten Moor und Amalia an den Rand rückt. Grund dafür ist der Grad der entsprechenden Knoten. Dabei handelt es sich um eine simple Metrik, die die Anzahl an Kanten misst, die ein Knoten auf sich vereint (vgl. Newman 2010, 168–169). Während Schwarz und Grimm alle 25 möglichen Verbindungen zu anderen Figuren realisieren, sind es bei Franz nur deren zwölf. Selbst Moretti würde von Schwarz’ und Grimms Zentralität aber kaum darauf schließen, dass sie die Hauptfiguren des Stücks sind. Zu marginal ist ihr Einfluss auf die Handlung, zu gering sind ihre Redean-teile.

Der Mehrwert solcher Figurennetzwerke wird jedoch meist auf die Analyse größerer Textsammlungen verschoben, die schon aufgrund ihrer Menge nur schwer durch *close reading* erschließbar erscheinen (vgl. Trilcke, Fischer 2018, Kap. 3). Hinsichtlich der vier Dimensionen würde das einer Hörskalierung des Untersuchungsgegenstands entsprechen. Ziel ist es hierbei, die als Netzwerke modellierten literarischen Texte durch mathematische Metriken – wie den Grad der Knoten – in einen Vergleichszusammenhang zu stellen. Davon sind auch die Analyseeinheiten betroffen. Die Kopräsenzen der Figuren sind zwar weiterhin die Grundlage der Analyse, als infratextuelle Merkmale sollen sich durch sie aber supratextuelle Muster identifizieren lassen. Hierdurch könnten sich, so die Hoffnung, entweder neue Einsichten in die Literaturgeschichte ergeben oder existierende Hypothesen anhand umfassender Textsammlungen nachvollzogen werden.

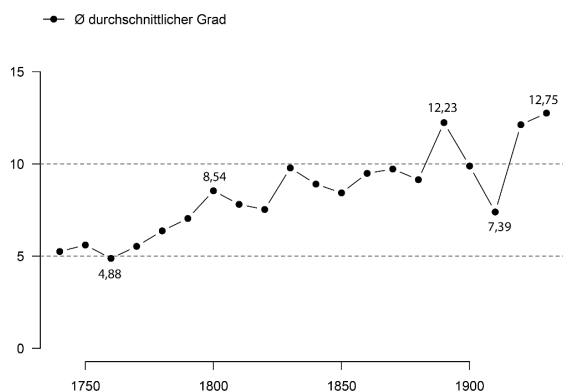


Abbildung 2: Durchschnittlicher Grad von 583 deutschsprachigen Dramen. Die Abbildung zeigt die Mittelwerte pro Dekade.

Abbildung 2 zeigt ein Beispiel für eine diachrone Analyse anhand 583 deutschsprachiger Dramen, die zwischen 1730 und 1930 veröffentlicht oder uraufgeführt wurden (German Drama Corpus). Die Abbildung reproduziert eine Untersuchung von Trilcke und Fischer (2018, Abbildung 6). Wie Trilcke und Fischer habe ich aus dem durchschnittlichen Grad der einzelnen Dramen die Mittelwerte für jedes Jahrzehnt von 1730 bis 1930 ermittelt. Rein deskriptiv ist festzuhalten, dass der durchschnittliche Grad ab dem späten 18. Jahrhundert langsam ansteigt. Zwischen 1830 und 1880 sind dann nur relativ geringe Schwankungen zu erkennen, ehe auf einen Anstieg bis etwa 1890 ein abrupter Fall und ein erneuter starker Anstieg folgen. Trilcke und Fischer haben diese Werte als Indikator dafür gedeutet, dass Dramatiker:innen mit ihren Stücken „auf die gesellschaftliche Modernisierung und Ausdifferenzierung seit der zweiten Hälfte des 18. Jahrhunderts“ reagieren. Sie weisen im Anschluss gleichwohl darauf hin, dass diese Erkenntnis nichts Neues sei (Trilcke, Fischer 2018, Kap. 4.1). Mit Fotis Jannidis (2019, 65) gesprochen lässt sich diese Art der Wissenskonsolidierung als Form der „Kreuzpeilung“ begreifen.

Wie überzeugend ist diese Deutung der Werte aber? Vergleicht man den Werteverlauf des durchschnittlichen Grads in Abbildung 2 mit der Zahl der auftretenden Figuren aus Abbildung 3, scheint ein Zusammenhang zu bestehen. Die Berechnung der Korrelation zwischen Grad und Figurenzahl bestätigt diese Relation: Spearman’s ρ beträgt 0,75. Aus konzeptioneller Perspektive erscheint der Zusammenhang schlüssig. Je größer die Zahl der auftretenden Figuren, umso höher ist der maximal mögliche Grad einer Figur. Gleichzeitig sinkt mit steigender Figurenzahl die Wahrscheinlichkeit, dass alle möglichen Kanten tatsächlich realisiert werden. Folglich sinkt die Dichte des Netzwerks. Die diachrone Analyse des durchschnittlichen Grads beruht also zu großen Teilen auf der sich verändernden Anzahl auftretender Figuren in den Dramen. Ob und wie die Interaktion der Figuren davon beeinflusst wird, müsste jedoch genauer untersucht werden. Dabei könnten Netzwerkmetriken, die die Zentralität eines Knotens möglichst unabhängig von der Netzwerkgröße ermitteln, künftig eine wichtige Rolle einnehmen (vgl. Szemes, Vida 2023 [in Vorb.]).

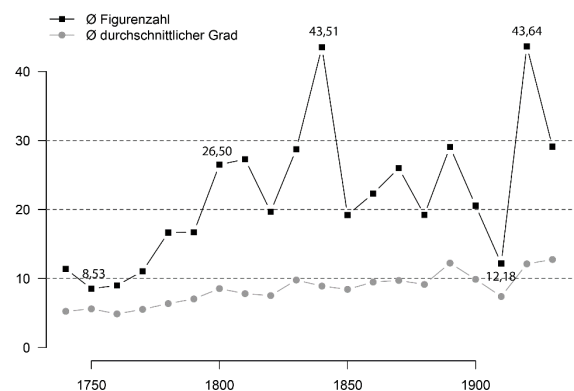


Abbildung 3: Zahl der Figuren (schwarz) und durchschnittlicher Grad (grau) von 583 deutschsprachigen Dramen. Die Abbildung zeigt die Mittelwerte pro Dekade.

Fazit

Die vier von mir beschriebenen Dimensionen des *scalable reading* können helfen, Möglichkeiten und Grenzen der Skalierung mit Bezug auf spezifische Forschungsfragen oder das methodische Vorgehen auszuloten. Wie meine Ausführungen gezeigt haben, sind die Textgrundlage, die Größe des Untersuchungsgegenstandes, die Analyseeinheiten und die eingesetzten Methoden nicht als voneinander unabhängige Variablen zu denken. So ist es einerseits nicht selbstverständlich, dass Kopräsenznetzwerke für die typische literaturwissenschaftliche Meso-Skala produktiv gemacht werden können. Bei der Höherkalierung des Untersuchungsgegenstandes ist es andererseits wichtig, die Werte der Netzwerkmetriken zu kontextualisieren und zu reflektieren, da die Analyseeinheiten als stark abstrahierte Datenwerte repräsentiert werden. Die immer wieder als zentral erachtete Verbindung von Einzeltextlektüren mit der Analyse größerer Textsammlungen (vgl. Weitin 2017, 1) steht somit nicht nur vor der Herausforderung, qualitative und quantitative Methoden zu kombinieren. In Abhängigkeit davon müssten zeitgleich unterschiedliche Textgrundlagen, unterschiedlich skalierte Untersuchungsgegenstände und unterschiedliche Analyseeinheiten gewinnbringend zusammengeführt werden.

Fußnoten

1. Weniger optimistisch zeigt sich Weitin in seiner kürzlich erschienenen Monografie *Digitale Literaturgeschichte*. Dort konstatiert er, dass sich Muellers Metapher des skalierbaren Lesens als „hauptsächlich irreführend“ herauskristallisiert habe. Er differenziert *scalable reading* deshalb zu *reading at scale*, das die einmalige Skalenjustierung vor der Analyse fokussiert (Weitin 2021, 116).
2. Friederike Schruhl kritisiert mit Blick auf die Anwendungspraxis, dass meist unklar bleibe, „was eigentlich in einem [...] *scalable reading* ganz konkret getan werden müsste.“ (Schruhl 2018, Kap. 5).
3. In einem vor kurzem publizierten Aufsatz bezweifelt Moretti ganz grundsätzlich, dass sich qualitative und quantitative Methoden hinsichtlich derselben Fragestellung überhaupt sinnvoll synthetisieren ließen (vgl. Moretti 2020, S. 133–136).
4. Die Analogie zu *Google Earth* und die Metaphorik des Zoomens wurden durchaus kritisch betrachtet (vgl. Weitin 2017, 2). Marcus Willand (2017, 84) problematisiert etwa, dass Mueller eine „graduelle Skalierbarkeit [...] des menschlichen Lesens und [...] des computergestützten Analysierens“ impliziere.
5. Mueller veranschaulicht dies beispielhaft an Bibelkonkordanzen (vgl. Mueller 2012 und 2013).
6. <https://dracor.org/>.
7. Es sei denn, man trainiert selbst ein Modell, das die Auf- und Abtritte automatisch identifiziert.

Bibliographie

- Arnold, Frederik und Benjamin Fiechter. 2022. „Lesen, was wirklich wichtig ist. Die Identifikation von Schlüsselstellen durch ein neues Instrument zur Zitationsanalyse.“ In *DHd 2022. Konferenzabstracts*: 162–166. DOI: 10.5281/zenodo.6304590.
- Fischer, Frank, Anna Busch, Angelika Hecht, Peer Trilcke und Andreas Vogel. 2020. „Besuch im ‚Mars-theater‘ – Eine Netzwerkmodellierung von Karl Kraus’ Riesendrama ‚Die letzten Tage der Menschheit‘.“ In *DHd 2020. Konferenzabstracts*: 278–280. DOI: 10.5281/zenodo.3666690.
- Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hecht, Christopher Kittel, Carsten Milling und Peer Trilcke. 2019. „Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor.“ In *DHd 2019. Konferenzabstracts*: 194–197. DOI: 10.5281/zenodo.2596095.
- Herrmann, J. Berenike. 2018. „In a Test Bed with Kafka. Introducing a Mixed-Method Approach to Digital Stylistics.“ In *Digital Humanities Quarterly* 11/4. <http://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html> (zugegriffen: 15.12.2022).
- Horstmann, Jan und Rabea Kleymann. 2019. „Alte Fragen, neue Methoden – Philologische und digitale Verfahren im Dialog. Ein Beitrag zum Forschungsdiskurs um Entsagung und Ironie bei Goethe.“ In *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2019_007.
- Jannidis, Fotis. 2017. „Netzwerke.“ In *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein, 147–161. Stuttgart: J.B. Metzler.
- Jannidis, Fotis. 2019. „Digitale Geisteswissenschaften: Offene Fragen – schöne Aussichten.“ In *Zeitschrift für Medien- und Kulturforschung* 10/1: 63–70.
- Kleymann, Rabea. 2022. „Datendiffraktion. Von Mixed zu Entangled Methods in den Digital Humanities.“ In *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/sb005_008.
- Krause, Marcus und Nicolas Pethes. 2017. „Scholars in Action. Zur Autoreferentialität philologischen Wissens im Wandel medialer Praktiken.“ In *Deutsche Vierteljahrschrift für Literaturwissenschaft und Geistesgeschichte* 91/1: 73–108.
- Krautter, Benjamin und Marcus Willand. 2020. „Close, Distant, Scalable. Skalierende Textpraktiken in der Literaturwissenschaft und den Digital Humanities.“ In *Ästhetik der Skalierung*, hg. von Carlos Spoerhase, Steffen Siegel und Nikolaus Wegmann, 77–97. Hamburg: Felix Meiner Verlag.
- Krautter, Benjamin und Marcus Willand. 2021. „Vermessene Figuren. Karl und Franz Moor im quantitativen Vergleich.“ In *Schillers Feste der Rhetorik*, hg. von Peter-André Alt und Stefanie Hundehege, 107–138. Berlin, Boston: De Gruyter.
- Marcus, Solomon. 1973 [1970]. „Mathematische Poetik.“ Übers. von Edith Mândroiu. Frankfurt a. M.: Athenäum.
- Moretti, Franco. 2000a. „The Slaughterhouse of Literature.“ In *Modern Language Quarterly* 61/1: 207–227.
- Moretti, Franco. 2000b. „Conjectures on World Literature.“ In *New Left Review* 1: 54–68.

Moretti, Franco. 2011. „Network Theory, Plot Analysis.“ In *Pamphlets of the Stanford Literary Lab 2*: 1–12. <https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf> (zugegriffen: 15.12.2022).

Moretti, Franco. 2020. „The Roads to Rome. Literary Studies, Hermeneutics, Quantification.“ In *New Left Review* 124 (2020): 125–136.

Mueller, Martin. 2012. *Scalable Reading*. <https://web.archive.org/web/20211201185120/sites.northwestern.edu/scalablereading/2020/04/26/scalable-reading/> (zugegriffen: 15.12.2022).

Mueller, Martin. 2013 [2008]. *Morgenstern's Spectacles or the Importance of Not-Reading*. <https://web.archive.org/web/20220120104041/scablereading.northwestern.edu/2013/01/21/morgensterns-spectacles-or-the-importance-of-not-reading/> (zugegriffen: 15.12.2022).

Mueller, Martin. 2014. „Shakespeare His Contemporaries. Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment.“ In *Digital Humanities Quarterly* 8/3. <http://digitalhumanities.org:8081/dhq/vol/8/3/000183/000183.html> (zugegriffen: 15.12.2022).

Newman, Mark E. J. 2010. *Networks. An Introduction*. Oxford, New York: Oxford University Press.

Trilcke, Peer. 2013. „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft.“ In *Empirie in der Literaturwissenschaft*, hg. von Philip Ajouri, Katja Mellmann und Christoph Rauen, 201–247. Münster: Mentis.

Trilcke, Peer und Frank Fischer. 2018. „Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen.“ In *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/sb03.

Trilcke, Peer, Frank Fischer und Dario Kampkaspar. 2015. „Digital Network Analysis of Dramatic Texts.“ In *Book of Abstracts DH 2015*. DOI: 10.5281/zenodo.3627711.

Schruhl, Friederike. 2018. „Objektumgangsnormen in der Literaturwissenschaft.“ In *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/sb003_012.

Spoerhase, Carlos. 2020. „Skalierung. Ein ästhetischer Grundbegriff der Gegenwart.“ In *Ästhetik der Skalierung*, hg. von Carlos Spoerhase, Steffen Siegel und Nikolaus Wegmann, 5–15. Hamburg: Felix Meiner Verlag.

Szemes, Botond und Bence Vida. 2023 (in Vorb.). „Tragic and Comical Networks. Clustering Dramatic Genres According to Structural Properties.“ In *Computational Drama Analysis. Reflecting Methods and Interpretations*, hg. von Melanie Andresen und Nils Reiter. Berlin, Boston: De Gruyter.

Viehhauser, Gabriel. 2017. „Digitale Gattungsgeschichten. Minnesang zwischen generischer Konstanz und Wende.“ In *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2017_003.

Weitin, Thomas. 2015. „Thinking slowly. Literatur lesen unter dem Eindruck von Big Data.“ In *LitLab Pamphlet 1*: 1–17. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2019/06/p01_weitin-thinking-slowly_de.pdf (zugegriffen: 15.12.2022).

Weitin, Thomas. 2017. „Scalable Reading.“ In *Zeitschrift für Literaturwissenschaft und Linguistik* 47/1: 1–6.

Weitin, Thomas. 2021. *Digitale Literaturgeschichte. Eine Versuchsreihe mit sieben Experimenten*. Berlin, Heidelberg: J.B. Metzler.

Willand, Marcus. 2017. „Hermeneutische Interpretation und digitale Analyse: Versuch einer Verhältnisbestimmung.“ In *Lektüren. Positionen zeitgenössischer Philologie*, hg. von Luisa Banki und Michael Scheffel, 77–98. Trier: Wissenschaftlicher Verlag Trier.

Willand, Marcus und Nils Reiter. 2017. „Geschlecht und Gattung. Digitale Analysen von Kleists ‚Familie Schrockenstein‘.“ In *Kleist-Jahrbuch: 177–195*.

Synoptische Interfaces Digitaler Editionen

Herbst, Yannik

yannik.herbst@stud-mail.uni-wuerzburg.de
Zentrum für Philologie und Digitalität (ZPD), Universität
Würzburg, Deutschland

Roeder, Torsten

torsten.roeder@uni-wuerzburg.de
Zentrum für Philologie und Digitalität (ZPD), Universität
Würzburg, Deutschland

Reul, Christian

christian.reul@uni-wuerzburg.de
Zentrum für Philologie und Digitalität (ZPD), Universität
Würzburg, Deutschland

Das Prinzip Synopse

Digitale Editionen stehen häufig vor der Herausforderung, verschiedene Zustände eines Dokuments oder auch Varianten eines Textes in einer gegenüberstehenden Ansicht, also synchron anzeigen zu wollen. Das erstere Szenario ist dokumentenbezogen und kann beispielsweise dem Vergleich von Digitalisat, diplomatischer Transkription, Lesetext, XML-Text oder Plaintext dienen (Abb. 1), während das andere Szenario textbezogen ist und der Vermittlung von unterschiedlichen Überlieferungszuständen oder textgenetischen Verwandtschaften dient (Abb. 2). Dem jeweiligen Dokument oder Text können ferner Kommentare, Annotationen, Apparate und Metadaten anbeigestellt werden.

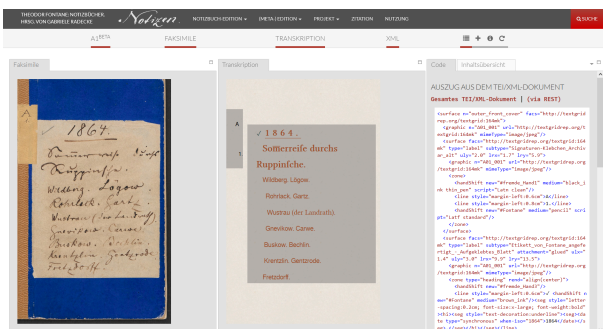


Abb. 1: Dokumentbezogene synoptische Darstellung. Exemplarischer Screenshot: Theodor Fontane, "Notizbuch A1". *Theodor Fontane: Notizbücher. Digitale genetisch-kritische und kommentierte Edition.* https://fontane-nb.dariah.eu/edition.html?id=/xml/data/1zz-dk.xml&page=outer_front_cover (zugegriffen: 14. Dezember 2022).



Abb. 2: Textbezogene synoptische Darstellung mit flexiblen Spaltenmanagement. Exemplarischer Screenshot: Märchen "Rumpelstilzchen" (nach den *Kinder- und Hausmärchen* der Gebrüder Grimm). *LERA – Locate, Explore, Retrace and Apprehend complex text variants.* <https://lera.uzi.uni-halle.de/editions/3/?lang=de> (zugegriffen: 14. Dezember 2022).

Interfaces mit mehrspaltigem Layout

Interfaces digitaler Editionen adressieren diese Herausforderungen fast immer mithilfe eines mehrspaltigen Layouts. Die sogenannte Synopse (abgeleitet von griech. *syn-optikós* "das Ganze zusammensehend") ist eine klassische Darstellungsform, bei der zueinander in Beziehung stehende Dokumente oder Texte in zwei oder mehr Spalten nebeneinander angezeigt werden (vgl. das Konzept der "Composite Design Patterns", dazu z. B. Javed/Elmqvist 2012). Die Beziehung zwischen den gegenübergestellten Ressourcen drückt sich in dem jeweiligen Punkt der Synchronisierung aus: Diese kann beispielsweise an Seiten, Absätzen oder auch an einzelnen Zeilen vorgenommen werden. Das Layout erzeugt aus diesen Synchronisations-Punkten die 'ersichtliche' Verbindung zwischen den abgebildeten Ressourcen (zum Prinzip der "Juxtaposition" und die Kombination mit anderen Abbildungsmethoden vgl. Gleicher et al. 2011). Hingegen ist in den Daten diese visualisierte Verbindung nicht immer explizit abgelegt, sondern kann auch in einem 'soften' Konstrukt bestehen, beispielsweise indem regulär aufgebaute IDs und konsequente Nummerierung verwendet werden, um Digitalisat und XML-Dateien abzugleichen. Denkbar ist zudem, dass zwei nebeneinanderstehende Texte aus ein und demselben Dokument generiert werden, wie etwa eine diplomatische Umschrift mit einem aufgeschlüsselten Lesetext daneben, oder auch zwei

Textfassungen auf der Grundlage eines textkritischen In-line-Apparats.

Divergenz der Ansätze

Synoptische Ansichten werden aktuell fast ausschließlich individuell für die jeweiligen Editionsprojekte konzipiert und programmiert. Eine Ursache dafür ist darin zu sehen, dass die jeweils gegenüberzustellenden Inhalte von Projekt zu Projekt extrem unterschiedlich ausfallen können, sowohl in der Art ihrer Beziehung zueinander, als auch in der eigenen Ausprägung. Dokumentenzentrierte Ansätze wie z. B. im *Deutschen Textarchiv* (Abb. 3) und in *DiScholEd* (Abb. 4) stellen einzelnen Digitalisaten verschiedene Derivate der dazugehörigen XML-Codierung gegenüber.

Keller, Gottfried: Der grüne Heinrich. Bd. 1. Braunschweig, 1854.



Abb. 3: Dokumentenzentrierter Ansatz. Exemplarischer Screenshot: Gottfried Keller, "Der grüne Heinrich", *Deutsches Textarchiv*. https://www.deutschestextarchiv.de/book/view/keller_heinrich01_1854?p=9 (zugegriffen: 14. Dezember 2022).

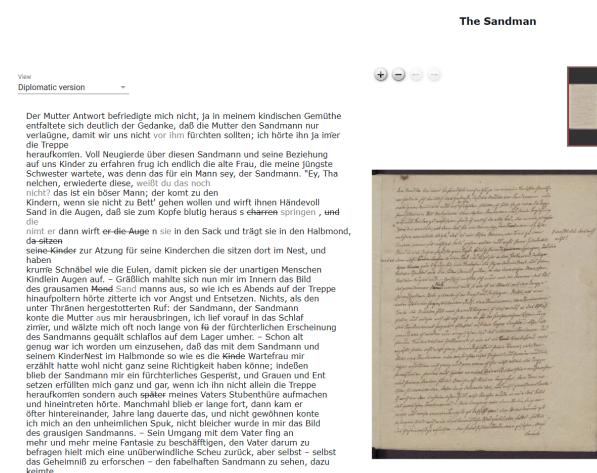


Abb. 4: Dokumentenzentrierter Ansatz mit nutzerseitiger Auswahl der Spalteninhalte. Exemplarischer Screenshot: E. T. A. Hoffmann, "Der Sandmann", hg. von Anna Busch, *Briefe und Texte aus dem intellektuellen Berlin um 1800*, hg. von Anne Baillet. <https://discholed.huma-num.fr/exist/apps/discholed/bi/corpus/Sandmann.xml?root=3.4.2.2.24> (zugegriffen: 14. Dezember 2022).

Projekte mit Fokus auf Textgenese stellen indessen mehrere Textdokumente gegenüber und setzen diese in der Regel nach Strukturabschnitten (Kapitel, Absätze, Zeilen) in Beziehung – was einen völlig anderen Präsentationsansatz darstellt. In diesem Fall gehen die Ansätze zudem in divergente Richtungen, deren Pole einerseits durch eine platzintensive Rasterdarstellung wie z. B. in LERA (Abb. 2) und andererseits durch ein auf eine Spalte (!) verdichtetes Konzept wie z. B. in der "spinal edition" des *Frankenstein Variorums* (Abb. 5) gebildet werden. Es besteht somit der Bedarf an einem allgemeinen Modell synoptischer Präsentationskonzepte, welches die Grundlage eines generischen Interface-Frameworks für Digitale Editionen bilden könnte.



Abb. 5: Verdichtung der Synopse auf eine einzelne Spalte. Exemplarischer Screenshot: *Frankenstein Variorum*, hg. von Elisa Beshero-Bondar, Rikk Mulligan und Raffaele Viglianti. <https://frankensteinvariorum.github.io/viewer/viewer> (zugegriffen: 14. Dezember 2022).

Des Weiteren erlauben bestehende Ansätze bereits die nutzerseitige Auswahl der Spalteninhalte wie bei DTA (Abb. 3) und *DiSchoEd* (Abb. 4), teils sogar mit der zusätzlichen Möglichkeit des Spaltenmanagements wie bei den *Fontane Notizbüchern* (Abb. 1) und im generischen *EVT Viewer* (Abb. 6). Dies ist insbesondere mit steigender Zahl der Spalten notwendig (vgl. LERA, Abb. 2).



Abb. 6: Synoptische Darstellung mit flexiblen Spaltenmanagement, hier im EVT Viewer [Abk. für Edition Visualization Technology]. Exemplarischer Screenshot: *Codice Pelavicino*, hg. von Enrica Salvatori und Edilio Riccardini. <http://pelavicino.labcd.unipi.it/evt> (zugegriffen: 14. Dezember 2022).

Als nicht nur technische, sondern vor allem auch konzeptionelle Herausforderung kommt noch Responsivität hinzu. Das Nutzungskonzept der Edition muss sich an unterschiedliche Endgeräte – und damit an die mögliche Zahl der sichtbaren (!) Spalten – anpassen können. Letzteres hat einen unmittelbaren Einfluss auf die Zugänglichkeit und Nutzbarkeit der Ressourcen. Dies hat außerdem zur Folge, dass die oben genannten divergierenden Präsentationskonzepte für die Gegenüberstellung verschiedener Textvarianten behandelt werden müssen, was schließlich auf ein Konzept hinausläuft, welches grundsätzlich von einer variablen Spaltenzahl ausgeht.

Einen weiteren Aspekt bringt das Projekt TAPAS (Abb. 7) ein, welches dem Nutzer die Auswahl der dem Layout zugrundeliegenden XML-Renderingskripte erlaubt, die zudem jeweils responsiv agieren. Dies verdeutlicht nicht nur, dass die Grundlage für anpassbare Lösungen in der Trennung zwischen Daten- und Präsentationsschicht besteht, sondern gerade auch, dass dieser Umstand prinzipiell für Interfaces nutzbar ist, sofern die spezifische Abbildungslogik einmal formalisiert wurde. Der hier aufgezeigte Weg ist z. B. mithilfe von XSLT und ggf. ODD vollständig innerhalb des X-Technologiestacks umsetzbar (näheres bei Flanders/Hamlin 2013).



Abb. 7: Synoptische Präsentation von TAPAS. Exemplarischer Screenshot: François-Joseph Bérardier de Bataut, "Sixième entretien. Narration historique", *Essai sur le récit*, hg. von Christof Schöch. TAPAS. <http://www.tapasproject.org/berardier/files/sixi%C3%A8me-entretien-narration-historique> (zugegriffen am 14. Dezember 2022).

Die Software "Synopticon"

Das Software-Entwicklungsprojekt *Synopticon* (vgl. Herbst 2022, auf GitHub verfügbar) verfolgt das Ziel, die typischen Anforderungen spaltenorientierter Interfaces digitaler Editionen in einer generischen Lösung zusammenzufassen, so dass es perspektivisch in bestehende Webseiten digitaler Editionen mit geringem Aufwand eingebunden werden kann. Auch die Einbindung in Editions-Frameworks wird verfolgt: So ist *Synopticon* bereits mit der modularen Architektur von *ediarum* (Berlin-Brandenburgische Akademie der Wissenschaften 2022), das ebenfalls auf einen X-Technologiestack aufbaut, kombinierbar.

Um den unterschiedlichen Ansprüchen verschiedener Editions-Nutzungs-Konzepte zu entsprechen, können in der individuellen Konfiguration von *Synopticon* Beschreibungen der Text- und Dokumentenrelationen abgelegt werden. Dabei wird zwischen Textansichten (bspw. diplomatische Ansichten, konstitutive Lesetexte, Textvarianten etc.) und Zusatzinformationen (bspw. kritischer Apparat, Informationen zum Textumfeld etc.) unterschieden. Das Projekt ging hervor aus der Bachelor-Arbeit von Yannik Herbst (2021), die am Zentrum für Philologie und Digitalität (ZPD) der Universität Würzburg bereut wurde. Im Jahr 2022 erhielt das Projekt eine Förderung vom Universitätsbund Würzburg und bildet inzwischen eine Basiskomponente im Portfolio des ZPD.

Herausforderungen

Eine allgemeine Herausforderung stellt zunächst die Synchronisierung der Ansichten in Abhängigkeit von verschiedenen User-Interaktionen dar (Blättern, Scrollen etc.). Diese müssen projektspezifisch und in Abhängigkeit von dem jeweiligen Editions-konzept konfigurierbar sein, bilden also einen Teil der generischen *Synopticon*-Konfiguration. Zudem müssen dort weitere User-Interaktionen einbezogen werden, die ebenfalls eine Veränderung des Spalteninhalts hervorrufen oder sogar eine Resegmentierung des Textes zur Folge haben können (ähnlich LERA, vgl. Pöckelmann et al. 2022). Eine weitere Herausforderung besteht in den Konfigurationsmöglichkeiten der Spalten in Abhängigkeit von dem allgemeinen Nutzungskonzept der digitalen Edition. Daran knüpft sich zugleich die responsiv bedingte Variierbarkeit der sichtbaren Spalten und die jeweiligen Anpassungen des Nutzungskonzepts. Diese zahlreichen teils nutzer-teils projektspezifischen Anpassungen, die ermöglicht werden sollen, werfen Fragen der Referenzierbarkeit auf. Im Kontext von digitalen Editionen müssen Ansichten derselben reproduzierbar und zitierfähig sein.

Umsetzung als Software-Architektur

Für Frontend und Backend von *Synopticon* kommen unterschiedliche Open-Source-Frameworks zum Einsatz (Abb. 8). Im Frontend sorgt *Vue.js* für die Organisation der verschiedenen Informationen und Strukturen. Auf Benutzerwunsch (bspw. Auswahl einer Textvariante)

werden Anfragen an die von *ediarum.Web* gestellte Schnittstelle geschickt. Diese liefert als Antwort entweder HTML-Bruchstücke oder JSON, wobei die HTML Bruchstücke durch XSLT Transformationen ("Views") der zugrunde liegenden XML Dateien erzeugt werden.

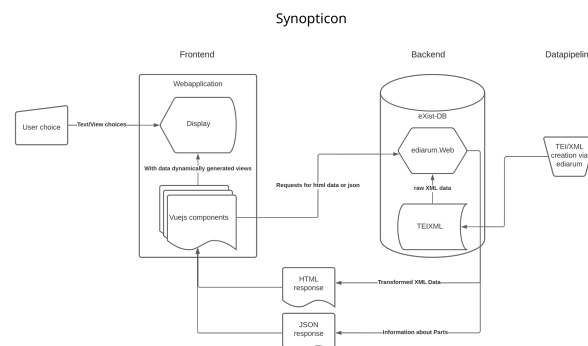


Abb. 8: Makrostruktur einer mittels *Synopticon* umgesetzten digitalen Editions-Webapplikation. Grafik: Yannik Herbst.

Das Frontend wurde mit dem JavaScript-Framework *Vue.js* (You 2022) realisiert. Um zu ermöglichen, dass das Frontend in viele unterschiedliche Projekte aufgenommen werden kann, wurde es als Single-Page-Anwendung konzipiert, welche mittels einer JavaScript-, einer CSS- und einer Konfigurations-Datei in bestehende Webseiten eingebunden wird. Der modulare und auf wiederverwertbaren Komponenten aufbauende Ansatz von *Vue.js* sowie die Konfiguration von *Synopticon* ermöglicht es, projektspezifische Anforderungen mithilfe bereits bestehender Konfigurations- und Softwarekomponenten aus anderen Projekten zu realisieren. Somit wird der Weg für ein nachhaltiges und nachnutzbares Softwaredesign geebnet.

Als Framework kommt aktuell *ediarum.WEB* (Fechner 2022) zum Einsatz, das im Bereich der digitalen Editionen bereits verbreitet ist. Dieses dient der Kommunikation zwischen Frontend und der XML-Datenbank *eXist* (Meier 2022) des Backends und enthält die dazu notwendigen Schnittstellen und Konfigurationsmöglichkeiten. Außerdem beinhaltet es notwendige Konsistenzchecks und Error-Handling. Das Ökosystem von *ediarum* (.BASE.edit, .REGISTER.edit, .WEB, .DB) bietet außerdem die Möglichkeit, den gesamten Workflow von Datenakquise bis hin zur Präsentation zu integrieren. Dieser Workflow, inklusive der Präsentation mittels *Synopticon*, wird derzeit in verschiedenen am ZPD betreuten Projekten erprobt.

Aktuelle Beispielprojekte

Drei aktuelle Usecases decken diverse Synchronisations-Beziehungen ab. Ein Beispiel ist das DFG-geförderte Projekt *Narragonia Latina* (Baier/Hamm 2021), in dessen BMBF-Vorgängerprojekt *Narragonien Digital* (Burricher/Hamm 2021) eine projektspezifische Synopse entwickelt wurde, die als Ausgangspunkt bzw. als Inspiration für die Entwicklung der generischen Lösung *Synopticon* diente. Das Projekt erarbeitet eine kommentierte zweisprachige Hybridedition der beiden lateinischen "Narrenschiffe" von Jakob Locher (1497) und Jo-

docus Badius (1505), die verschiedene Druckausgaben, insbesondere Übersetzungen desselben Originalwerkes gegenüberstellt (Abb. 9) und die beiden "Narrenschiffe" mittels eines digitalen Kapitel-Kommentars interdisziplinär und vergleichend erläutert.

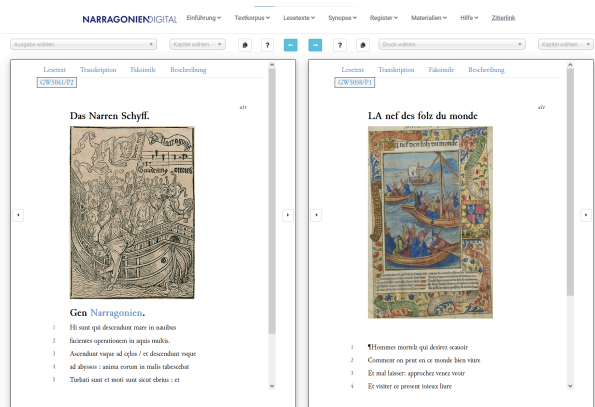


Abb. 9: Synoptische Vergleichsansicht zweier Ausgaben desselben Originalwerkes im Projekt *Narragonien digital* (Burrichter/Hamm 2021). Exemplarischer Screenshot: "Das Narren Schyff / La nef des folz du monde", <https://www.narragonien-digital.de/exist/synopsis.html?book1=GW5041&chap1=GW5041n2&book2=GW5065&chap2=GW5065n1> (zugegriffen: 14. Dezember 2022).

Ein weiteres Beispiel ist das interdisziplinär ausgerichtete Akademieprojekt *Richard Wagner Schriften*, dessen Ziel es ist, erstmals die gesamte Hinterlassenschaft der rund 230 Texte Richard Wagners mit ca. 5.000 Seiten Umfang auf Basis aller überlieferter historischer Textzeugen philologisch zu erschließen und in einer vollständigen, historisch-kritischen Edition der Schriften mit umfassendem Kommentar zu bündeln. Dabei werden einem Text sowohl sein Apparat, sein Textumfeld und ggf. vorhandene Digitalisate gegenübergestellt (Abb.10).

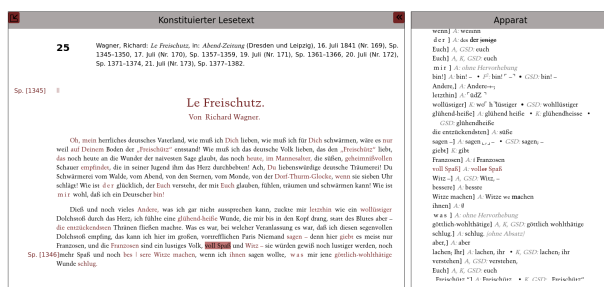


Abb.10: Projektspezifische Anpassung der durch Synopticon zur Verfügung gestellten Funktionalität für die Gegenüberstellung von Text mit einem dazugehörigen textkritischen Apparat. Exemplarischer Screenshot: Richard Wagner, "Le Freischütz". In *Richard Wagner Schriften*. <https://wagner-schriften.de> (zugegriffen: 14. Dezember 2022).

Zusammenfassung und Ausblick

Synopticon bündelt mehrere verschiedene konzeptionelle Anforderungen spaltenorientierter Layouts digitaler Editionen. Vor dem Hintergrund zahlreicher Beispiele aus bereits bestehenden, projektspezifischen Lösungen betont der vorliegende Ansatz die Notwendigkeit einer verallgemeinerten Formalisierung von Interfaces bezüglich deren Layouts und deren Funktionalitäten. Hinsichtlich der Nachhaltigkeit der Infrastrukturen für digitale Editionen und deren Interfaces kann in einem solchen Modell eine Schlüsselfunktion bestehen.

Seitens des ZPDs wird *Synopticon* bereits als Standardinstrument für synoptische Interfaces digitaler Editionen genutzt und kontinuierlich weiterentwickelt. Es wird angestrebt, im Rahmen zukünftiger Editionsprojekte sowohl die Funktionalität als auch die Interoperabilität von *Synopticon* zu erweitern: Perspektivisch soll *Synopticon* zunächst modular kompatibel zur Editionssoftware *ediarum* werden, bevor die Kombinierbarkeit mit anderen Frameworks weiter ausgelotet wird. Als weiterer konzeptioneller Bestandteil wird die generische Einbindung von Digitalisaten durch das *Image Interoperability Framework* (IIIF) angestrebt. Des Weiteren zeigten Jänicke et al. (2017) bereits auf, dass synoptische Ansichten nicht nur Nutzungsszenarien des klassischen *close reading* bedienen, sondern durchaus auch in der Lage sind, auch Distant Reading oder kombinierte Methoden zu befördern, so dass mit hinzukommenden Kooperationsprojekten eine weitere Flexibilisierung von *Synopticon* in Aussicht steht. Ferner könnten zukünftig auch Schnittstellen zu Textanalyse-Tools (z. B. *Voyant*, *LERA*, etc.) und nach den Richtlinien der *Text Encoding Initiative* (TEI) ausgezeichnete Textkorpora (ggf. via NFDI Text+) eingebunden werden.

Bibliographie

Baier, Thomas; Hamm, Joachim. 2021. *Narragonia latina: Kommentierte zweisprachige Hybridedition der lateinischen Narrenschiffe von J. Locher und J. Badius (1497/1505)*. <https://www.narragonia-latina.de> (zugegriffen: 14. Dezember 2022).

Burrichter, Brigitte; Hamm, Joachim. 2021. *Narragonien digital. Digitale Textausgaben von europäischen 'Narrenschiffen' des 15. Jahrhunderts*. <https://narragonien-digital.de> (zugegriffen: 14. Dezember 2022).

Berlin-Brandenburgische Akademie der Wissenschaften. 2022. *ediarum - Digitale Editionen erstellen und publizieren*. <https://www.ediarum.org> (zugegriffen: 14. Dezember 2022).

Fechner, Martin. 2022. *ediarum.WEB Version 1.13.4*. Zenodo, 10.5281/zenodo.5940545.

Flanders, Julia; Hamlin, Scott. 2013. "TAPAS: Building a TEI Publishing and Repository Service". In *Journal of the Text Encoding Initiative* 5, 10.4000/jtei.788.

Gleicher, Michael; Albers, Danielle; Walker, Rick; Jusufi, Ilir; Hansen, Charles D.; Roberts, Jonathan C. 2011. "Visual comparison for information visualization." In *Information Visualization* 10(4): 289–309, 10.1177/1473871611416549.

Herbst, Yannik. 2021. "Implementierung eines generischen Frameworks zur Visualisierung synoptischer Ansichten in digitalen Editionen". Bachelor-Thesis, Julius-Maximilians-Universität Würzburg (unveröffentlicht).

Herbst, Yannik. 2022. Synopticon. <https://github.com/zpd-digital-editions/Synopticon> (zugegriffen: 14. Dezember 2022).

Jänicke, Stefan; Franzini, Greta; Cheema, Muhammad F.; Scheuermann, Grik. 2017. "Visual text analysis in digital humanities." In *Computer Graphics Forum* 36(6): 226-250, 10.1111/cgf.12873.

Javed, Waqas; Elmqvist, Niklas. 2012. "Exploring the design space of composite visualization." 2012 *IEEE Pacific Visualization Symposium*, 1-8, 10.1109/PacificVis.2012.6183556.

Meier, Wolfgang. 2022. *eXist-db*. <http://exist-db.org/> (zugegriffen: 14. Dezember 2022).

Pöckelmann, Marcus; Medek, André; Ritter, Jörg; Molitor, Paul. 2022. "LERA—an interactive platform for synoptical representations of multiple text witnesses". In *Digital Scholarship in the Humanities*, 30. Juni. Oxford University Press, 10.1093/llc/fqac021.

You, Evan. 2022. *Vue.js*. <https://vuejs.org> (zugegriffen: 14. Dezember 2022).

Textliche Relationen maschinenlesbar formalisieren: Systeme der Intertextualität

Horstmann, Jan

jan.horstmann@uni-muenster.de
Westfälische Wilhelms-Universität Münster,
Deutschland

Lück, Christian

christian.lueck@uni-muenster.de
Westfälische Wilhelms-Universität Münster,
Deutschland

Normann, Immanuel

immanuel.normann@uni-muenster.de
Westfälische Wilhelms-Universität Münster,
Deutschland

Intertextualität als theoretisches Konzept in der (digitalen) Forschung

Wie können intertextuelle Beziehungen formalisiert und annotiert werden? Was wäre ein kohärentes Kate-

goriensystem der Intertextualität und welche Formalisierung ist geeignet, um es computergestützt berechenbar zu machen, ohne seine Aussagekraft zu verlieren. Intertextualität ist eine komplexe und zugleich sehr zentrale Kategorie in der Literaturanalyse. Per Definition betrifft sie nicht nur *ein* (literarisches) Artefakt, sondern mindestens zwei und beschreibt die Beziehung zwischen ihnen (vgl. Pfister 1985, 11). Diese Beziehungen können auf zahlreichen Ebenen zu finden sein und je nach persönlichem Verständnis von „Intertextualität“ können entweder Grenzen gezogen werden oder Beziehungen überall zu finden sein. Etliche literarische Gattungen konstituieren sich durch ihren inhärenten Verweischarakter: Persiflage, Parodie, Pastiche, Cento (Flickengedicht), Travestie, Stilkopien usw. Unterhalb der Ebene der Konstitution von Gattungen lässt sich Intertextualität als feingranuläres Geflecht von Bezugnahmen in den meisten, wenn nicht gar allen Texten feststellen. Auf dieser Ebene wird Intertextualität mit Begriffen wie Hypolepse (textuelle Reaktionen auf andere Texte durch z.B. Zustimmung, Ablehnung, Weiterführung, Korrektur usw.) oder Anspielung beschrieben. Zudem können hier Bezugnahmen in Form von z.B. nicht-gattungskonstituierenden Stilkopien oder im parodistischen Modus einer Textpassage vorkommen. Generell unterscheidet man dabei Einzeltextreferenzen von sog. Systemreferenzen, d.h. Referenzen auf ganze „Systeme“ wie etwa Gattungen.

In der Forschung finden sich daher zahlreiche Ansätze, das von Julia Kristeva (1967) benannte theoretische Konzept Intertextualität als einen Beschreibungsbegriff für die Beziehung zwischen Texten zu systematisieren. Zu nennen sind in diesem Zusammenhang insbesondere Barthes (1984), Genette (1982), Pfister (1985), oder in digitaler Hinsicht Scheirer et al. (2016), Schlupkoth und Nantke (2019) und Burghardt und Liebl (2020). Die einzelnen Ansätze und Systematisierungen haben in der Regel verschiedene theoretische oder praxeologische Hintergründe (z.B. Strukturalismus, Poststrukturalismus oder Digital Humanities) und damit verbunden verschiedene Fokusse. Für Kristeva (1967) etwa bildet das Konzept Intertextualität einen theoretischen Zugang zur Dialogizität literarischer Texte. Mit Bezug auf Michail Bachtin entwickelt sie ein Verständnis von Text als „Mosaik von Zitaten“ (Kristeva 1967, 348). Genette (1982) beschreibt – ebenfalls mit dem Ziel einer Gattungstypologie – textuelle Bezugnahmen als Transformation oder Nachahmung hypertextueller Textgattungen. Er differenziert fünf Formen: 1. Intertextualität durch Zitate, Plagiate oder Anspielungen, 2. Paratextualität, womit er Rahmungen wie Titel, Genreklassifikationen, Autorname etc. meint, 3. Metatextualität (d.h. kritische Kommentare), 4. Architextualität (d.h. externe, z.B. durch Kritiker zugewiesene Rahmungen) sowie 5. Hypertextualität, bei der ein späterer Text (Hypertext) ohne einen vorherigen Bezugstext (Hypotext) nicht denkbar ist. Hypertextualität unterteilt er schließlich in Transformation (James Joyce transformiert in *Ulysses* die *Odyssee* Homers – „dasselbe anders sagen“) und Nachahmung (Vergil ahmt in der *Aeneis* die *Odyssee* nach – „etwas anderes auf dieselbe Weise sagen“; Genette 1982, 17). Pfister (1985) schließlich bietet ein graduelles System an, um die beiden Pole der sehr weiten und sehr engen (Inter)textualitätskonzeption miteinander zu verbinden.

In Operationalisierungsansätzen der Digital Humanities wird Intertextualität als *text reuse* (vgl. Burghardt und Liebl 2020) oder im Sinne von *event alignments* (vgl. Reiter und Frank 2015) verstanden. Scheirer et al. (2016) versuchen mithilfe des *Latent Semantic Indexing* (LSI) darüber hinausgehend semantisch-thematische intertextuelle Bezüge zwischen Ausgangs- und Bezugstexten zu modellieren. Diese Ansätze zielen darauf ab, den Weg für die Entwicklung eines automatischen Detektors für intertextuelle Beziehungen zu ebnen. Dies ist nicht unser Ziel. Die Wiederverwendung von Texten, das automatisierte Auffinden von Zitaten oder ähnlichen Textpassagen in einem Textkorpus – an sich schon anspruchsvoll aus Sicht der Informatik – erscheint aus der Perspektive der Literaturwissenschaft, die sich traditionell mit semantisch viel komplexeren Formen intertextueller Beziehungen beschäftigt, häufig unterkomplex.

Ziel des Beitrags ist – statt von bestimmten digitalen Verfahren auszugehen – die theoriegeleitete Modellierung eines maschinenlesbaren Schemas, eines Kategoriensystems, das strukturell und grundlegend Analysen von Intertextualität, wie sie in literaturwissenschaftlichen und -theoretischen Abhandlungen zu finden sind, repräsentieren kann. Schlupkothén und Nantke (2019) verfolgen ein ähnliches Ziel. Bei ihnen ist aber nicht weiter ausgearbeitet, wie sich das Vorhaben, analytisch-interpretatorische Lektürepraktiken zu repräsentieren, zur eingesetzten Technologie X-Link verhält und welche Beziehung diese zu der von den Autoren ins Spiel gebrachten Situationslogik hat.

Formale Methoden und maschinelle Interpretierbarkeit

Ob ein logisches System (Prädikatenlogik, Beschreibungslogik, Situationslogik etc.) für die Formalisierung eines Forschungsgegenstandes geeignet ist, hängt von den Zielen ab, die mit der formalen Repräsentation des Gegenstandes erreicht werden sollen. Ist der Gegenstand formal repräsentiert, lassen sich durch ein formales Kalkül Entscheidungsfragen hinsichtlich ihres Wahrheitswertes auswerten und (neue) Aussagen aus dem Formalisierten ableiten, worunter auch das Abfragen der 'Fakten'-Basis zählt. Ganz allgemein sind bei der Wahl eines logischen Systems folgende Aspekte ausschlaggebend: Es sollte so ausdrucksmächtig sein, dass es für die formale Repräsentation des Gegenstandes geeignet ist (Ausdrucksmächtigkeit). Und für die DH ist es wünschenswert, dass der formale Kalkül von einem Computer ausgeführt werden kann (Implementierung), und zwar zudem effizient (Komplexität).

Ziel unserer Formalisierung der Domäne Intertextualität ist die Repräsentation von Intertextualitätsanalysen. Eine Einschränkung auf eine bestimmte Intertextualitätstheorie oder auf eine bestimmtes Teilphänomen, etwa werkästhetisch manifeste Intertextualität, soll zunächst nicht erfolgen. Stattdessen versuchen wir, einen gemeinsamen Kern von Intertextualitätskonzepten freizulegen, und zwar so, dass er für spezielle Theorien erweiterbar ist. Ziel ist also, intertextuelle Relationen zu annotieren, abzufragen und ggf. Aussagen abzuleiten.

Formalisierung des Kerns

Die Formalisierung wird zunächst mit halbformalen Mitteln durchgeführt: mit einer Liste dessen (der Aspekte oder Merkmale), was repräsentiert werden soll. Einen ersten Zugang zum gemeinsamen Kern der verschiedenen kursierenden Intertextualitätstheorien bietet eine Analyse des Wortes Intertextualität. Es besteht aus den lexikalischen Morphemen *inter* und *text* sowie dem Derivationsmorphem *-alität*, fr. *-alité*, lat. *-alitas*. Während *text* gegenständlich ist, ist *inter* präpositional. Also einfach zwischen Text bzw. zwischen Texten? Ganz so „durchsichtig und selbsterklärend“ (Adamzik 2004, 96) ist das Wort jedoch nicht, denn die Präposition *inter* setzt nicht nur zwei (oder mehrere) Entitäten miteinander in Beziehung, sondern gibt dem Zwischenraum auch ein eigenes Sein; siehe z.B. 'interlineare Annotationen'. Möglicherweise liegt genau hier der Designfehler des Neologismus, denn Theorien des globalen Intertexts, von dem im Singular gesprochen wird, rücken anscheinend diesen Zwischenraum ins Zentrum ihres Sprachspiels, ohne weiter zu bestimmen, was dort ist. Hinzu kommt, dass die Präposition *inter*, anders als z.B. *trans* ungerichtet ist. Das steht im Gegensatz dazu, dass das theoretische Konzept auch entwickelt worden ist, um Texten eine historische Dimension zu geben, nämlich auf ihre Avant-Texte hin, so schon bei Kristeva und insbesondere in Genettes Palimpsest-Metapher und seinem Transtextualitätskonzept (vgl. Dosse 1997, II, 446f.). In der historischen Dimension gibt es aber Gerichtetheit; und ein Bezug auf zukünftige Texte kann unschwer als gegenwärtige Vorstellung zukünftiger Texte angesehen werden. Im Kern, so scheint uns, rückt die Untersuchung von Intertextualität immer einen späteren Text ins Zentrum und untersucht seine Bezüge zu früheren Texten; andersherum handelt es sich um Wirkungs- und Rezeptionsforschung. Im Kern, so halten wir fest, zielt Intertextualität auf die Beziehung (1a) *zwischen Texten*, und zwar derart, dass eine intertextuelle Beziehung (1b) *anti-chronologisch gerichtet* ist. Auch wenn Beschreibungen von Intertextualität mitunter Komplexe von vielen Texten in den Blick nehmen, so können solche Komplexe doch als eine Menge (1c) *binärer* intertextueller Beziehungen repräsentiert werden, sofern die Möglichkeit gegeben ist, Vermittlung durch ein Drittes sowie transitive Relationen zu modellieren. Der Formalismus muss also erlauben, dass (2) Vermittlungsinstanzen notiert werden können. Solche können wiederum binäre intertextuelle Relationen sein, jedoch muss die Kategorie der Vermittlungsinstanz auf eine Vielzahl von Phänomenen hin erweiterbar sein. Selbstredend muss eine intertextuelle Relation zudem auch mit den (3) *speziellen Kategorien der verschiedenen Theorien* bestimmt- bzw. beschreibbar sein, wobei diese Bestimmung nach Kategorien auch Grade (vgl. Pfister 1985) umfassen können muss. Das Textkonzept als solches, bis hin zu der Frage, ob Text manifeste Schrift in einem Zeichensystem ist oder ein weiteres Verständnis vorliegt, gehört aber nicht zum Kern von Intertextualität, sondern zur theoretischen Ausprägung, also aus unserer Perspektive zur Peripherie. Der Aspekt der Markiertheit auf der Textoberfläche, den Pfister und insbesondere die Textlinguistik interessieren, könnte sich nicht zuletzt im Hinblick auf Weiternutzung durch Textmining als fruchtbar erweisen und ist allge-

meiner Art. Wir halten folgende Form intertextueller Relationen fest und notieren dabei zugleich die Anzahl der genannten Aspekte, wobei n eine natürliche Zahl ist und n_0 eine natürliche Zahl oder Null: TextHier/1, TextDort/1, Vermittlungsinstanz/ n_0 , Bestimmung/ n , Marker/ n_0 .

Der auf diese halb-formale Art entworfene Kern von Intertextualität lässt sich nicht mit allen formalen Methoden repräsentieren. Zur Formalisierung von Intertextualität ist ein Ausdrucksmittel von Relationalität erforderlich. Damit scheidet die Aussagenlogik aus und es muss auf eine Prädikatenlogik zurückgegriffen werden. Die allgemeine Prädikatenlogik erster Stufe ist wiederum ausdrucksstärker als für unser Vorhaben nötig. Als ausreichend ausdrucksstark erweist sich die Beschreibungslogik, die in einfacher Ausprägung im Wesentlichen der Prädikatenlogik entspricht mit der Einschränkung auf ein- und zweistellige Prädikate (vgl. Baader et al., Hg. 2010, Harmelen et al., Hg. 2008). Für sie steht in RDF eine Implementierung zur Verfügung. Wir geben hier nur ein Beispiel für eine in RDF notierte intertextuelle Relation (für ihre Form) und verweisen für die Ontologie in RDFS/OWL auf unser Github-Repository.¹

```
@prefix : https://intertextuality.org/abstract#> .
@prefix ex: https://example.org/my-inter-
textual-findings/> .
ex:i a :IntertextualRelation;
:here ex:t1 [a :Reference]; # TextHier/1
:there ex:t2 [a :Reference]; # TextDort/1
:mediatedBy ex:md [a :Mediator]; # Vermitt-
lungsinstanz/ $n_0$ 
:specifiedBy ex:s [a :IntertextualSpecifica-
tion]; # Bestimmung/ $n$ 
:markedBy ex:mrk [a :Marker ]. # Marker/ $n_0$ 
```

In der Ontologie ist ausgedrückt, dass intertextuelle Relationen Mediatoren sein können: `:IntertextualRelation rdfs:subClassOf :Mediator`. Auf diese Weise wird die Struktur rekursiv bzw. können vermittelte, transitive Relationen repräsentiert werden. Die Klasse `Reference`, welche für die Texte verwendet wird, zwischen denen die intertextuelle Relationen beschrieben wird, wird mit der Ontologie von Web Annotations modelliert.² Dies gilt potentiell auch für die Klasse `Marker`, jedoch sind auch andere Marker, z.B. Typen aus einem System von Markierungen, realisierbar. Die Metadaten der Referenzen, insbesondere das Datum der Ersterscheinung, ermöglichen, die derart formalisierten Aussagen hinsichtlich ihrer anti-chronologischen Gerichtetheit auf Konsistenz zu prüfen (Reasoning).

Erweiterungen: Genette

Der Kern ist erweiterbar, indem die Klassen `Mediator`, `IntertextualSpecification` und `Marker` durch Bildung von Subklassen differenziert und spezifiziert werden. Beispielhaft soll dies hier für Genettes Beschreibungskategorie Hypertextualität durchgeführt werden. Verglichen mit Kristevas Intertextualitätsbegriff bezeichnet sie ein sehr enges Feld, das durch eine 1-zu-1-Beziehung zweier Texte, des Hypertextes B zu einem vor-

hergehenden Hypotext A, gekennzeichnet ist. Die 1-zu-1-Beziehung wird dabei weniger durch Einzelstellen der Texte gestiftet. Vielmehr setzen paratextuelle Signale (z.B. der Titel des Hypertextes) die Texte als Ganzes in hypertextuelle Beziehung. Dennoch manifestiert sie sich auch im Verhältnis einzelner Textstellen. Eine Formalisierung dieser auf einer Relation beruhenden Relationen, die bislang ein Desiderat der DH geblieben ist, wird auf Grundlage des vorgeschlagenen Kerns möglich. Wir schlagen vor, das paratextuelle Signal des Titels als Vermittlungsinstanz der vielen intertextuellen Relationen zwischen Einzelstellen aufzufassen.

```
@prefix g:https://intertextuality.org/ex-
tensions/genette/hypertextuality#> .
@prefix pt:https://intertextuality.org/ex-
tensions/genette/paratext#> .
@prefix :https://intertextuality.org/abs-
tract#> .
@prefix rdfs:http://www.w3.org/2000/01/rdf-
schema#> .
@prefix owl:http://www.w3.org/2002/07/owl#> .
<https://
intertextuality.org/extensions/genette/hy-
pertextuality> a owl:Ontology .
<https://
intertextuality.org/extensions/genette/para-
text#> a owl:Ontology .
pt:ParatextualSignal rdfs:subClassOf :Media-
tor .
pt:title a pt:ParatextualSignal .
```

Vor allem aber gewinnt Genette aus der Analyse der Hypertextualität eine Gattungstypologie. Dazu unterteilt er diese Beziehung nach zwei Gesichtspunkten: Es kann sich entweder im Hinblick auf den Relationstyp um eine Imitation oder um eine Transformation handeln, und sie kann im Hinblick auf die Art und Weise entweder spielerisch, satirisch oder ernst sein. Daraus gewinnt er durch Kombination sechs Gattungen. In RDFS/OWL formalisiert heißt das:

```
g: HypertextualRelation rdfs:subClassOf :In-
tertextualSpecification ;
owl:disjointUnionOf (g:RelationalType g:Mo-
dalType) .
g:RelationalType owl:ObjectOneOf (g:trans-
formation g:imitation) .
g:ModalType owl:ObjectOneOf (g:playfully
g:satirically g: seriously) .
```

Die Formalisierung wäre dadurch zu ergänzen, dass eine solche hypertextuelle Relation auch über eine Vermittlungsinstanz verfügen muss.

Fazit und mögliche Anschlussforschung

Die hier vorgeschlagene Formalisierung auf Grundlage der Beschreibungslogik kommt an bestimmten Stellen an ihre Grenzen. Eine davon steckt im Begriff Situation: Eine der fundamentalen Unterscheidungen intertextuel-

ler Relationen ist die zwischen solchen, die werkästhetisch manifest sind (etwa durch paratextuelle Signale oder andere Marker), und solchen, die rezeptionsästhetisch gefunden worden und damit unklarer sind (vgl. Pfister 1985, 23f.). Mit einem rezeptionstheoretischen Hintergrund schlagen Schlupkothén und Nantke (2019) die Situationslogik nach Barry und Parwise als angemessene formale Methode vor. Die hier vorgeschlagene auf der Beschreibungslogik bzw. OWL basierende Formalisierung bietet die Möglichkeit, die Situation durch Metadaten (Name, Datum) zu kodieren und einem Datensatz anzuhängen. Allerdings ist das keine Implementierung der Situationslogik, bei dem es insbesondere um eine Formalisierung des Zusammenhangs von Situation und Konsistenz geht. Denselben Einwand wird man auch gegenüber dem Beitrag von Schlupkothén und Nantke einwenden können, denn die von ihnen eingesetzt Technologie X-Link ist ebenfalls keine Implementierung der Situationslogik.

Unser Beitrag hat das Ziel, in der Pluralität unterschiedlicher Konzeptionen von Intertextualität einen Kern von Intertextualität herauszuschälen und so zu formalisieren, dass er durch Theorien erweiterbar ist. Er bahnt damit einen Weg zur Repräsentation intertextueller Beziehungen, welche einerseits der Komplexität der literaturtheoretischen Konzepte gerecht wird und die andererseits Berechenbarkeit gewährleistet.

Der Beitrag richtet sich damit einerseits an Intertextualitätstheoretiker*innen und -praktiker*innen. Erstere können durch unsere Formalisierung ihren Intertextualitätsbegriff schärfen: durch eine weitere Verfeinerung des von uns vorgeschlagenen Modells oder durch eine klare Abgrenzung des eigenen Intertextualitätsbegriffs. Intertextualitätspraktiker*innen wird ein (erweiterbares) Modell an die Hand gegeben, um Intertextualität zu identifizieren/zu annotieren und zu analysieren (z.B. mittels Netzwerkvisualisierung, Netzwerkanalyse oder durch eine synoptische Gegenüberstellung von Textpassagen mit intertextuellem Bezug). Andererseits kann unsere theoretische Konzeption die Grundlage für die Architektur einer möglichen Forschungsumgebung bilden, die den Intertextualitätsforschenden sowohl eine Weiterentwicklung oder Anpassung des vorgeschlagenen Intertextualitätsmodells als auch die Erforschung der Intertextualität eines annotierten Textkorpus auf Basis dieses Modells erlaubt.

Fußnoten

1. Vgl. <https://github.com/janhorstmannn/intertextuality> (Zugriff: 12.12.2022).
2. Vgl. <https://w3c.github.io/web-annotation/model/wd/> (Zugriff: 12.12.2022).

Bibliographie

- Adamzik, Kirsten. 2004. *Textlinguistik. Eine einführende Darstellung*. Tübingen: Niemeyer.
- Baader, Franz et al. (Hg.). 2010. *The Description Logic Handbook. Theory, Implementation, Applications*, 2. ed., New York: Cambridge UP.

Barthes, Roland. 1984. "La mort de l'auteur." In Roland Barthes: *Le bruissement de la langue*. Paris: Seuil.

Burghardt, Manuel und Bernhard Liebl. 2020. "'The Vectorian' – Eine parametrisierbare Suchmaschine für intertextuelle Referenzen." 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.4621836>.

Dosse, François. 1997. *Geschichte des Strukturalismus*. 2 Bde. Frankfurt/M: Fischer.

Genette, Gérard. 1982. *Palimpsestes: la littérature au second degré*. Collection Poétique. Paris: Seuil.

Harmelen, Frank van et al. (Hg.). 2008. *Handbook of Knowledge Representation*. Amsterdam: Elsevier.

Kristeva, Julia. 1967. "Bakhtine, le mot, le dialogue et le roman." *Critique* 23: 438–465.

Pfister, Manfred. 1985. "Konzepte der Intertextualität." In *Intertextualität. Formen, Funktionen, anglistische Fallstudien*, hg. von Ulrich Broich und Manfred Pfister, 1–30. Tübingen: Niemeyer.

Reiter, Nils und Anette Frank. 2015. "Computerlinguistische Verfahren zur Aufdeckung struktureller Ähnlichkeiten in Narrativen." In *DHd 2015 Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation. 2. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2015)*. Graz. <https://doi.org/10.5281/zenodo.4623234>.

Scheirer, Walter, Christopher W. Forstall und Neil Cofee. 2016. "The sense of a connection: Automatic tracing of intertextuality by meaning." *Digital Scholarship in the Humanities* 31/1: 204–217. <https://doi.org/10.1093/llc/fqu058>.

Schlupkothén, Frederik und Julia Nantke. 2019. "Formlit: Eine multimodale Arbeitsumgebung zur systematischen Erfassung literarischer Intertextualität." In *DHd 2019 Digital Humanities multimedial und multimodal. 6. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2019)*. Frankfurt am Main, Mainz. <https://doi.org/10.5281/zenodo.4622106>.

Tool Studies 2.0 – Zum Potenzial von Transformern für die Erkennung und Klassifikation von Software-Tools in DH- Publikationen

Burghardt, Manuel

burghardt@informatik.uni-leipzig.de
Computational Humanities, Universität Leipzig

Ruth, Nicolas

nicolas.ruth@web.de

Computational Humanities, Universität Leipzig

Niekler, Andreas

aniekler@informatik.uni-leipzig.de

Computational Humanities, Universität Leipzig

Einleitung: Zur Rolle von Tools in den Digital Humanities

Software-Tools spielen in den Digital Humanities eine zentrale Rolle, ja, sind geradezu genre-prägend für diese Disziplin. Diese wichtige, praktische Rolle ist u.a. belegt durch die Existenz diverser Tutorials und vor allem auch von Tool-Katalogen (vgl. Tab. 1) sowie Versuchen der einheitlichen Kategorisierung von Tools, etwa der TaDi-RAH-Taxonomie (Borek et al., 2016).

Tabelle 1: Übersicht zu verschiedenen Tutorials und Katalogen für Tools in den Digital Humanities.

Name	Typ	URL
Programming Historian	Tutorials	https://programminghistorian.org/
forTEXT	Tutorials	https://fortext.net/
CLARIAH-DE Tutorial Finder	Tutorials	https://teaching.clariah.de/search/
TAPoR	Katalog	https://tapor.ca/home
Digital Methods Initiative	Katalog	https://wiki.digitalmethods.net/Dmi/ToolDatabase
Alan Liu's DH Toychest	Katalog	http://dhresourcesforproject-building.pbworks.com
SSH Open Marketplace	Katalog	https://marketplace.sshopencloud.eu/
NFDI4Culture	Katalog	https://riajournal.com/article/57036/instance/5947376/ ; Tabelle 9, Appendix

Neben diesen stärker praxeologischen Aspekten finden sich auch diverse Diskurse um die Rolle und Implikationen von Tools in den Digital Humanities. Ein Thema ist dabei etwa „Tool Criticism“, also der kritische Umgang mit Tools, insbesondere deren spezifischer Funktionsweise und den Effekten, die diese auf die letzten Ergebnisse haben (Koolen et al., 2019; Traub & van Ossenbruggen, 2015; van Es et al., 2018). Ein weiteres Themenfeld findet sich im Bereich der Mensch-Maschine-Interaktion, also den speziellen Anforderungen, die geisteswissenschaftliche Forschende an die Funktionalität und Usability von Software-Tools mitbringen (Burghardt & Wolff, 2014; Wolff, 2015).

Besonders stark ausgeprägt ist die Diskurslinie, die sich mit der Frage um die epistemologischen Effekte – also die unmittelbare Auswirkung von digitalen Tools auf den geisteswissenschaftlichen Erkenntnisprozess – beschäftigen (Burghardt et al., 2022; Dalbello, 2011; Kaden, 2016; Ramsay & Rockwell, 2012).

Dabei erstrecken sich die genannten Diskussionen vornehmlich auf Einzelbeispiele, größere empirische Untersuchungen gibt es bislang nur wenige. Dazu zählt u.a. eine Studie aus dem Umfeld des Research Software Engineering, bei der die Art und Häufigkeit von Softwarezitationen in DHd-Abstracts analysiert wurde (Henny-

Krahmer & Jettka, 2022). Weiterhin sind hier die zahlreichen Experimente von Frank Fischer und Kollegen, die u.a. DH-Abstracts und Tutorials des Programming Historian auf Tool-Vorkommen hin untersucht haben (Barbot et al., 2019; Fischer & Moranville, 2020b, 2020a; Zarei, Alireza et al., 2022) sowie auch erste eigene Experimente (Burghardt et al., 2022) zu nennen. Ein erster Austausch zwischen den unterschiedlichen Akteur*innen im deutschsprachigen Raum fand weiterhin im Rahmen einer gemeinsamen Veranstaltung mit dem Titel „Die Werkbänke der Digital Humanities: Zur Rolle von Tools und Software für die Forschungsarbeit“ bei der vDHD (2021) statt.

Als grundlegende Methoden zur Erkennung von Tools finden sich in den genannten Studien vor allem zwei Ansätze: (1) lexikonbasierte Ansätze, bei denen mithilfe bestehender Tool-Listen, wie sie bspw. in den in Tab. 1 genannten Ressourcen verfügbar sind, ein einfacher look-up in einem Zielkorpus von DH-Publikationen erfolgt. Dieser Ansatz lässt sich sehr schnell umsetzen, leidet aber unter den üblichen Einschränkungen lexikonbasierter Verfahren, bspw. der Unvollständigkeit von statischen Wortlisten und gleichzeitig die Ambiguität einzelner Einträge (R, Python, Gate, etc.). (2) Der zweite Ansatz versucht diese Einschränkungen zu überwinden, indem auf verschiedene Verfahren des maschinellen Lernens mit dem Ziel eines Klassifikationstasks gesetzt wird. Hier gibt es einerseits vortrainierte Modelle aus dem Bereich der Software Entity Recognition (Patrice & Romary, 2015; Schindler et al., 2022), die aber zumeist aus dem naturwissenschaftlichen Bereich kommen und deshalb nur mäßig für den Einsatz in Digital Humanities-Publikationen geeignet sind (vgl. Henny-Krahmer & Jettka, 2022). In einer aktuellen Publikation experimentieren Zarei et al. (2022) mit einem auf Prodigy und dem spaCy-Framework basierenden NER-Ansatz, welcher für ein sehr klar umrissenes Anwendungsszenario mit vorhergehendem Training gute Ergebnisse bringt. Für einen Ansatz, der in der Lage ist auch Tools zu erkennen, die nicht vortrainiert wurden, schlagen die Autoren weiterführende Ansätze mit aktuellen Transformer-Modellen wie bspw. BERT (Devlin et al., 2019) vor. Erste Experimente mit BERT für eine binäre Klassifikation von Sätzen mit / ohne Tools wurden von uns bereits im Rahmen einer Vorstudie (Burghardt et al., 2022) erfolgreich durchgeführt und zeigten sich nach einer ersten qualitativen Evaluation als sehr vielversprechend. Das vorliegende Paper knüpft hier nahtlos an und präsentiert Erkenntnisse aus aktuellen Experimenten mit dem RoBERTa-Modell (Liu et al., 2019), einer optimierten Variante des bekannten BERT-Modells. Der von uns verfolgte Ansatz ist neben einer grundlegenden Identifikation von Tools über deren Embedding-Vektoren auch in der Lage die Tools größeren Kategorien, wie bspw. *Textanalysetool* oder *Visualisierungstool*, zuzuordnen. Wir glauben, dass wir mit einem solchen Ansatz einen wichtigen Beitrag zu den bestehenden Tool-Diskursen in den DH leisten können und hoffen damit weitere empirische „Tool Studies“ zu befördern.

Tool-Identifizierung und -Klassifikation mit RoBERTa Tool-Embeddings

Der Task der Tool-Identifizierung und -Klassifikation lässt sich methodisch als Problem der Text-Klassifikation einordnen. Beim gewählten Ansatz handelt es sich um ein Verfahren des überwachten Maschinellen Lernens, welches konkret aus einer binären und einer mehrklassigen Klassifikationsaufgabe besteht. Die grundlegende Idee beim nachfolgenden Vorgehen ist, dass die Erwähnung eines Tools in einem Paper in einem spezifischen sprachlichen Kontext steht, der sich in Form eines Sequenz-Embeddings mithilfe eines transformer-basierenden Sprachmodells abbilden und klassifizieren lässt.

Korpus, Trainings- und Testdaten

Am Anfang steht die Erstellung spezieller, gelabelter Trainings- und Testdatensätze. Der Ausgangsdatensatz, der ebenfalls als Untersuchungsgegenstand für die folgende Analysen dient, besteht aus 3.737 englischsprachigen Zeitschriftenpublikationen aus dem Bereich der Digital Humanities, die zwischen 1966 und 2020 veröffentlicht wurden (Luhmann & Burghardt, 2021). Eine manuelle Extraktion von Sequenzen in Papers, die die Nennung eines Tools beinhalten, ist in Anbetracht der vorliegenden Datenmengen nicht durchführbar. Stattdessen wurden automatisch bekannte Tools im Korpus gesucht, um entsprechende Sequenzen ausfindig machen zu können. Um qualitativ hochwertige Trainingsdaten zu erhalten, haben wir zunächst nach besonders populären Tools gesucht. Dazu haben wir insgesamt acht Listen und Tutorials mit Toolnennungen (vgl. Tab. 1) durchsucht und nur diejenigen Tools ausgewählt, die in mindestens zwei unterschiedlichen Listen genannt wurden. Aus diesen Tools haben wir dann zur weiteren Qualitätssteigerung diejenigen entfernt, die ein hohes Maß an Ambiguität aufwiesen. Dieses reduzierte Lexikon, mit insgesamt 246 Tools, dient als Ausgangspunkt für die Erstellung der Trainingsdaten. Im nächsten Schritt wurden sodann alle Papers des Korpus tokenisiert, mit dem Tool-Lexikon durchsucht und bei einem Treffer als Textausschnitt mit 15 Tokens vor und 15 Tokens nach dem Treffer als Sequenz-Label-Paar in den Trainingsdatensatz übernommen. Anschließend wurde die Menge der gefundenen Sequenzen durch zufällig gewählte und gleichmäßig über alle Papers verteilte, weitere Sequenzen ergänzt, die als Negativbeispiele für Toolnennung dienen.

Als Label wird hier für den binären Trainingsdatensatz „Tool“ und „Kein Tool“ verwendet und für den mehrklassigen Trainingsdatensatz die Zugehörigkeit des Tools zur bestehenden Taxonomie von Alan Liu's „DH Toychest“, da diese fast alle Tools in unseren Trainingsdaten bereits enthält und aus unserer Sicht auch gut nachvollziehbar ist. Langfristiges Ziel ist hier die Übernahme einer standardisierten Taxonomie wie TaDiRAH. Anschließend wurden die Datensätze für jeden der beiden Klassifikationstasks nach dem Prinzip einer 5-fold cross-validation in fünf Trainings- und Testsätze geteilt. Final besteht der

Datensatz aus 3.780 Sequenz-Label-Paaren in jedem der fünf folds. Für jede Runde der Kreuzvalidierung wurden vier folds für das Training und der fünfte fold als Testdatensatz für die Validierung benutzt.

Modell-Evaluation

Als transformer-basiertes Sprachmodell wurde eine RoBERTa Implementierung von Hugging Face gewählt.¹ Aus Effizienzgründen wurde konkret „distilroberta-base“ als „case-sensitive“- und „knowledge distilled“-Variante verwendet. Die Ergebnisse dieser Modell-Evaluation sind sehr vielversprechend. So wird für den Klassifikations-task der binären Tool-Detektion für beide Klassen bei den 5 folds im Schnitt ein F1-Score von 0,99 erreicht. Für den Task der Tool-Kategorisierung wurden ebenfalls gute F1-Scores im Bereich 0,94 - 0,99 erzielt, die sich von Kategorie zu Kategorie geringfügig unterscheiden (vgl. Tab. 2).

Tabelle 2: Ergebnisse des mehrklassigen Klassifikationsansatzes. Kategorien mit zu geringen Vorkommen für aussagekräftigen Klassifikationswert wurden entfernt (Testsamples > 100). *Mittelwert über alle Runden der Kreuzvalidierung

Class	F1-Score*
Authoring / Annotation / Editing / Publishing Platforms & Tools (including collaborative platforms)	0,97
Exhibition/Collection/Editorial Platforms & Tools	0,96
Platforms and Communication	0,99
Programming Languages/Packages	0,98
Text Analysis Tools	0,97
Visualization Tools	0,94
No Tool	0,99

Analysen

Die sehr guten Ergebnisse aus der Evaluation sollen im Folgenden im Rahmen einer beispielhaften Analyse weiter diskutiert werden. Dazu wurde der Ursprungsdatensatz mit den 3.737 DH-Papers mit einem „Sliding Window“-Ansatz und einer Fenstergröße von 31 Tokens (das entspricht der Größe unserer Trainingssequenzen, siehe oben) und einer Überlappung von fünf Tokens in insgesamt 753.210 Sequenzen geteilt. Diese wurden anschließend klassifiziert und in eine neue Datenbank geschrieben.

Ergebnisse der binären Toolidentifikation

Für den binären Klassifizierungstask der Tool-Detektion werden 14.561 Sequenzen mit potenziellen Toolnennungen vorhergesagt. Obwohl die Trainingsätze auf Basis von nur 246 unterschiedlichen Tools erstellt wurden, ist der Anteil an gefundenen Sequenzen, die genau eines dieser Tools enthalten mit 45,5% überschaubar – mehr als die Hälfte der Sequenzen enthalten andere Tools.

An dieser Stelle soll weiterhin ein Vergleich zu TAPoR, dem mit über 1.600 Einträgen größten Tool-Katalog, angestellt werden. 54,7% der von uns identifizierten Tool-Sequenzen enthalten Tools, die auch in TAPoR gelistet sind und die auch schon von vorhergehenden Lexikonan-

sätzen gefunden wurden (vgl. Barbot et al., 2019; Fischer & Moranville, 2020b, 2020a; Burghardt et al., 2022). In den restlichen 45,3% der Sequenzen finden sich allerdings neue Tools, also solche, die nicht schon in der sehr umfangreichen TAPoR-Liste dokumentiert sind. Mithilfe eines POS-Taggers wurden aus diesen Sequenzen Eigennamen gefiltert, um einen Überblick über die *neuen* Tools zu erhalten.

Es sind dies einerseits Programmiersprachen (*Java*, *SNOBOL4*, *Swift*, *Pascal*, *NetLogo*), Datenbanken (*SQL*) und Markupsprachen (*TEI*, *SGML*), aber auch diverse Belege aus dem Online-Bereich, etwa Social Media (*YouTube*, *Wikipedia*, *Facebook*, *Instagram*, ...) und Web Browser (*Netscape*, *Mozilla*) sowie auch diverse Beispiele aus dem Bereich des Desktop Publishing (*WordPerfect*, *Microsoft Word*, *WordStar*, *PageMaker*). Daneben finden sich zahlreiche weitere, teils antikierte Tools, die nicht ohne Weiteres zu größeren Gruppen zusammengefasst werden können, etwa: *TreeForm*, *Storyspace*, *MtScript*, *Seshat*, *FarsiTag*, *PlotVis*, *LexStat*, *Galgo*, *Neurolingo*, *AustLit*, *XyWrite*, *StoryTrek*, u.v.m.

Es zeigt sich also, dass unser Ansatz das Tool-Inventar bestehender Kataloge wie TAPoR deutlich erweitern kann, und über die Embeddings auch weitere, neue Tools gefunden werden.

Betrachtet man den jüngsten Aufruf (April 2022)² von TAPoR, bei dem Mitglieder der DH-Community gebeten werden, weitere Tools in TAPoR zu ergänzen, dann könnte unser Ansatz hier im großen Stil automatisiert weitere Vorschläge unterbreiten, indem systematisch verschiedene DH-Publikationen klassifiziert werden.

Ergebnisse der multi-class Toolkategorisierung

In einer zweiten Analyse wurden die Ergebnisse der automatischen Toolkategorisierung diachron ausgewertet (vgl. Abb. 1). Die Charts sind zur Paperanzahl pro Jahr im Datensatz normalisiert und zeigen einen zeitlichen Verlauf des Vorkommens der jeweiligen Kategorie.

Dabei fällt bspw. auf, dass (a) Tools für „Authoring / Annotation / Editing / Publishing“ ebenso wie (b) Tools für „Exhibition / Collection / Edition Platforms“ und (c) Kommunikationsplattformen alle erst Anfang der 2010er an Fahrt aufgenommen haben. Gleichzeitig zeigt sich im Falle der Programmiersprachen, dass diese in den Anfängen der DH, also den 1965er - 1985er Jahren, besonders populär waren, danach nimmt deren Nennung in Publikationen allerdings deutlich ab.

Tools für die (e) Textanalyse hatten ihren ersten Peak in der Mitte der 1970er, für etwa 10 Jahre, und dann nochmals besonders extrem in den frühen 2000ern und später wieder von 2010-2020. Für eine weitergehende Analyse dieser Konjunkturen ist ein close reading der entsprechenden Publikationen vonnöten, um so zu untersuchen, ob bspw. besonders populäre Einzeltools wie *Voyant* (Release 2003) für diese Spitzen verantwortlich sind. Spannend – und ggf. als Ursache für den Rückgang der Programmiersprachen zu interpretieren – ist weiterhin die steile Karriere von (f) Visualisierungstools, ab Anfang 2010.

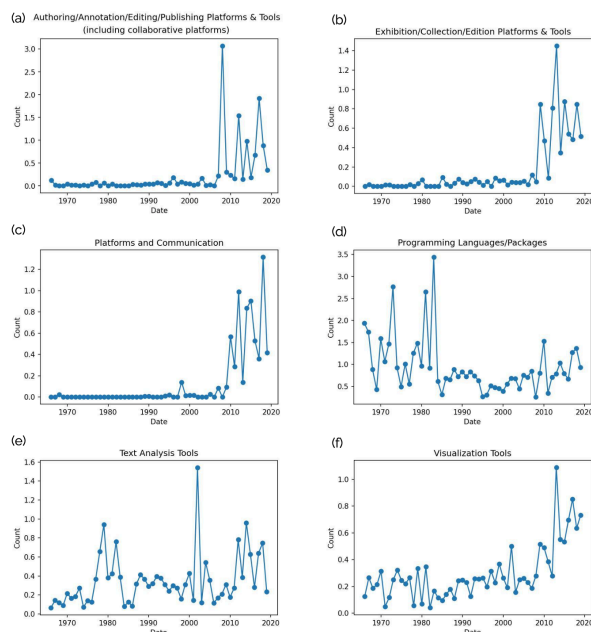


Abbildung 1. Diachrone Ergebnisse der multi-class Toolkategorisierung

Ausblick: Tool Studies 2.0

In diesem Beitrag haben wir dargestellt, dass es vielfältige Diskurse zur Rolle und Funktion von Tools in den DH gibt. Wir sehen großes Potenzial bei systematischen, empirischen Analysen von Tools, um die bestehenden Diskussionen zu ergänzen. Die bisherige, erste Welle der Tool Studies setzt primär auf lexikonbasierte Verfahren (vgl. Barbot et al., 2019; Burghardt et al., 2022; Fischer & Moranville, 2020b, 2020a). Vereinzelt wurden auch einfache Ansätze aus dem Bereich des maschinellen Lernens (vgl. Burghardt et al., 2022; Henny-Krahmer & Jettka, 2022; Zarei, Alireza et al., 2022) erprobt. Mit ersten Experimenten, die das große Potenzial von Transformer-Architekturen in Kombination mit Kontext-Embeddings aufzeigen, hoffen wir bestehende Tool Studies einen Schritt weiter zu bringen. Wir planen weitere Experimente, bei denen wir zum einen das Korpus erweitern wollen, aber auch weitere Modelloptimierungen vornehmen wollen. U.a. soll künftig ein Klassifikationsschema nach dem Vorbild von TaDiRAH trainiert werden. Neben einer diachronen Vermessung von Tooltrends in den Digital Humanities, soll in letzter Instanz mit einem ausreichend generalisierten Klassifikator ein großes interdisziplinäres Korpus nach dem Vorbild von Luhmann & Burghardt (2021) bezüglich der dort vorkommenden Tools analysiert werden. Ziel wird es sein aufzuzeigen, welche Tools ggf. von anderen Disziplinen in die DH importiert wurden, welche Tools aus den DH erfolgreich exportiert wurden, und welche Tools ggf. DH-spezifisch sind, und in keiner anderen Disziplin vorkommen.

Fußnoten

1. Epochen: 1.0; weitere Konfiguration: „train_batch_size“: 16, „eval_batch_size“: 64, „warmup_steps“: 500, „weight_decay“: 0.01.

2. Tweet: <https://twitter.com/tapordotca/status/1517564033519345664?s=21&t=Mj6Hk76pigAxGt-K8iaUuKA>

Bibliographie

Barbot, L., F. Fischer, Y. Moranville & I. Pozdniakov. 2019. “Which DH Tools Are Actually Used in Research?” In *Weltliteratur.Net – A Black Market for the Digital Humanities* (blog). <https://weltliteratur.net/dh-tools-used-in-research/>.

Borek, Luise, Quinn Dombrowski, Jody Perkins & Christof Schöch. 2016. “TaDiRAH: A Case Study in Pragmatic Classification.” *Digital Humanities Quarterly* 10, No. 1.

Bradley, John. 2019. “Digital Tools in the Humanities: Some Fundamental Provocations?” *Digital Scholarship in the Humanities* 34, No. 1: 13–20. https://academic.oup.com/HTTH/Handlers/Sigma/LoginHandler.ashx?error=login_required&state=395478f1-3ea2-4078-8eb8-b5c3bf51d898redirecturl%3Dhttpsazjzacademiczwoupzwcomzjdshzjarticlez34zj1zj13zj5063425.

Burghardt, Manuel, Jan Luhmann & Andreas Niekler. 2022. “Tools as Epistemologies in DH? A Corpus-Based Exploration.” Book of Abstracts of the ADHO Digital Humanities Conference. Tokyo.

Burghardt, Manuel & Christian Wolff. 2014. “Humanist-Computer Interaction: Herausforderungen Für Die Digital Humanities Aus Perspektive Der Medieninformatik.” Universität Regensburg. <https://doi.org/10.5283/EPUB.35716>.

Bush, Vannevar. 1945. “As We May Think.” *The Atlantic*.
Dalbello, Marija. 2011. “A Genealogy of Digital Humanities.” *Journal of Documentation* 67, No. 3: 480–506. <https://doi.org/10.1108/00220411111124550>.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. “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*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>.

Fischer, F. & Y. Moranville. 2020a. “DH Tools Mentioned in ‘The Programming Historian?’” In *Weltliteratur.Net – A Black Market for the Digital Humanities* (blog). <https://weltliteratur.net/dh-tools-programming-historian/>.

Fischer, F. & Y. Moranville. 2020b. “Tools Mentioned in DH2020 Abstracts.” In *Weltliteratur.Net – A Black Market for the Digital Humanities* (blog). <https://weltliteratur.net/tools-mentioned-in-dh2020-abstracts/>.

Henny-Krahmer, Ulrike & Daniel Jettka. 2022. “Softwarezitation Als Technik Wissenschaftskultur: Vom Umgang Mit in Den Digital.” In *DHd2022: Kulturen Des Digitalen Gedächtnisses. Konferenzabstracts*, 203–6. Potsdam.

Kaden, Ben. 2016. “Zur Epistemologie Digitaler Methoden in Den Geisteswissenschaften”, In *Ber-*

liner Beiträge zu Digital Humanities. <https://doi.org/10.5281/ZENODO.50623>.

Kittler, Friedrich. 1993. “Es Gibt Keine Software.” In *Draculas Vermächtnis: Technische Schriften*, 1st ed., 225–42. Leipzig.

Koolen, Marijn, Jasmijn van Gorp & Jacco van Ossenbruggen. 2019. “Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice.” *Digital Scholarship in the Humanities* 34, No. 2: 368–85. <https://doi.org/10.1093/llc/fqy048>.

Licklider, J. C. R. 1960. “Man-Computer Symbiosis.” *IRE Transactions on Human Factors in Electronics* HFE-1, No. 1: 4–11. <https://doi.org/10.1109/THFE2.1960.4503259>.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. <https://doi.org/10.48550/ARXIV.1907.11692>.

Lopez, Patrice & Laurent Romary. 2015. “GROBID - Information Extraction from Scientific Publications.” *ER-CIM News* 100, no. 100.

Luhmann, Jan & Manuel Burghardt. 2022. “Digital Humanities—A Discipline in Its Own Right? An Analysis of the Role and Position of Digital Humanities in the Academic Landscape.” *Journal of the Association for Information Science and Technology* 73, No. 2: 148–71. <https://doi.org/10.1002/asi.24533>.

Ramsay, Stephen & Geoffrey Rockwell. 2012. “Developing Things: Notes toward an Epistemology of Building in the Digital Humanities.” In *Debates in the Digital Humanities*. Minneapolis; London: University of Minnesota Press.

Schindler, David, Felix Bensmann, Stefan Dietze & Frank Krüger. 2022. “The Role of Software in Science: A Knowledge Graph-Based Analysis of Software Mentions in PubMed Central.” *PeerJ Computer Science* 14, No. 8. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8771769/>.

Traub, Myriam C. & Jacco van Ossenbruggen. 2015. “Workshop on Tool Criticism in the Digital Humanities.” CWI Techreport.

Unsworth, John. 2003. “Tool-Time, or ‘Haven’t We Been Here Already?’: Ten Years in Humanities Computing”.

Van Es, Karin, Maranke Wieringa & Mirko Tobias Schäfer. 2018. “Tool Criticism: From Digital Methods to Digital Methodology.” In *Proceedings of the 2nd International Conference on Web Studies*, 24–27. New York, NY, USA: Association for Computing Machinery, 2018. <https://doi.org/10.1145/3240431.3240436>.

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)*, 932–38. Tallinn, Estonia: IEEE. <https://doi.org/10.1109/EDUCON.2015.7096085>.

Zarei, Alireza, Yim Seung-Bin, Frank Fischer, Matej Ďurčo & Philipp Wieder. 2022. “Measuring the Use of Tools and Software in the Digital Humanities: A Machine-Learning Approach for Extracting Software Mentions from Scholarly Articles.” In *Book of Abstracts, ADHO DH Conference*. Tokyo.

Zarei, Alireza, Yim Seung-Bin, Matej Ďurčo, Klaus Illmayer, Laure Barbot, Frank Fischer & Edward Gray. 2022.

“Der SSH Open Marketplace: Kontextualisiertes Praxiswissen Für Die Digital Humanities.” In *DHd2022: Kulturen Des Digitalen Gedächtnisses. Konferenzabstracts*. Potsdam. <https://doi.org/10.5281/ZENODO.6327975>.

Understanding the impact of three derived text formats on authorship classification with Delta

Du, Keli

duk@uni-trier.de

Universität Trier, Deutschland

Introduction

Due to copyright law, Text and Data Mining (TDM) with copyrighted texts faces a lot of restrictions in terms of storage, publication and follow-up use of the resulting corpora, which, however, is against the spirit of open data in digital humanities (DH). As a solution to the problem, the concept of derived text formats (DTFs) has been suggested and discussed (see e.g., Lin et al. 2012, Bhattacharyya et al. 2015, Jett et al. 2020, Schöch et al. 2020). In DTFs, although some information (primarily copyright-relevant features) has been removed from the texts, the texts can still be used for various relevant TDM tasks in DH, such as authorship attribution or topic modeling. Schöch et al. (2020) also provides a very detailed examination of several DTFs from the perspectives of Computational Literary Studies, Computer Science, memory institutions and law. DTFs are extremely meaningful for the DH community, because they match the spirit of open data and make it possible for researchers and libraries to provide more text data for DH research. It also supports the pursuit of open science by encouraging researchers to publish their research data without worrying about violating copyright laws.

However, as far as I know, there is not much research dedicated to the question, how much the loss of information caused by DTFs affects the TDM results. Eder (2013) presented an empirical study of verifying the impact of unwanted noise in texts on authorship attribution and emphasizing that the usefulness of damaged texts should not be underestimated in stylometric studies. He brought noise into texts by (a) randomly replacing a portion of characters, (b) increasing standard deviation of word counts and (c) randomly replacing original words with other words in the same corpus, in order to show the correlation between a dirty corpus and the attribution accuracy. The presented paper did a similar empirical study

by transforming texts into token-based DTFs and provide a review on the correlation between information loss caused by these DTFs and the loss of accuracy in authorship classification.

Token-based DTFs

In Schöch et al. (2020), three kinds of token-based DTFs are introduced, to enable the free reusability of text data:

- Simple document-term-matrix: The idea is to transform a corpus into a matrix, which only saves the frequency of each term in each text in the corpus.
- Sequence randomization in segments: The idea is to split a text into segments, randomize the sequence of words in each segment, and reassemble all the segments into a text.
- Selectively reduced information on individual tokens: The idea is to replace a portion of the words (e.g., all the function words) in text with their POS-tags.

Applying the first and the second DTFs to frequency-based authorship attribution does not present any challenge. Take the most well-known method in authorship attribution Burrows's Delta (Burrows, 2022) as an example: Delta test follows the “bag of words” model for representing documents and only requires the frequency of each word in each text to distinguish between authors. The sequence information of words in texts is not necessary. Therefore, the first and the second DTFs keep all the information needed for the Delta test and the transformation does not affect the test results. As a matter of fact, if one only wants to publish a corpus so that the reported classification results of authorship could be verified, all one must do is to publish the document-term-matrix and the metadata table of the corpus.

In comparison, if the texts are transformed into the third DTF, although the frequency information of some words in the text will not match the original situation, the sequence information of words could be kept. This opens the possibility of using the data in this form for other TDM tasks such as sentiment analysis or named entity recognition that require the sequence information of words. If a corpus is published with the expectation that it can be applied to multiple TDM tasks, it makes more sense to prepare the corpus in this format. And of course, it is important to understand how much this format will affect the outcome of different TDM tasks. Therefore, this paper evaluates the usefulness of the third token-based DTF on authorship attribution as a start. In the next sections, the method and the results of the evaluation are reported.

Method

For the evaluation, three corpora representing different languages and text types have been constructed: deu_DraCor (German plays), fra_ELTeC (French novels) and eng_RSC (English journal articles). The relevant information about the corpora is shown in Table 1.

Table 1: Overview of the corpora.

corpus	corpus size (million words)	average text length (words)	no. of texts	no. of authors	period	language	text type
deu_DraCor (Fischer et al., 2019)	5.69	18237	312	55	1650 - 1928	German	play
fra_ELTeC (Odebrecht et al., 2021)	11.33	80370	141	30	1840 - 1912	French	novel
eng_RSC (Kermes et al., 2016)	7.92	6206	1276	69	1665 - 1869	English	journal article

The test is designed as follows: First, for each document in a corpus, a certain percentage of words (0%, 10%, ..., 100%) were randomly selected and replaced by their corresponding POS-tags. Since function words are crucial to authorship attribution, instead of only replacing function words as suggested in Schöch et al. (2020), any kind of word (including punctuation) may be replaced. The next step is the standard procedure for Delta test: creating a document-term-matrix, computing the z score of each value in it and then select the most frequent word types as feature to classify documents. The classification was done by a linear SVM classifier with 5-fold cross-validation. Following the results in Evert et al. (2017), the 2000 most frequent word types in each corpus were taken as feature for all the classification tasks. A further test was also performed: All POS-tags that replaced their corresponding words were removed from the text and the Delta test were then conducted again. By comparing the classification results of these two tests, we can also determine the contribution of POS-tags to the results of the classification. Since the words in texts are randomly replaced or removed which could introduce some random variation into the results, each of the above-described tests is repeated 10 times.

Before presenting the classification results, three text passages are prepared to give an impression of readability of the texts in DTF. The original text, the texts with 10% and 50% of words replaced by their corresponding POS-tags, are listed in Table 2.

Table 2. Text passages in original format and DTF (The percentage value indicates the proportion of words replaced or removed.).

percentage	text
0% (original)	The members of this new group of alkaloids are so numerous, their department is so singular, and their derivatives ramify in so many directions, that I have not as yet been able to complete the study of these substances in all their bearings; nor is it my intention to go fully into the chemistry of the subject in the present communication, my object being merely to establish the existence of these bodies, and to give a general outline of their connection with the volatile bases, and of their most prominent chemical and physical properties, reserving a detailed description of their salts and derivatives to a future memoir .
10%	The members of this new group of alkaloids are so numerous, their department is so singular, and their derivatives ramify in so many directions, that I have not as yet been able to complete the study of these substances in all their bearings; nor is it my intention to go fully into DT chemistry of DT subject in the present communication, PP\$ object being merely TO establish DT existence of these bodies, and to give a general outline of their connection IN the volatile bases, and of their most prominent chemical and physical properties, reserving a detailed description of their salts and derivatives to a JJ memoir.
50%	DT members IN DT JJ NN IN NNS VBP so JJ, PP \$ department is RB JJ, CC their NNS ramify in RB many NNS, WDT PP VHP RB RB yet been JJ TO VV DT study IN these NNS IN all PP\$ NNS : CC is it my NN TO go RB into DT chemistry IN the subject IN DT JJ communication, my object NN merely to VV DT NN of DT NNS, and TO VV DT JJ outline IN PP\$ connection IN DT volatile bases, CC IN their RBS prominent chemical and physical properties, reserving DT JJ NN IN their NNS and derivatives TO a future NN.

Results

The classification results on the German play collection, the English article collection, and the French novel collection are presented in Figures 1, 2 and 3, respectively. The y-axis is the F1(macro)-score, and the x-axis shows the portion of words that are replaced or removed. The blue boxplots and the yellow boxplots represent the classification results, when the words in texts are replaced with POS-tags or removed, respectively. As the reference value, the classification results for the original data are also shown in the figures. The Welch's t-test is also performed to determine the difference in classification results. The "ns" in the figures means non-significance.

In all the three figures, the same trend can be observed: Step by step, the median of F1-score distributions get worse as the percentage increases. Especially when more than half of the words were replaced or removed, the tendency for the classification results to become worse became particularly obvious. In addition, the variance of the F1-scores always becomes larger, if a certain percentage of words in texts are replaced or removed. According to the Welch's t-test, in all cases, whether the words are replaced or removed does not affect the classification. This observation indicates that the POS-tags do not contribute to the distinction of authorships.

Another interesting observation is, when all words in texts are replaced by POS-tags, the classification results improve, relative to a reduction of 90%, in the case of the German and English data, but not for the French data. To understand this situation, the change in the number of word types in each corpus was checked. As presented in Figure 4, when 90% of the words are replaced or removed, there are still around 20,000 word types in each text collection. But when all the words are replaced, only a few dozen types remain. Their number becomes so small

that it looks like it is reduced to zero in the Figure 4. Since the classification is based on the most frequent 2000 types, although 90% of the words are replaced or removed, the 2000 features used for classification are still mostly from the remaining 10% of words. In the German and English collections, these words bring apparently noise to the classification task. In contrast, the remaining 10% of words in the French corpus are still able to guarantee a relatively good classification result. From the data in Table 1 we can see that number of authors in the French corpus is smaller, which indicates the classification task on the French corpus is easier. More importantly, as presented in previous studies (e.g., Eder 2015, Romanov & Grallert 2022), that pulling random samples of at least 5000 words length out of texts will be sufficient for ensuring reliable authorship attribution. Considering the average length of the French novels is over 80,000 words, when 90% of the words are replaced or removed, the remaining 10% (that is, about 8000 words) is still sufficient to guarantee a good classification result. To clarify this issue, further research would certainly be of great interest.

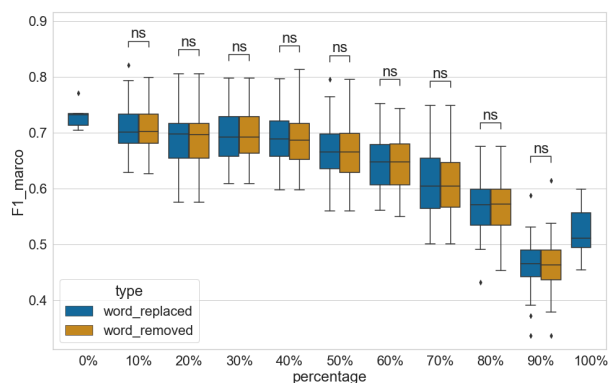


Figure 1. Authorship classification on the German play collection

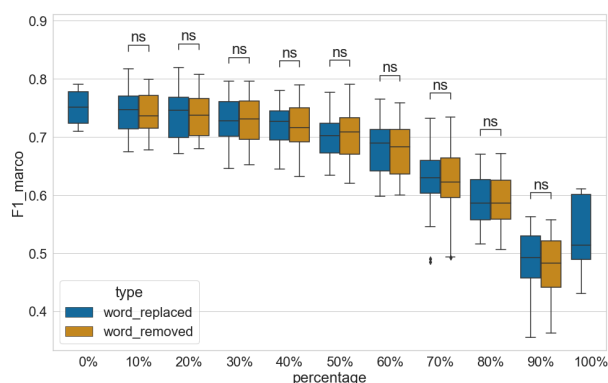


Figure 2. Authorship classification on the English article collection

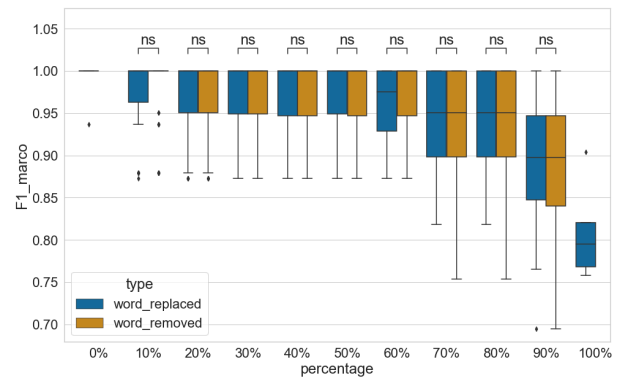


Figure 3. Authorship classification on the French novel collection

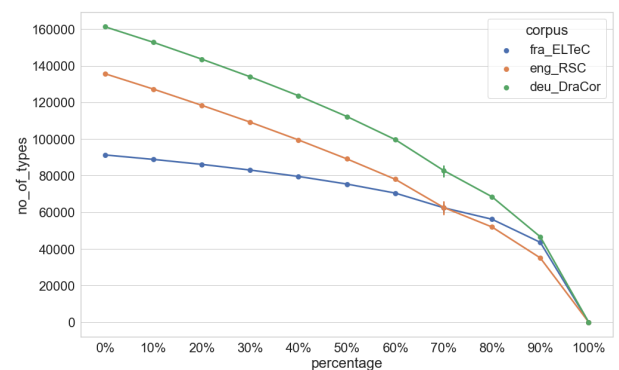


Figure 4. Change of the number of word types in three text collections

Conclusion

This paper provides an exploration of the usefulness of three token-based DTFs for frequency-based authorship classification with Delta. As presented, selectively reducing information on individual tokens could ensure, to a certain extent, that the authorship classification results are not affected too much. The impact of token-based DTFs on the results of Delta test can be reduced by considering only replacing or removing content words, while all function words remain unchanged. But this limits the application of the texts on other TDM tasks such as topic modeling. For the future work, a series of tests are planned on evaluating the usefulness of token-based DTFs on other TDM tasks. The goal is to find DTFs that could balance various factors (e.g. word frequency, sequence information, content vs. function words, copyright) so that texts could be published and used for as many TDM tasks as possible without violating copyright law.

Bibliographie

Bhattacharyya, Sayan, Peter Organisciak, und J. Stephen Downie. 2015. „A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features“. *Interdisci-*

plinary Science Reviews 40 (1): 61–77. <https://doi.org/10.1179/0308018814Z.000000000105>.

Burrows, John. 2002. „Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship“. *Literary and Linguistic Computing* 17 (3): 267–87. <https://doi.org/10.1093/lilc/17.3.267>.

Eder, Maciej. 2013. „Mind your corpus: systematic errors in authorship attribution“. *Literary and Linguistic Computing* 28 (4): 603–14. <https://doi.org/10.1093/lilc/fqt039>.

Eder, Maciej. 2015. Does size matter? Authorship attribution, small samples, big problem, *Digital Scholarship in the Humanities*, Volume 30, Issue 2, Pages 167–182. <https://doi.org/10.1093/lilc/fqt066>.

Evert, Stefan, Fotis Jannidis, Thomas Proisl, Steffen Pielström, Thorsten Vitt, Christof Schöch, und Isabella Reger. 2017. „Understanding and Explaining Distance Measures for Authorship Attribution“. *Digital Scholarship in the Humanities*. https://academic.oup.com/dsh/article-pdf/32/suppl_2/ii4/21298943/fqx023.pdf.

Fischer, F., Börner, I., Göbel, M., Hecht, A., Kittel, C., Milling, C. and Trilcke, P. 2019. Programmable corpora: Introducing DraCor, an infrastructure for the research on European drama. *Digital Humanities 2019*: 5 doi: [10.5281/zenodo.4284002](https://doi.org/10.5281/zenodo.4284002).

Jett, Jacob, Capitanu Boris, Kudeki Deren, Cole Timothy, Hu Yuerong, Organisciak Peter, Underwood Ted, Koehl Eleanor Dickson, Dubnick Ryan, Downie J. Stephen. 2020. „The HathiTrust Research Center Extracted Features Dataset (2.0)“. HathiTrust Research Center. <https://doi.org/10.13012/R2TE-C227>.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and Teich, E. 2016. The royal society corpus: From uncharted data to corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. pp. 1928–31.

Lin, Yuri, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, und Slav Petrov. 2012. „Syntactic Annotations for the Google Books N-Gram Corpus“. In *Proceedings of the ACL 2012 System Demonstrations*, 169–74. Jeju Island, Korea: Association for Computational Linguistics. <https://aclanthology.org/P12-3029>.

Odebrecht, C., Burnard, L. and Schöch, C. 2021. European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels. Zenodo doi: [10.5281/ZENODO.4662444](https://doi.org/10.5281/ZENODO.4662444). <https://zenodo.org/record/4662444> (accessed 9 December 2022).

Romanov, Maxim, Grallert, Till. 2022. ‘Establishing Parameters for Stylometric Authorship Attribution of 19th-Century Arabic Books and Periodicals’. *DH2022*, Tokyo, 23 July 2022. <https://dh-abstracts.library.virginia.edu/works/11858>.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen“. *Zeitschrift für digitale Geisteswissenschaften (ZfdG)* 5. http://dx.doi.org/10.17175/2020_006.

Vom Heben verborgener Schätze – Literarische Blogs als Ressource

Schenk, Nicolas

nicolas.schenk@dla-marbach.de
Deutsches Literaturarchiv Marbach

Blessing, André

andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Hein, Pascal

pascal.hein@ilw.uni-stuttgart.de
Universität Stuttgart, Institut für Literaturwissenschaft

Hess, Jan

jan.hess@dla-marbach.de
Deutsches Literaturarchiv Marbach

Jung, Kerstin

kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Schlesinger, Claus-Michael

claus-michael.schlesinger@ilw.uni-stuttgart.de
Universität Stuttgart, Institut für Literaturwissenschaft

30 Mio. Token, 140.000 Blogposts, über 200 aufbereitete Blogs

Bereits seit Ende der 90er Jahre gewannen Weblogs als Medium zur öffentlichen Darstellung unterschiedlicher Themen und Inhalte immer mehr an Popularität. Findige Literatur- und Kulturschaffende zögerten nicht lange, um das neue Medium auch für literarische Zwecke umzunutzen.¹ Blogs wie *Die Dschungel*. *Anderswelt*² von Alban Nikolai Herbst, *Abfall für alle*³ von Rainald Goetz, Wolfgang Herrndorfs *Arbeit und Struktur*⁴ oder das kollaborativ betriebene Blog *Die Riesenmaschine*⁵ werden seither zwar regelmäßig als Gegenstand wissenschaftlicher Untersuchungen herangezogen (vgl. Fassio 2021, Giacomuzzi 2008, Knapp 2014, Knapp 2012), wie viele andere Formen von Literatur im Netz im Vergleich zu ihren genuin analogen Pendanten jedoch immer noch durchaus

stiefmütterlich behandelt. Hinzu kommt, dass sich die Verfasser:innen dieser Blog-Studien in erster Linie klassisch hermeneutisch-literaturwissenschaftlicher Methoden bedienen und nur in seltenen Fällen auf computergestützte Analysemethoden und -werkzeuge zurückgreifen (vgl. zuletzt: Fassio 2021), obwohl die Weblogs schon ihrem Begriff nach born-digital sind und damit eine ‚digitale‘, computergestützte Form der Analyse zunächst auf der Hand zu liegen scheint.⁶

In dem Beitrag zur Jahrestagung der DHd 2022 (Blesing et al 2022) haben die Autor:innen des vorliegenden Beitrags bereits am Beispiel des u. a. von der Autorin Kathrin Passig ins Leben gerufenen Techniktagebuch⁷ verschiedene computergestützte Möglichkeiten sowie geeignete Werkzeuge für die Analyse literarischer Blogs vorgestellt. Der in der End-Anwendung später kaum sichtbare Aufwand an textuellen Preprocessing-Schritten, der bereits an diesem vermeintlich einfachen Fallbeispiel offenbar wurde, gepaart mit den Reaktionen und dem großen Interesse aus der wissenschaftlichen Community an der grundlegenden Methodik zum Umgang mit WARC-Dateien, lassen bereits die Gründe erahnen, weshalb in der Blog-Forschung digitale Methoden bislang vergleichsweise selten zum Einsatz kamen. Im hier nun vorliegenden Beitrag sollen – als Fortsetzung und Erweiterung des vorangehenden Beitrags auf der DHd 2022 – nicht nur exemplarisch ebendiese Herausforderungen, sondern vielmehr auch Lösungsmöglichkeiten im Umgang mit (archivierten) Blogs aufgezeigt werden. Im Fokus der Untersuchung steht dabei nicht mehr nur ein einzelnes Blog, sondern vielmehr ein insgesamt aus über 200 aufbereiteten Blogs mit ca. 140.000 Blogposts und 30 Millionen Token bestehender Fundus literarischer Blogs. Eben solche literarische Weblogs sammelt die Bibliothek des Deutschen Literaturarchivs Marbach neben literarischen Zeitschriften und Objekten der Netzliteratur bereits seit 2008.⁸ Bislang erfolgt die Bereitstellung dieser Sammlungsobjekte über die Plattform Literatur-im-Netz⁹, 2023 werden diese zusammen mit der hier vorgestellten Ressource über das SDC4Lit-Repository bereitgestellt.¹⁰

Bausteine von Weblogs

Typische Bausteine in Blogs sind Blogposts als (meist datierte) Inhaltseinheiten verschiedenster Länge und unter Verwendung unterschiedlicher Modalitäten wie Text, Bild, Animation, Video, Referenz (z. B. Hyperlinks), etc. Weiterhin finden sich auf der Ebene der Blogposts oft Kommentarformulare und Kommentare sowie eine Etikettierung von Einträgen in Form von Tags, die der Gruppierung und Beschreibung der Einträge dienen und sich auf Themen, Stimmungen, Autoren, etc. des Blogposts beziehen können (vgl. zu den Bausteinen von Blogs: Ernst 2010, 286f.). Zusätzliche Elemente wie Übersichtsseiten, die als Einstiegsseiten der jeweiligen Blogs die einzelnen Blogposts (zusätzlich) in einer bestimmten Reihenfolge auflisten, oder Archivseiten, die den Nutzer:innen einen Zugang zu älteren Blogposts ermöglichen sollen, prägen die Struktur der Blogs und damit ggf. deren Analyse.

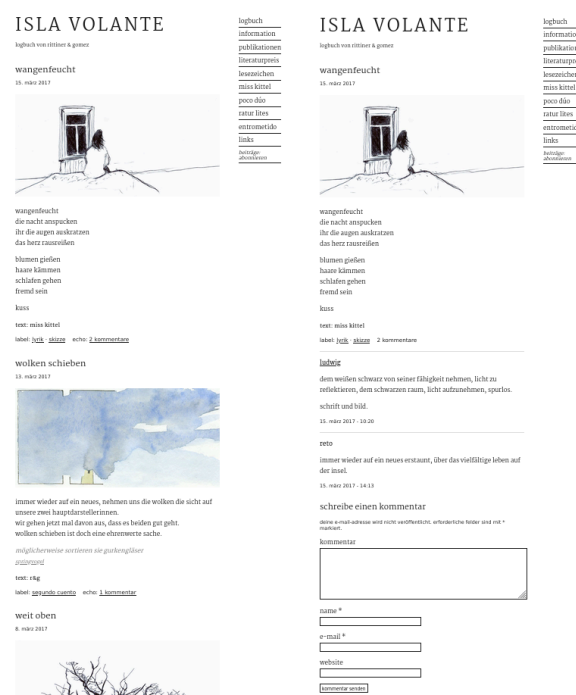


Abb.1: Links: Übersichtsseite (Home) mit mehreren Blogposts. Rechts: Seite des ersten Blogposts mit Tags, Kommentaren und Kommentarfunktion. Beispiele aus "Logbuch Isla volante": Bilder und Texte von der Insel". Spiegelung 2017.03.17_01, URN: urn:nbn:de:bsz:mar1-dd001-fe31dc38-7c43-4da2-ae3d-fd9619f88ea45.

Die Spiegelungen der Blogs, die am DLA durchgeführt werden, erfolgen zunächst mit einem Crawling-Vorgang, der der Hyperlinkstruktur im Blog folgt und die clientseitig (wie durch einen Browser) empfangenen Daten gemeinsam mit einigen Metadaten zum Crawlingprozess im Web Archive- oder kurz: WARC-Format ablegt. Das aus dem ARC-Format des Internet Archive weiterentwickelte Archivformat hat sich inzwischen als internationaler Standard für die Archivierung von Webinhalten etabliert (IIPC, n.d.). Beim Crawling können Inhalte, die ggf. nicht Teil des Blogs sind, wie z. B. zufällig eingebundene Werbeanzeigen oder externe Inhalte, auf die von diesen Werbeanzeigen verwiesen wird, Teil des Archivobjekts werden. Aber auch bezüglich der tatsächlichen Blog-Elemente sind im Archivobjekt Inhalte (z. B. bestimmte Textpassagen) so oft abgelegt, wie der Crawler ihnen auf verschiedenen Seiten begegnet ist. Der Inhalt eines Blogposts kann an verschiedenen Stellen für den Crawler erreichbar sein: auf der Übersichtsseite, auf der eigenen Seite des Posts, auf der Seite jedes Schlagworts, mit dem der Post versehen ist, im Archiv, etc. Aber auch Textpassagen aus Strukturelementen wie Kopf- und Fußzeilen führen zu mehrfachen Vorkommen von Textpassagen oder Begriffen.

Werkzeuge, die Strukturelemente ausblenden und (Text-)Duplikate erkennen, sind daher bei der Vorverarbeitung der Daten für Analysen notwendig, arbeiten oft aber statistisch und müssen ggf. auf jedes zu untersuchende Blog neu angepasst werden, was im Spannungsfeld mit dem maschinell unterstützten Distant Reading steht. Im Folgenden wird daher das Vorgehen bei der Aufbereitung der Blogs beschrieben, das notwendig ist,

Tab. 1: Übersicht der resultierenden Ressourcengrößen nach der schrittweisen Aufbereitung für vier Beispielblogs.

Blog	WARC-Records	HTML-Seiten	Blogposts	Token
Die.Dschungel.Anderwelt ¹⁶	123.859	65.014	14.106	7,8 Mio
Techniktagebuch ¹⁷	310.131	127.532	5.233	1,6 Mio
Lux autummalis ¹⁸	11.121	6.426	3.092	1,3 Mio
Henrikes Tagebuch ¹⁹	499.878	168.856	861	0,2 Mio

Evaluation

Die Zahlen zur Ressourcengröße aus Tabelle 1 veranschaulichen, dass die Aufbereitung der Blogs nicht vollständig manuell validiert werden kann. Trotzdem ist eine Aussage über die Qualität der aufbereiteten Blog-Daten wichtig. Zu diesem Zweck wurde ein Annotations-Jupyter-Notebook entworfen, mit dessen Hilfe automatisch aus den jeweiligen CMS-Familien wie beispielsweise Wordpress oder Blogger repräsentative Datenmengen entnommen werden, die anschließend von mehreren Annotator:innen bewertet werden. Im ersten Schritt geht es um das Erkennen der Blog-Posts in Abgrenzung zu Übersichtsseiten oder weiteren Seiten wie beispielsweise dem Impressum. Wenn es sich bei der Seite um einen Post handelt, dann wird im zweiten Schritt die Qualität der Metadaten- und Textextraktion bewertet. Dabei wird u. a. darauf geachtet, ob Datum und Überschrift richtig extrahiert wurden und ob der extrahierte Text vollständig ist oder beispielsweise Teile aus der Seitennavigation enthalten sind.

Es hat sich gezeigt, dass die beschriebene manuelle Bewertung auch für Menschen nicht immer trivial ist. Daher ist der Annotationsprozess aktuell noch nicht abgeschlossen. Die Evaluationsergebnisse werden jedoch mit dem Release der Daten bereitgestellt.

Nutzungsszenarien für die aufbereiteten Blog-Daten

Nach den oben beschriebenen Aufbereitungsschritten sind die daraus resultierenden Daten (nahezu) frei von Dubletten und Fremdinhalten, sodass sie für verschiedene Formen der Textanalyse sowie für weitere Prozessierungsschritte verwendet werden können. Dazu gehören korpuslinguistische Untersuchungen, wie sie mit CQPweb durchgeführt werden können, das Erstellen von Sprachmodellen (z. B. embeddings, topic-models) oder automatische Keyword-Erkennung. Auch die Untersuchung der in den Blogposts enthaltenen Metadaten (Datum, Autor:in, Verschlagwortung) kann für die Forschung von Interesse sein. Beispielhaft erwähnt sei an dieser Stelle die Verschlagwortung der einzelnen Blogs durch die jeweiligen Blog-Autor:innen, die mit End-Anwendungen wie Keshif²⁰ erst auf Basis der aufbereiteten Blog-Daten überhaupt sinnvoll visualisiert und analysiert werden können (vgl. Abb. 4).

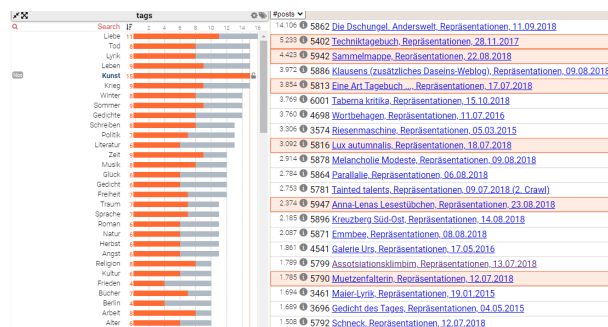


Abb. 4: Screenshot aus einer Keshif-Testinstanz. Rechts werden alle Blogs hervorgehoben, in denen das Schlagwort Kunst vergeben wurde, links befindet sich eine Liste aller Schlagworte.

In Blessing et al (2022) wurde am Fallbeispiel des Techniktagebuchs bereits exemplarisch aufgezeigt, welche komplexeren computergestützten Analysen durch die Verwendung der extrahierten Text- und Metadaten vorgenommen werden können. Unter anderem wurde bereits untersucht, welche Zusammenhänge zwischen Inhalt – repräsentiert durch automatische Keyword-Erkennung und die Verschlagwortung durch die Autor:innen – und Form bzw. Sprache erkennbar werden.

Schätze heben und nutzen

30 Millionen Zeichen, 140.000 Blogposts, über 200 Blogs: Schon wegen seiner Größe ist das hier vorgestellte, aufbereitete Korpus für viele Bereiche der Digital Humanities, beispielsweise die Computational Literary Studies, die Digital History oder für NLP-Untersuchungen, eine wichtige Quelle für Inhaltsanalysen oder das Trainieren von Sprachmodellen. Wie in diesem Beitrag gezeigt stellt vor allem die Struktur der Weblogs eine Herausforderung dar, enthalten die WARC-Dateien, in denen die Blogs zunächst vorliegen, doch sehr viel Redundantes, das für eine Vielzahl von Inhaltsanalysen nicht nur uninteressant, sondern sogar hinderlich ist. Mit den im Zuge der Veröffentlichung der SDC4Lit-Plattform 2023²¹ zur Verfügung gestellten Blog-Daten in aufbereiteter Form wird eine robuste Ressource bereitgestellt, die neben den Rohdaten im WARC-Format auch das bereinigte Textkorpus in Form der inhaltlich relevanten Blogposts sowie die zugehörigen Metadaten zu jedem Post enthält. Dank der ebenfalls über SDC4Lit zur Verfügung gestellten WARC-Volltextsuche und WARC-Player wie SolrWayback können die Blogs bzw. die einzelnen Blogposts zudem – unabhängig von allen weiteren Analyseschritten – möglichst originalgetreu in ihrer ursprünglichen Repräsentationsform angesehen und erforscht werden, auch wenn die originalen Webseiten bereits nicht mehr vorhanden sind oder geändert wurden. Die Implementierung der Aufbereitung wird in Form von dokumentierten Jupyter-Notebooks bereitgestellt, dank der auch weitere, über das hier präsentierte Korpus hinausgehende (literarische) Blogs aufbereitet und damit für weitere DH-Bereiche zugänglich gemacht werden können, sodass die nunmehr bereits gehobenen Weblog-Schätze künftig nicht die einzigen bleiben.

Fußnoten

1. In Bezugnahme auf Ernst (2010, 294 –297); Giacomuzzi (2012, 183) und Jürgensen (2011, 407) liefert Fassio (2021, 97f.) eine ausführliche Diskussion der bisherigen Definitions- und Typisierungsversuche literarischer Weblogs.
2. <https://dschungel-anderswelt.de/> , (zugegriffen: 14. Dezember 2022).
3. Das Blog erschien zunächst in einer Folge von 343 Blogposts auf rainaldgoetz.de (Quelle offline), anschließend auch in Buchform: Goetz, Rainald. 2015. "Abfall für alle. Roman eines Jahres." Frankfurt a. M.: Suhrkamp.
4. <https://www.wolfgang-herrndorf.de/> , (zugegriffen: 14. Dezember 2022) ; auch als Print-Ausgabe: Wolfgang Herrndorf. 2015. "Arbeit und Struktur." Rowohlt: Reinbek.
5. <http://riesenmaschine.de/> , (zugegriffen: 14. Dezember 2022).
6. Neben hermeneutischen Untersuchung en einzelner literarischer Blogs (Ainetter 2006, Knapp 2012, 2014) oder theoretischer bzw. struktureller Überlegungen (Ernst 2010) stellt ein Großteil der Studien zu (literarischen) Weblogs die Beziehung bzw. Abgrenzung verwandter Textsorten wie Flugblatt, Zeitung, Autobiographie oder – am häufigsten – dem Tagebuch in den Fokus (Augustin 2015, Flüh 2017, Jürgensen 2011, Michelbach 2019).
7. <https://techniktagebuch.tumblr.com/> , (zugegriffen: 14. Dezember 2022).
8. Das Deutsche Literaturarchiv Marbach sammelt deutschsprachige Literatur von 1750 bis zur Gegenwart, sodass unser Korpus vorwiegend aus deutschsprachigen Blogs besteht. Allerdings finden sich in unserer Sammlung auch an manchen Stellen nicht-deutschsprachige Absätze oder gar ganze Blogbeiträge. Die Technik zur Extraktion der Texte lässt sich zu weiten Teilen auch auf andere Sprachen übertragen, sodass die entwickelte Pipeline auch für andere Sprachen vollständig oder zumindest größtenteils nachnutzbar ist.
9. <http://literatur-im-netz.dla-marbach.de> , (zugegriffen: 14. Dezember 2022).
10. Im Rahmen des Projekts *SDC4Lit – Aufbau eines nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung* (<https://www.sdc4lit.de/> , zugegriffen: 14. Dezember 2022) entsteht seit 2019 die SDC4Lit-Plattform, über die die gesammelten Blogs zusammen mit den sonstigen Beständen von Literatur im Netz nicht nur in einem Repositorium verfügbar gemacht, sondern auch (explorative) Zugangs- und Analysemöglichkeiten aufgezeigt und bereitgestellt werden.
11. Bei einigen Blog-Hostern stehen APIs zur Verfügung, über die der Download einer inhaltsorientierten Version der Blogs möglich ist. Diese Repräsentationen sind bisher nicht Teil der Sammlung.
12. Der ClueWeb09-Datensatz umfasst 1.040.809.705 Webseiten. <http://lemurproject.org/clueweb09/> (zugegriffen: 14. Dezember 2022).
13. <https://github.com/adbar/trafilatura> , (zugegriffen: 14. Dezember 2022), eine Python-Bibliothek zur Boilerplate-Entfernung und Metadatenerkennung (Barbaresi 2019).
14. <https://cqpwweb.lancs.ac.uk> , (zugegriffen: 14. Dezember 2022).

15. <https://github.com/netarchivesuite/solrwayback> , (zugegriffen: 14. Dezember 2022).
16. <https://dschungel-anderswelt.de> , (zugegriffen: 14. Dezember 2022).
17. <https://techniktagebuch.tumblr.com> , (zugegriffen: 14. Dezember 2022).
18. <http://www.luxautumnalis.de> , (zugegriffen: 14. Dezember 2022).
19. <http://henrikeheiland.blogspot.de/> , (zugegriffen: 14. Dezember 2022).
20. <https://github.com/adilyalcin/Keshif> , (zugegriffen: 14. Dezember 2022).
21. <https://www.sdc4lit.de/> , (zugegriffen: 14. Dezember 2022).

Bibliographie

- Ainetter, Sylvia. 2006. "Blogs - literarische Aspekte eines neuen Mediums. Eine Analyse am Beispiel des Weblogs Miagolare". Wien: Lit Verlag.
- Augustin, Elisabeth. 2015. "BlogLife. Zur Bewältigung von Lebensereignissen in Weblogs." Bielefeld: transcript.
- Barbaresi, Adrien. 2019. "Generic Web Content Extraction with Open-Source Software". In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*: 267–268.
- Blessing, André, Jan Hess und Kerstin Jung. 2022. "Ja, jetzt ist das langweilig. Aber in zwanzig Jahren!" - Bereitstellung, Zugang und Analyse literarischer Blogs am Beispiel des Techniktagebuchs." *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum"* (DHd 2022), Potsdam. Zenodo: <https://doi.org/10.5281/zenodo.6322488>; <https://doi.org/10.5281/zenodo.6328029>.
- Cormack, Gordon V., Mark D. Smucker und Charles L. A. Clarke. 2011. "Efficient and effective spam filtering and re-ranking for large web datasets." In *Information retrieval 14*: 441–465.
- Ernst, Thomas. 2010. "Weblogs. Ein globales Medienformat." In *Globalisierung und Gegenwartsliteratur. Konstellationen - Konzepte - Perspektiven*, hg. von Wilhelm Amann, Georg Mein und Rolf Parr, 281–302. Heidelberg: Synchron Wissenschaftsverlag der Autoren.
- Fassio, Marcella. 2021. "Das literarische Weblog. Praktiken, Poetiken, Autorschaften". Bielefeld: Transcript.
- Flüh, Thorsten. 2017. "Flugblatt – Zeitung – Blog. Materialität und Medialität als Literaturen." Wien: Passagen Verlag.
- Giacomuzzi, Renate. 2008. "Die ‚Dschungel. Anderswelt‘ und A. N. Herbsts ‚Poetologie des literarischen Bloggens‘." *Die Horen* 53: 137–149.
- Giacomuzzi, Renate. 2012. "Deutschsprachige Literaturmagazine im Internet. Ein Handbuch." Innsbruck: Studien-Verlag.
- 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. n. d. "The WARC Format." <https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/> (zugegriffen: 14. Dezember 2022).
- Jürgensen, Christoph. 2011. "Ins Netz gegangen – Inszenierungen von Autorschaft im Internet am Beispiel

von Rainald Goetz und Alban Nikolai Herbst." In *Schriftstellerische Inszenierungspraktiken – Typologie und Geschichte*, hg. von Christoph Jürgensen und Gerhard Kaiser, 405–422. Heidelberg: Winter.

Knapp, Lore. 2012. "Christoph Schlingensiefels Blog. Multimediale Autofiktion im Künstlerblog." In *Narrative Genres im Internet: Theoretische Bezugsrahmen, Mediengattungstypologie und Funktionen*, hg. von Ansgar Nünning und Jan Rupp, 117–132. Trier: Wissenschaftlicher Verlag.

Knapp, Lore. 2014. "Künstlerblogs. Zum Einfluss der Digitalisierung auf literarische Schreibprozesse (Goetz, Schlingensiefel, Herrndorf)." Berlin: Ripperger & Kremers.

Manning, Christopher D., Prabhakar Raghavan und Hinrich Schütze. 2010. "Introduction to information retrieval." In *Information retrieval* 13: 192–195. <https://doi.org/10.1007/s10791-009-9115-y>.

Michelbach, Elisabeth. 2019. "Poetik des autobiografischen Blogs." Dissertation, Universität Göttingen.

Vom sprachlichen Indikator zum komplexen Phänomen?

Jacke, Janina

janina.jacke@uni-goettingen.de

Georg-August-Universität Göttingen, Deutschland

Zum Verhältnis zwischen computationeller und traditioneller Literaturwissenschaft

Operationalisierungsprobleme in der computationellen Literaturwissenschaft am Beispiel des unzuverlässigen Erzählens

Im Feld der Digital Humanities hat sich die computationelle Literaturwissenschaft als Teildisziplin etabliert. Von literaturwissenschaftlicher Seite wird allerdings immer wieder angemahnt, die computationelle Auseinandersetzung mit Literatur sei zu reduktionistisch – unter anderem weil Textanalysen nur statistisch-deskriptiv und weitgehend kontextfrei möglich seien (vgl. Gius/Jacke 2022). Derartigen Bedenken lässt sich auf unterschiedliche Weise begegnen. Eine Möglichkeit besteht in der Akzeptanz, dass die computationelle und die traditionelle Literaturwissenschaft schlichtweg unterschiedliche Fragen an Texte stellen. Soll die Idee eines Brückenschlags zwischen traditioneller und computationeller Literaturwissenschaft dagegen nicht verworfen werden, gibt es zwei weitere Möglichkeiten: Es kann der ambitionierte Versuch unternommen werden, auch für komplexere literaturwissenschaftliche Fragestellungen (vollständig) computationelle Lösungen zu finden. Solche Versuche sind zwar wünschenswert, sollten aber

(durch den großen Aufwand und die oft eingeschränkten Erfolgsaussichten) nicht die einzige Möglichkeit darstellen, computationelle Modellierung für traditionelle literaturwissenschaftliche Fragen fruchtbar zu machen.

Der vorliegende Beitrag stellt eine andere Möglichkeit vor, wie ein Brückenschlag aussehen kann: Es kann es fruchtbar sein, dezidiert nur eine computationelle Teil-Operationalisierung komplexer literarischer Phänomene anzustreben und möglichst genau zu ergründen und zu explizieren, welcher Status den entwickelten computationellen Analysemethoden im Zusammenhang mit komplexeren literarischen Phänomenen und Fragestellungen zukommt. Computationelle Modelle können deskriptiv-quantitativ Textmerkmale feststellen, die als Indikatoren für komplexere literaturwissenschaftlich interessante Phänomene verstanden werden können. Sinnvoll verwertbar sind solche Analysen, wenn das Indikationsverhältnis (also die genaue Beziehung zwischen Indikator und komplexem Phänomen) spezifiziert wird. Dabei geht es zum einen um die (aus literaturwissenschaftlicher Perspektive nachvollziehbare) *Erklärung der Indikationsbeziehung* („Warum indiziert das sprachliche Merkmal das komplexe literarische Phänomen?“), zum anderen um die *Bestimmung der Indikationskraft* („Wie eng ist der Zusammenhang zwischen Indikator und komplexem Phänomen?“). Für beide Fragen kann die Identifikation und Analyse von 'Zwischenphänomenen' sinnvoll sein – also von Phänomenen mittlerer Komplexität, die (inhaltlich-logisch) als Verbindungsglieder zwischen einfachem sprachlichen Indikator und komplexem literarischen Phänomen zu verstehen sind.

Die vorgeschlagenen Ideen zur Spezifikation des Indikationsverhältnisses basieren auf ersten Erkenntnissen aus dem Versuch der Operationalisierung und computationellen Teil-Modellierung des literaturwissenschaftlichen Konzepts „unzuverlässiges Erzählen“ im Rahmen des Projekts CAUTION.¹ U nzuverlässiges Erzählen liegt dann vor, wenn der Erzählfigur einer fiktionalen Erzählung insofern ‚nicht zu trauen‘ ist, als ihre Aussagen über die fiktive Welt teilweise falsch sind oder relevante Aussagen fehlen bzw. die moralischen Ansichten der Erzählfigur in Diskrepanz zu durch das Werk vermittelten Werten stehen (vgl. Kindt 2008, 53). In diesem Beitrag wird unzuverlässiges Erzählen als illustratives Beispiel angeführt. Da die präsentierten Ideen zur Spezifikation des Indikationsverhältnisses auch im Rahmen anderer Vorhaben genutzt werden können, in denen komplexe literaturwissenschaftliche Konzepte (teilweise) computationell modelliert werden, versteht sich dieser Ansatz als theoretischer Beitrag, der zur Stärkung der Brücke zwischen traditioneller und computationeller Literaturwissenschaft beitragen möchte.

Im Folgenden sollen zunächst kurz die für den vorgestellten Ansatz zentralen Begriffe „Operationalisierung“ und „Komplexität“ diskutiert werden (Abschnitt 2). Im Anschluss wird die vorgeschlagene Analyse des Indikationsverhältnisses genauer expliziert (Abschnitt 3), die auf einer theoretischen Analyse des Unzuverlässigkeitskonzepts sowie ersten Annotations- und Analyseerfahrungen im Rahmen von CAUTION basiert.

Begriffsklärung: Operationalisierung und Komplexität

Die Operationalisierung geisteswissenschaftlicher Konzepte für die computationale Textanalyse ist bereits zum Thema extensiver Auseinandersetzung geworden (vgl. Moretti 2013, Döring/Bortz 2016, Pichler/Reiter 2021, Krautter et al. im Erscheinen). Unter der Operationalisierung von Begriffen ist die Angabe von Handlungsschritten zu verstehen, die ausgeführt werden müssen, um das Phänomen identifizieren (bzw. messen und quantifizieren) zu können. Im Feld der Digital Humanities wurde der Begriff von Moretti eingeführt (vgl. Moretti 2013). Seine literaturwissenschaftlichen Beispiele zeigen aber, dass die operationalisierten Modelle oft nicht oder nur lose an die zu operationalisierenden literaturwissenschaftlichen Konzepte angeknüpft sind (vgl. Krautter/Pichler/Reiter im Erscheinen). Es liegt der Schluss nahe, dass viele literaturwissenschaftliche Konzepte zu komplex sind, um sich unter Erhaltung ihrer ursprünglichen Bedeutung und Funktion vollständig und eindeutig operationalisieren bzw. computationell modellieren zu lassen. Dennoch scheinen die bisherigen Beiträge nicht grundsätzlich von dem Ziel Abstand zu nehmen, mit computationellen Operationalisierungen vollständige Übersetzungen bzw. direkte Entsprechungen komplexer Konzepte zu entwickeln. Reduzierte Ansprüche lassen sich lediglich insofern feststellen, als im Falle umstrittener Konzepte eine begründete Auswahl aus dem Definitionsangebot (vgl. Döring/Bortz 2016, 226) oder möglicher Kontexte (vgl. Pichler/Reiter 2021, 6) getroffen wird. Grundsätzlich wird aber davon ausgegangen, dass durch die Aufgliederung in Teilschritte (vgl. Pichler/Reiter 2021, 19–23) bzw. die Kombination unterschiedlicher Indikatoren (vgl. Döring/Bortz 2016, 229) die entwickelten Modelle direkt auf die komplexen geisteswissenschaftlichen Phänomene zielen. Es wird nicht in Betracht gezogen, dass es – um Reduktionismus zu vermeiden – notwendig oder sinnvoll sein könnte, mit dem computationellen Modell dezidiert nur einen Baustein zu ihrer Analyse beizutragen und dessen genaue Funktion zu explizieren. Das Kriterium der Validität (vgl. Drost 2011) computationeller Modelle wäre aus dieser Perspektive neu zu denken: Es ist nicht notwendig, dass die Modelle dasjenige messen, nach dem Literaturwissenschaftler:innen fragen – sofern die Relation zwischen Modell und Phänomen so expliziert wird, dass ein (nicht-computationelles) literaturwissenschaftliches Weiterarbeiten mit den erzielten Ergebnissen ermöglicht wird (vgl. hier auch Flick 2012 zu Triangulation).

Aber warum bzw. inwiefern kann die Komplexität literaturwissenschaftlicher Phänomene es notwendig oder sinnvoll machen, von einer vollständigen computationellen Modellierung abzusehen? In diesem Zusammenhang sind unterschiedliche Dimensionen der Komplexität literarischer Phänomene (bzw. der literaturwissenschaftlichen Konzepte, die diese Phänomene fassen sollen) zu beachten. Literarische Phänomene können zum einen dadurch komplex sein, dass sie zusammengesetzt sind (vgl. Gius 2019). In solchen Fällen müssen mehrere Teilphänomene vorliegen bzw. untersucht werden, um Aussagen über das komplexe Gesamtphänomen treffen zu können. Eine Operationalisie-

rung, die so ein zusammengesetztes Phänomen unter Erhaltung der literaturwissenschaftlichen Bedeutung in Teilschritte zur Erkennung übersetzen will, ist deshalb schwierig, weil die einzelnen Teilphänomene und ihre Beziehung zueinander identifiziert und diese dann jeweils operationalisiert und computationell modelliert werden müssen. Da literaturwissenschaftliche Konzepte häufig nicht derart analytisch aufbereitet sind, wäre die Rekonstruktion und behutsame Schärfung einer solchen Definition (vgl. Carnap 1959 zu Explikation), ebenso wie die Überprüfung, ob diese Definition mit der tatsächlichen Verwendung des Konzepts in der Literaturwissenschaft kompatibel ist, eine Aufgabe, die im Rahmen der Operationalisierung stattfinden müsste.

Unzuverlässiges Erzählen ist insofern zusammengesetzt, als es in distinkte Typen mit unterschiedlichen Eigenschaften zerfällt (vgl. Jacke 2020, 17–57). Zudem muss (zumindest laut einigen Definitionen) eine Kombination aus mehreren (Text-)Eigenschaften gegeben sein, damit unzuverlässiges Erzählen vorliegt. Diese Probleme ließen sich allerdings noch durch begründete Auswahl und ein Zergliedern in Analyseschritte adressieren, wie von Döring/Bortz bzw. Pichler/Reiter vorgeschlagen.

Zum anderen können literarische Phänomene dadurch komplex sein, dass ihre Feststellung in einem Text interpretationsabhängig ist. Interpretationsabhängigkeit ist dabei als gradierbare Eigenschaft zu verstehen – der Grad bemisst sich danach, in welchem Maße nicht-wahrheitserhaltende Schlüsse und strittige (Kontext-)Annahmen notwendig sind, um das Vorliegen des Phänomens festzustellen. Auch bei der Feststellung des Vorliegens interpretationsabhängiger Phänomene spielen aber in der Regel der literarische Text selbst bzw. konkrete identifizierbare Texteeigenschaften eine zentrale Rolle – die alleinige Bezugnahme auf sie ist aber eben nicht ausreichend, um für ihr Vorliegen zu argumentieren. Bei starker Interpretationsabhängigkeit ist es in der Regel extrem kompliziert (bzw. möglicherweise unmöglich), ein literaturwissenschaftliches Konzept vollständig (computationell) zu operationalisieren. Dies ergibt sich zum einen durch die Schwierigkeit, die im Rahmen von Interpretation stattfindenden zahlreichen und schwer zu fassenden Prozesse überhaupt zu rekonstruieren, zum anderen durch die Herausforderung, extratextuelles Wissen in computationellen Analysen abzubilden.

Einige zentrale Aspekte, die für das Vorliegen unzuverlässigen Erzählens notwendig sind, sind interpretationsabhängig. Beispielsweise ist es für faktenbezogene Unzuverlässigkeit notwendig, dass eine Erzählfigur falsche Aussagen über die erzählte Welt tätigt (oder relevante Informationen auslässt, vgl. Kindt 2008, 53). Eine derartige Diagnose erfordert eine Entscheidung darüber, was in der erzählten Welt wahr und relevant ist. Für Entscheidungen dieser Art sind zwar Textargumente (vgl. Descher/Petraschka 2019, 88–93) sehr wichtig. Aber zum einen müssen diese Textargumente unter Umständen in komplizierter Weise gegeneinander abgewägt werden und zum anderen sind – gerade bei potenziell unzuverlässig erzählten Texten – Kontextannahmen in der Regel unerlässlich für die Rekonstruktion der fiktiven Welt.

Während eine vollständige (computationelle) Operationalisierung unzuverlässigen Erzählens also wenig aussichtsreich erscheint, bietet die Relevanz von Textargumenten dennoch einen aussichtsreichen Aus-

gangspunkt für die Entwicklung computationeller Modelle, die für die Analyse von Unzuverlässigkeit in Texten unterstützend herangezogen werden können. So werden in der (nicht-computationellen) Unzuverlässigkeitsforschung auch tatsächlich Indikatorenlisten zusammengestellt, die unter anderem auch wenig komplexe sprachliche Phänomene enthalten, deren computationelle Modellierung aussichtsreich bzw. bereits umgesetzt ist. Die Indikatoren reichen von konkreten sprachlichen Einzelphänomenen („Ausrufe, Ellipsen, Wiederholungen“) und nicht spezifizierten linguistischen Sammelphänomenen („linguistische Signale für Expressivität und Subjektivität“) über Eigenschaften bzw. Zustände von Erzähler:innen („Hinweise auf kognitive Einschränkungen“) sowie Sprachhandlungen und Absichten (versuchte „Rezeptionslenkung durch den Erzähler“, Nünning 1998, 27–28) bis hin zu inhaltlich-strukturellen (verschiedene Arten von Widersprüchen) und inhaltlich-kontextuellen Phänomenen (stark unwahrscheinliche oder unmögliche Aussagen). Eine Analyse der Indikationsbeziehungen wird in der Unzuverlässigkeitsforschung allerdings nicht vorgenommen.

Zum Status von Indikatoren für komplexe literaturwissenschaftliche Phänomene

Um Klarheit über die Relevanz konkreter computationell modellierter sprachlicher Indikatoren im Zusammenhang mit komplexen literarischen Phänomenen zu erlangen, sollte zum einen reflektiert und kommuniziert werden, welche Funktion dem Modell zukommen soll, also ob es beispielsweise als Heuristik zum Auffinden für eine Forschungsfrage potenziell relevanter Texte (bzw. zur Exploration von Korpora), in argumentativen Zusammenhängen genutzt (vgl. Gerstorfer 2020) oder in anderen Funktionen eingesetzt werden soll. Hiervon ist abhängig, wie stark die Indikationsbeziehung überhaupt sein muss, um valide (oder: plausible) Ergebnisse erzielen zu können (vgl. Gius/Jacke 2022).²

Zwei weitere Aspekte der Indikationsbeziehung, die für eine literaturwissenschaftliche Verwertbarkeit computationeller Modelle im Zusammengang mit komplexen Phänomenen analysiert und kommuniziert werden sollten, werden im Folgenden etwas genauer vorgestellt.

Erklärung der Indikationsbeziehung

Grundsätzlich gilt, dass computationelle Modelle, die für die Textanalyse eingesetzt werden, für viele Literaturwissenschaftler:innen insbesondere dann interessant sind, wenn (zumindest ansatzweise) nachvollziehbar wird, *warum* sie funktionieren. Es ist deswegen lohnenswert zu fragen, warum bestimmte einfache sprachliche Textmerkmale auf das Vorliegen eines bestimmten komplexen Phänomens hinweisen können – und ob sich zwischen diesen beiden Polen möglicherweise Phänomene mittlerer Komplexität identifizieren lassen, die dieses Indikationsverhältnis (logisch-inhaltlich) besser nachvollziehbar machen.

Warum, beispielsweise, soll ein sprachliches Phänomen wie Ausrufe unzuverlässiges Erzählen indizieren? Ausrufe sind ein Merkmal expressiver Sprache. Expressive Sprache weist auf eine emotional aufgewühlte Erzählinstanz hin. Eine emotional aufgewühlte Erzählinstanz neigt dazu, etwas durcheinanderzubringen. Eine Erzählinstanz, die etwas durcheinanderbringt, ist disponiert, inkorrekte Aussagen zu treffen. Eine Erzählinstanz, die inkorrekte Aussagen tätigt, lässt sich als unzuverlässige Erzählinstanz einordnen.

Während diese Reihe das Fortschreiten von nicht zu stark interpretationsabhängigen Texteigenschaften illustriert, sollten insbesondere bei dem hier gewählten Beispiel auch kausale Zusammenhänge bzw. Richtungen beachtet werden. Eine entsprechende Analyse zeigt, dass das Vorliegen sprachlicher Merkmale wie Ausrufe im Text und das Vorliegen unzuverlässigen Erzählens nicht im eigentlichen Sinn kausal verbunden sind, sondern dass beide Phänomene (intrafiktionaler Logik folgend) die gleiche Ursache haben können – nämlich eine emotional aufgewühlte Erzählinstanz (siehe auch Reichenbachs *common cause principle*, vgl. Hitchcock/Rédei 2020).

Art und Stärke der Indikationsbeziehung

Konkrete sprachliche Texteigenschaften, die als Indikatoren für komplexere literarische Phänomene betrachtet werden, können diese Phänomene in unterschiedlicher Art und mit unterschiedlicher Stärke indizieren:

(1) Das Vorkommen bestimmter sprachlicher Indikatoren (ggf. mit einer bestimmten Frequenz, in bestimmten Kombinationen oder an bestimmten Stellen in einem Text) kann notwendig oder (meist in Kombination mit anderen Indikatoren) hinreichend für das Vorliegen eines bestimmten komplexen Phänomens sein. Ein Indikator ist dann notwendig für ein Phänomen, wenn er vorliegen muss, sofern das Phänomen vorliegt; hinreichend ist er dann, wenn sein Vorliegen das Vorliegen des Phänomens garantiert (vgl. Brennan 2022). Während ein einzelnes sprachliches Phänomen in der Regel weder notwendig noch hinreichend für das Vorliegen eines komplexen literarischen Phänomens ist, lassen sich unter Einbeziehung der oben genannten Zwischenphänomene als Verbindungsglieder zwischen sprachlichem Indikator und komplexem Phänomen aussagekräftigere Ergebnisse erzielen.³ Für jede dieser Indikationsbeziehungen lässt sich analysieren, ob der Indikator (möglicherweise auch in Konjunktionen oder Disjunktionen mit weiteren Indikatoren) notwendig oder hinreichend ist. Beispielsweise lässt sich feststellen, dass die Disjunktion aus Ausrufen und bestimmten weiteren sprachlichen Einzelphänomenen (wie beispielsweise expressiven Adjektiven etc., vgl. Gutzmann 2019) notwendig für expressive Sprache ist. Derartige Analysen lassen sich etwa auf der Basis von begriffsanalytischen Überlegungen und literaturwissenschaftlichen Argumentationsanalysen durchführen.⁴

(2) Basierend auf dem Vorkommen bestimmter sprachlicher Indikatoren (ggf. mit einer bestimmten Frequenz, in bestimmten Kombinationen oder an bestimmten Stellen in einem Text) kann dem Vorliegen eines bestimmten komplexen Phänomens eine hohe Wahr-

scheinlichkeit zugeschrieben werden. Solche bedingte Wahrscheinlichkeit lässt sich zum einen ebenfalls auf einer Mikroebene analysieren: So ist beispielsweise anzunehmen, dass expressive Sprache zwar weder notwendig noch hinreichend für eine emotional aufgewühlte Erzählinstanz ist, diese aber mit hoher Wahrscheinlichkeit indiziert.⁵ Zum anderen kann aber auch die Wahrscheinlichkeit auf der Makroebene interessant sein, mit der ein einfaches sprachliches Phänomen ein komplexes literarisches indiziert. Derartige Untersuchungen lassen sich nur mithilfe von Studien durchführen, bei denen Literaturwissenschaftler:innen das Vorliegen relevanter mittelkomplexer und komplexer Phänomene in einem Testkorpus beurteilen und manuell auszeichnen, in dem zugleich sprachliche Indikatoren für das Phänomen computationell identifiziert und ausgewertet werden. Derartige Studien werden im Rahmen des Projekts CAUTION durchgeführt.⁶

Auch bei der Analyse von Wahrscheinlichkeitszusammenhängen ist es wichtig, Kausalitätsrichtungen zu beachten, um die Relevanz eines computationellen Modells richtig einzuschätzen: Selbst wenn beispielsweise Emotionalität der Erzählinstanz sich mit sehr hoher Wahrscheinlichkeit in Form (automatisch messbarer) emotionaler Sprache niederschlägt, und wenn dieselbe Emotionalität ebenfalls mit hoher Wahrscheinlichkeit unzuverlässiges Erzählen hervorbringt, muss untersucht werden, ob die sprachlichen Indikatoren und unzuverlässiges Erzählen nicht mit ebenso hoher Wahrscheinlichkeit jeweils andere (verschiedene) Ursachen haben können.

Insgesamt kann die Identifikation von mittelkomplexen Phänomenen als Verbindungsgliedern zwischen sprachlichen Indikatoren und komplexen literarischen Phänomenen also nicht nur die Relevanz computationell modellier- und auswertbarer Texteigenschaften für literaturwissenschaftlich interessante Phänomene logisch-inhaltlich besser begreifbar machen. Sie ermöglicht auch eine aussagekräftigere (theoretische und methodologische) Analyse des komplexen literaturwissenschaftlichen Konzepts sowie eine Analyse der Indikationsverhältnisse und schafft die Grundlage für Teilerfolge (bspw. falls zwar letztlich kein signifikantes Indikationsverhältnis zwischen sprachlichem Indikator und komplexem Phänomen festgestellt werden kann, wohl aber zwischen sprachlichem Indikator und literaturwissenschaftlich ebenfalls relevanten Zwischenphänomenen). Der hier vorgeschlagene Weg, komplexe literarische Phänomene im Rahmen computationeller Zugänge bewusst nur partiell zu operationalisieren und zu modellieren, stellt auf diese Weise eine Möglichkeit dar, wie die Untersuchung literarischer Phänomene von den Vorteilen computationeller Analysen (u.a. Exaktheit und Reichweite) profitieren kann, ohne dass der Vorwurf des Reduktionismus angebracht ist. Denn nicht nur wird der genaue (eingeschränkte) Stellenwert der Analysen explizit gemacht – dies geschieht zudem auf eine Weise, die den Zusammenhang zwischen sprachlichen Merkmalen und komplexen interpretationsabhängigen Phänomenen auch inhaltlich expliziert und dadurch geisteswissenschaftlich anknüpfbar macht.

Fußnoten

1. „CAUTION“ steht für „Computer-Aided Analysis of Unreliability and Truth in Fiction – Operationalizing and Interconnecting Narratology“. Das Projekt ist assoziiert mit dem DFG-Schwerpunktprogramm Computational Literary Studies (vgl.).
2. Im Zusammenhang mit unzuverlässigem Erzählen scheinen die sprachlichen Indikatoren zumindest eine heuristische Funktion einzunehmen: Durch sie können Leser:innen auf die Idee gebracht werden, es mit unzuverlässigem Erzählen zu tun zu haben. Welche dieser Indikatoren auch für die Stützung von Unzuverlässigkeitsdiagnosen genutzt werden können, soll im Rahmen des Projekts CAUTION erforscht werden.
3. Ein ähnliches Vorgehen lässt sich im Rahmen von Mackies Rekonstruktion von Kausalität finden (vgl. Mackie 1974, 59–87).
4. Konkret kann dies beispielsweise bedeuten, Definitionen relevanter Begriffe wie „Emotion“ heranzuziehen oder literaturwissenschaftliche Argumentationen zu untersuchen, die für die Emotionalität einer Erzählfigur argumentieren, um herauszufinden, welche Texteigenschaften in diesem Zusammenhang relevant sind.
5. Für die genauere Analyse der involvierten Wahrscheinlichkeitswerte ließen sich probabilistische Logiken einsetzen, die beispielsweise mit dem Konzept der Wahrscheinlichkeitserhaltung (statt dem der Wahrheitserhaltung) arbeiten, vgl. Demey/Sack 2019.
6. Zum Zeitpunkt der Verfassung dieses Beitrags ist die Annotation noch nicht so weit fortgeschritten, dass erste Ergebnisse im Hinblick auf Korrelation und Wahrscheinlichkeitswerte präsentiert werden können.

Bibliographie

- Brennan, Andrew. 2022. „Necessary and Sufficient Conditions.“ *The Stanford Encyclopedia of Philosophy*, hg. von Edward N. Zalta. <https://plato.stanford.edu/archives/fall2022/entries/necessary-sufficient/> (zugegriffen: 1. August 2022).
- Carnap, Rudolf. 1959. *Induktive Logik und Wahrscheinlichkeit*. Wien: Springer.
- Demey, Lorenz und Joshua Sack. 2019. „Logic and Probability.“ *The Stanford Encyclopedia of Philosophy*, hg. von Edward N. Zalta. <https://plato.stanford.edu/entries/logic-probability/> (zugegriffen: 15. Dezember 2022).
- Descher, Stefan und Thomas Petraschka. 2019. *Argumentieren in der Literaturwissenschaft. Eine Einführung*. Stuttgart: Reclam.
- Döring, Nicola und Jürgen Bortz. 2016. „Operationalisierung.“ *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, hg. von Nicola Döring und Jürgen Bortz. Berlin, Heidelberg: Springer, 221–289.
- Drost, Ellen. 2011. „Validity and Reliability in Social Science Research.“ *Education Research and Perspectives* 38 (1), 105–124.
- Gerstorfer, Dominik. 2020. „Entdecken und Rechtfertigen in den Digital Humanities.“ *Reflektierte algorithmische Textanalyse. Interdisziplinäre Arbeiten in der CRETA-Werkstatt*, hg. von Nils Reiter, Axel Pichler und Jonas Kuhn.

De Gruyter, 107–124. <https://www.degruyter.com/document/doi/10.1515/9783110693973-005/html> (zugegriffen: 1. August 2022).

Gius, Evelyn. 2019. „Computationelle Textanalysen als fünfdimensionales Problem. Ein Modell zur Beschreibung von Komplexität.“ LitLab Pamphlet #8, hg. von Thomas Weitin. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2019/12/pamphlet_gius_2.0.pdf (zugegriffen: 1. August 2022).

Gius, Evelyn und Janina Jacke. 2022, im Erscheinen. „Are Computational Literary Studies Structuralist?“ *Journal of Cultural Analytics*.

Gutzmann, Daniel. 2019. *The Grammar of Expressivity*. Oxford, New York: Oxford University Press.

Hitchcock, Christopher und Miklós Rédei. 2020. „Reichenbach's Common Cause Principle.“ *The Stanford Encyclopedia of Philosophy*, hg. von Edward N. Zalta. <https://plato.stanford.edu/entries/physics-Rpcc/> (zugegriffen: 15. Dezember 2022).

Jacke, Janina. 2020. *Systematik unzuverlässigen Erzählens. Analytische Aufarbeitung und Explikation einer problematischen Kategorie*. Berlin, Boston: de Gruyter.

Kindt, Tom. 2008. *Unzuverlässiges Erzählen und literarische Moderne. Eine Untersuchung der Romane von Ernst Weiß*. Tübingen: Niemeyer.

Mackie, John Leslie. 1974. *The Cement of the Universe. A Study of Causation*. Oxford: Clarendon Press.

Moretti, Franco. 2013. „Operationalizing‘ or, the function of measurement in modern literary theory.“ LitLab Pamphlet #6, hg. von Thomas Weitin. <https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf> (zugegriffen: 1. August 2022).

Nünning, Ansgar. 1998. „‘Unreliable Narration‘ zur Einführung. Grundzüge eines kognitiv-narratologischen Theorie und Analyse unglaublichen Erzählens.“ *Unreliable Narration. Studien zur Theorie und Praxisunglaublichen Erzählens*, hg. von Ansgar Nünning, Bruno Zerweck und Carola Surkamp. Trier: Wissenschaftlicher Verlag Trier, 3–39.

Krautter, Benjamin, Nils Reiter und Axel Pichler. Im Erscheinen. „Operationalisierung.“ *Zeitschrift für digitale Geisteswissenschaften* (Sonderband: Glossar der Begriffe).

Pichler, Axel und Nils Reiter. 2021. „Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists *Das Erdbeben in Chili*.“ *Journal of Literary Theory* 15 (1–2), 1–29.

Von A bis Z: Überlegungen zur Erstellung eines Wissensgraphen aus historischen Enzyklopädien

Hagen, Thora

thora.hagen@uni-wuerzburg.de

Universität Würzburg, Deutschland

Strukturierte Daten, insbesondere Wissensgraphen, gewinnen in vielen Forschungsfeldern zunehmend an Bedeutung. Sie bieten eine kondensierte Sicht auf menschliches Wissen, sei es allgemein oder domänen-spezifisch. Unter anderem können sie damit in Forschungsprojekten helfen, Textkorpora mit zusätzlichen Informationen anzureichern oder das Beantworten spezifischer Fragestellungen durch Inferenzbildungen erst möglich zu machen. Für den Erstellungsprozess eines Wissensgraphen gibt es allerdings keinen universal gültigen Lösungsweg (Kejriwal 2019, 4).

In diesem Beitrag geht es speziell um die Herausforderungen, einen Wissensgraphen aus historischen Enzyklopädien zu erstellen. Als Beispiel dafür dient das DFG-geförderte Projekt *EncycNet*.¹ Es soll hierbei nicht die Präsentation von Methoden und deren Ergebnisse im Vordergrund stehen, sondern es sollen eher die Besonderheiten des Vorhabens und die damit einhergehenden konzeptuellen Überlegungen, die letztendlich auch bei der Auswahl der Methoden eine Rolle spielen, genauer dargestellt werden. Im Folgenden soll deshalb zunächst ein kurzer Überblick über Wissensgraphen und deren Erstellung gegeben werden. Den Hauptteil leitet dann eine kurze Einführung zu *EncycNet* ein und schließlich werden die drei Herausforderungen des Projekts diskutiert.

Forschungskontext

Um zu verstehen, was die Umwandlung von verschiedenen Quellen in einen Wissensgraph bedeutet, sollte zunächst der Begriff „Wissensgraph“ geklärt werden. Es gibt zwei verschiedene Definitionen des Begriffs; eine traditionelle und eine moderne Verwendung. Traditionell wird davon ausgegangen, dass ein Wissensgraph nur Weltwissen enthält – ganz konkret nur Named Entities. Dabei wird auch erwartet, dass es eine schwergewichtige Ontologie zu dem Wissensgraph gibt. Geprägt ist diese traditionelle Sichtweise durch das erste Aufkommen des Begriffs durch den Google Knowledge Graph, also Wikidata, welcher genau diese Eigenschaften besitzt. Gemäß dieser Definition sind also WordNet oder GermaNet auch keine Wissensgraphen, sondern semantische Netzwerke. Die moderne Definition ist dem

gegenüber etwas weniger streng; hier können Wissensgraphen jegliche Art von Wissen abbilden und die Ontologie darf ebenso auch leichtgewichtig sein.

Die Erstellung von Wissensgraphen lässt sich in drei Schritten zusammenfassen: 1) der Aufbau einer Ontologie, 2) die Erkennung von Entitäten und deren Alignierung und 3) die Relationsextraktion. Zusammengefasst aus der neuesten Forschung in dem Feld (aus Chaves-Fraga et al. 2021 u. 2022) lässt sich sagen, dass hier hauptsächlich das Mapping von Datenstrukturen im Vordergrund steht bzw. das Vereinheitlichen verschieden strukturierter Daten, um einen Graphen zu erstellen. Das bedeutet: Für alle drei Schritte kann auf bereits strukturierte Information zurückgegriffen werden und die Re-Organisation ist Kern der Aufgabe (Schröder et al. 2021, Wu et al. 2020). Daneben stehen häufig auch domänenspezifische Anforderungen an den Graphen im Vordergrund, so wie zum Beispiel bei der Erstellung des Open Drug Knowledge Graphs (Mann et al. 2021).

Ein Teilbereich der Forschung wiederum beschäftigt sich auch mit dem Erstellen von Graphen aus Fließtext, wobei hier die traditionelle Sichtweise auf Wissensgraphen, also die Repräsentation von realen Entitäten, dominiert. Typische Methoden aus dem Bereich sind daher z.B. Named Entity Recognition und Entity Linking (Kejriwal 2019, 12, 33). Dies gilt auch für das Feld der Digital Humanities, eben besonders für das Beschreiben kultureller Objekte, so wie zum Beispiel dem Modellieren von Erzählorten in Romanen (Hinzmann et al. 2022) oder dem Modellieren von Künstlern und Werken aus kunsthistorischen Texten (Jain et al. 2022). Daneben können auch syntaktische Marker statt Named Entities die Graphmodellierung stützen (Perak 2020). Häufig wird auch auf strukturierte Daten, z.B. Wissensgraphen aus derselben Domäne, zurückgegriffen, um eine Basis für die Alignierung der Konzepte, Relationsauswahl und Ontologie zu schaffen (Jain et al. 2022, Clancy et al. 2019).

EncycNet: Herausforderungen und Chancen

Das Ziel von EncycNet ist es, einen Wissensgraphen aus sechs historischen Enzyklopädien (genauer: Konversationslexika, siehe Tabelle 1) zu erstellen.² Der Graph soll ein lexikalisch-semantisches Netzwerk sein, welcher einerseits die Einträge über alle Enzyklopädien hinweg aligniert und disambiguiert. Zum anderen müssen die relevanten Relationen aus den Einträgen enthalten sein. So sollen beispielsweise Themenbereiche zu Stichwörtern verzeichnet sein („Operation“ / „Mathematik“ und „Operation“ / „Medizin“), aber auch typische Informationen über Named Entities wie Geburtsdatum und Geburtsort von bekannten Personen. Der Graph soll zunächst eine Vogelperspektive auf die Daten bieten; insbesondere sollen so Bedeutungsverschiebungen eines Konzepts im Verlauf der Zeit (z.B. durch Veränderung der nächsten Nachbarn) sichtbar gemacht werden. Ebenfalls kann mithilfe von Inferenzen der Graph genauer analysiert werden: Etwa die Verwandtschaftsgrade von Konzepten über Pfade, Zentralität von Konzepten für eine bestimmte Zeit oder die Bildung von Communities. Darüber hinaus können die Graphdaten dazu dienen,

Textkorpora anzureichern oder historische Evaluationsdaten bereitzustellen.

Tabelle 1: Auflistung der Konversationslexika für die Grundlage des Wissensgraphen EncycNet

Lexikon	Anzahl Einträge	Anzahl Tokens
Brockhaus Konversations-Lexikon oder kurzgefaßtes Handwörterbuch (1809)	6.960	1.186.000
Brockhaus Bilder-Konversations-Lexikon (1837)	7.049	2.604.000
Brockhaus Kleines Konversations-Lexikon (1911)	82.780	2.434.000
Damen Conversations-Lexikon (1834)	7.099	1.461.000
Herders Conversations-Lexikon (1854)	39.755	2.256.000
Meyers Großes Konversations-Lexikon (1905)	156.264	17.437.000

Für die Erstellung des Graphen bedeutet der dargestellte Forschungskontext, dass ein solches Vorhaben verschiedene Forschungsnischen in sich vereint. Zum einen bedient sich EncycNet der modernen Definition für Wissensgraphen. Da in den historischen Enzyklopädien nicht nur Personen und Orte zu finden sind, sondern auch Objekte und Phänomene aller Art, muss der resultierende Wissensgraph diese Dinge auch abbilden können. Zum anderen liegen die Enzyklopädien nur semi-strukturiert (TEI-XML) vor.

³ – zum Großteil handelt es sich deshalb um eine Grapherstellung aus Fließtext. Und letztlich, durch die Diversität, die mit einem historischen Korpus einhergeht, muss die Erstellung zusätzlich auf heterogene Fließtextdaten abgestimmt sein. Auf diese drei Herausforderungen, die das Enzyklopädienkorpus mit sich bringt, soll im Folgenden genauer eingegangen werden.

Domänenübergreifendes Wissen

In den Enzyklopädien wird jegliche Art von Wissen über unterschiedlichste Wissensdomänen äußerst detailliert abgebildet. Insbesondere dann, wenn Einträge deutlich länger als nur eine Definition sind, findet sich dort einiges an Wissen, welches sich idealerweise auch in dem resultierenden Wissensgraphen widerspiegeln sollte. Allerdings sind solche längeren Einträge auch weniger standardisiert als die kürzeren, einfachen Begriffsdefinitionen in den Enzyklopädien. Trotzdem gibt es auch in längeren Einträgen bereits einige vorstrukturierte Elemente, gekennzeichnet durch die XML Annotation, die sich auf den ersten Blick zum extrahieren anbieten. Dazu gehören beispielsweise Hierarchien oder Aufzählungen, Vers, oder auch semi-strukturierte Formen im Fließtext wie z.B. Gleichungen, Angaben von Einheiten oder Ähnliches (einige Beispiele in Abbildung 1). Allerdings sind diese Elemente dann fast immer domänenspezifisch und sind damit nur in den wenigsten Einträgen zu finden.

<p>"Cam" in Herder 1854 Cam, ostind. Silbermünze = 4 Sgr. 9¹/₄ Pf. = 15¹/₂ kr. C.-M.</p>							
<p>"Pint" in Herder 1854 Pint, engl. und nordamerik. Hohlmaß; für Getreide = 28³/₈, für Flüssigkeit = 28¹³/₂₅ Par. Kubikzoll. P.e. Getreidemaß in der Lombardei = 50,4 Par. Kubikzoll; Flüssigkeitsmaß in Bergamo = 62,1, in Brescia = 69¹/₂, auf</p>							
<p>"Apotheke" in Herder 1854 worden sein. – Apothekergewicht in Deutschland: Pfund, As (lb) = 12 Unzen, (3), die U. = 8 Drachmen (3), die D. = 3 Skrupeln (3), der Skr. = 20 Gran (gr.), von denen also 5760 auf 1 Pf. gehen.</p>							
<p>"Asien" in Meyer 1905 Sprachen gibt für die Bevölkerung Asiens ungefähr folgende Gruppen: A. Nordasiaten. I. Inkgirisch. II. Korjakisch, Tschuktschisch. III. Sprachen von Kamtschatka und Kurilen (Aino). IV. Jensei-Ostjakisch und Kottisch. B. Mittel- oder Hochasiaten. I. Uralaltaische Sprachen. a) Samojedische Gruppe: Jurakisch,</p>							
<p>"Geschwindigkeit" in Brockhaus 1837 In einer Sekunde legen zurück:</p> <table> <tr> <td>Flüsse von mittlerer Geschwindigkeit</td><td>3–4 Fuß.</td></tr> <tr> <td>Winde von mäßiger Stärke</td><td>10 Fuß.</td></tr> <tr> <td>Ströme von größter Geschwindigkeit</td><td>12 Fuß.</td></tr> </table>		Flüsse von mittlerer Geschwindigkeit	3–4 Fuß.	Winde von mäßiger Stärke	10 Fuß.	Ströme von größter Geschwindigkeit	12 Fuß.
Flüsse von mittlerer Geschwindigkeit	3–4 Fuß.						
Winde von mäßiger Stärke	10 Fuß.						
Ströme von größter Geschwindigkeit	12 Fuß.						
<p>"Assonanz" in Brockhaus 1837 nichts Weicheres und Lieblicheres denken, als Verse wie diese: »Wonne weht von Thal und Hügel, Weht von Flur und Wiesenplan, Weht vom glatten Wasserspiegel, Wonne weht mit weichem Flügel Des Piloten Wange an.«</p>							

Abbildung 1: Beispiele für strukturierte bis semi-strukturierte Inhalte in den Einträgen der Enzyklopädien.

Um ein möglichst generisches Bild von dem Inhalt der Enzyklopädien zu erhalten, können die Einträge zunächst in generische Klassen unterteilt werden; für EncycNet ergaben sich die Klassen Personen, Orte, Objekte und Abstrakta. Diese Klassen wurden nach dem Vorbild von Wikidata (Personen und Orte als 2 Hauptkategorien) und WordNet („physical entity“ und „abstract entity“) als direkte Hyponyme des Wurzelements „entity“). Mehrere Artikel aus diesen Klassen können dann gesichtet werden und in Abstimmung mit Wikidata oder WordNet die wichtigsten Informationen oder thematische Segmente innerhalb der Einträge identifiziert werden.

So ergeben sich drei Gütekriterien für die regelbasierte Extraktion von strukturiertem Wissen aus historischen Enzyklopädien: 1) Wie generisch ist die Information; also auf wie viele Klassen und auf wie viele Einträge trifft sie insgesamt zu, 2) Wie stark ist die Information vorstrukturiert, also wie präzise kann eine Regel für die Information gefunden werden und 3) Wie relevant ist die Information im Hinblick auf die Ziele, die der Wissensgraph verfolgt? Recht generische Informationen sind beispielsweise Synonyme und Übersetzungen; in den meisten Einträgen werden diese direkt nach der Nennung des Konzepts aufgelistet und sind damit auch vergleichsweise

einfach zu extrahieren. Die Beispiele aus Abbildung 1 sind das Gegenteil: Zwar sind sie alle eher einfach zu extrahieren durch die vorstrukturierte Form, allerdings sind sie auch in nur wenigen Einträgen vorhanden und insbesondere Zahlen und Einheiten sind für das Ziel das EncycNet verfolgt, nämlich ein semantisches Netzwerk zu bilden, eher uninteressant. Es empfiehlt sich daher, zunächst alle möglichen Informationen nach diesen Kriterien zu sortieren und dann erst mit der Extraktion zu beginnen, um eben nicht nur Detailwissen zu extrahieren bzw. bestimmte Domänen zu bevorzugen.

Die domänenübergreifende Perspektive wirkt sich auch auf die Auswahl der Ontologie aus. Zusammen mit dem Ziel, nicht nur Entitäten sondern auch lexikalischen Wissen semantisch zu modellieren, schließt dies einige Standards aus. OntoLex (Cimiano et al. 2016) beispielsweise ist gerade dafür gedacht, lexikalisches Wissen aus Nachschlagewerken in einen Graphen umzuwandeln. Allerdings liegt hier der Fokus auf morphologischen statt semantischen Eigenschaften (Wortart, Genus, etc.), welche in Lexika nicht unbedingt aufgeführt werden. Die Struktur des Lexikons wird außerdem explizit beibehalten. So werden beispielsweise Referenzen auf andere Artikel als Kante „reference“ eingepflegt und nicht weiter typisiert. Daneben gibt es CIDOC-CRM (Doerr 2005), eine Standard-Ontologie zum Beschreiben von kulturellen Objekten, und Faktoide (Bradley und Short 2005) zur Abbildung von spezifischen Stellen in einer Quelle über Personen. Bei beiden Modellierungsarten sind Objekte und Eigenschaften eher auf Named Entities und weniger auf Lexeme ausgelegt. Alle drei Standards sind gut geeignet für Teilbereiche der Enzyklopädien, so wie zum Beispiel Faktoide für biographische Einträge. Die möglichst vollständige Abbildung aller Inhalte die EncycNet anstrebt, auch hauptsächlich von lexikalischen Eigenschaften (z.B. Synonyme oder Hyperonyme), ist aber nicht möglich.

Synonyme und Hyperonyme gehören bei der Bildung eines lexikalisch-semantischen Netzwerks zu den Grundbausteinen; sind also besonders wichtig für Punkt 3. Synonyme werden benötigt, um Synsets nach dem Vorbild von WordNet zu bilden (bzw. das Verknüpfen gleicher Konzepte) und Hyperonyme für den Aufbau einer Taxonomie. Allerdings ist die Extraktion einer vollständigen Taxonomie, welche alle Domänen umfasst, nur aus dem Enzyklopädienkörper weitestgehend unrealistisch, weswegen auf zusätzliches Material zurückgegriffen werden muss. Als bislang größte Online-Enzyklopädie kann Wikidata / Wikipedia, welche inzwischen nicht nur Entitäten-bezogenes Wissen sondern über die Integration von WordNet auch über lexikalisches Wissen verfügt, als Schnittstelle genutzt werden. Sie liefert einerseits die Taxonomie, aber auch andererseits über Wikipedia zusätzliches Textmaterial zu den Konzepten, welches ebenfalls für die Alignierung der Einträge über verschiedene Enzyklopädien hinweg von Nutzen sein kann (Hagen et al. 2022).

Heterogene Daten

Der Wissensgraph, der durch EncycNet entstehen soll, fasst das Wissen von 6 allgemeinen Enzyklopädien zusammen. Alle diese Enzyklopädien unterscheiden sich

hinsichtlich der Auswahl und Organisation der Konzepte, der inneren Struktur der Einträge, Umfang und Auslegung der Definitionen und dem Stil. Jede dieser Eigenheiten müssen für die Informationsextraktion berücksichtigt werden, was bedeutet, dass die Relationen für alle Enzyklopädien neu beurteilt werden müssen. Pragmatisch betrachtet heißt dies auch, dass die Informationsdichte im resultierenden Graph abnimmt, da weniger Relationen in Betracht genommen werden können. Insbesondere auf mögliche Inferenzbildungen durch den Graph wirkt sich das negativ aus.

Aus diesem Grund sollten generische Methoden für die Informationsextraktion hinzugezogen werden, sodass der Graph an Informationsdichte gewinnt. Hierbei kann auf typische Methoden in den Digital Humanities zurückgegriffen werden: Topic Modeling zum Identifizieren von übergreifenden Wissensbereichen, TF-IDF für distinktive Terme eines Eintrags, Komposita des Konzepts als verwandte Konzepte, Named Entity Recognition, oder spezifisch für Enzyklopädien die Extraktion von Referenzen auf andere Einträge. Nachteil dieser generischen Extraktion ist allerdings, dass das Mapping der extrahierten Terme auf eine Relation sich schwieriger gestaltet bzw. viele der Terme eine unspezifische Relation zum Konzept erhalten (z.B. in GermaNet „related_to“).

Auch für die Alignierung ergeben sich praktische Probleme bei der Grapherstellung, denn in den Enzyklopädien können die Konzepte unterschiedlich organisiert sein. Dies reicht von Betitelungskonventionen der Einträge (z.B. „Der Adler“ / „Adler“, oder „William Shakespeare“ / „Shakespeare“ / „Shakespeare, William“) bis hin zu der Möglichkeit, dass es Einträge gibt, die mehrere Konzepte zusammenfassen. In Meyer (1905) werden beispielsweise manche Entitäten thematisch gruppiert. So gibt es zum Beispiel genau zwei Einträge zu *Alexander*: einer gruppiert Fürsten und der andere griechische Schriftsteller mit dem Vornamen, die auch jeweils nochmal genauer beschrieben werden. In Herder (1854) dagegen gibt es drei Einträge: einen, der den Vornamen ohne Personenzuordnung nennt, einen zu Alexander I. (welcher auch Alexander II. umfasst) und einen zu Alexander III. Hier gilt es also zu klären, inwieweit die Konzepte getrennt werden können. In Meyer sind die Konzepte deutlich durch Paragraphen gekennzeichnet, in Herder werden sie sprachlich vermischt. Zusätzlich können allerdings, selbst wenn die Konzepte isoliert sind, diese auch unterschiedlich ausgelegt werden, auch historisch bedingt.

Historisches Wissen

Die Historizität der Daten wirkt sich damit ebenso auf die Alignierung der Konzepte aus. Auf der einen Seite werden teils archaische Begriffe oder Schreibweisen für Konzepte verwendet (z.B. „Irrenanstalt“, jetzt „psychiatrische Klinik“). Durch eine orthographische Normalisierung und durch die Verwendung von Wikipedia für die Alignierung, welche zum Teil auch archaische Begriffe umfasst und auflösen kann, kann diesem Problem entgegengekommen werden. Auf der anderen Seite können sich aber auch die Definitionen von Begriffen im Laufe der Zeit stark verändert haben, etwa weil sich ein Konzept

in ein anderes verwandelt hat oder weil Konzepte zu unterschiedlichen Zeiten anders ausgelegt werden.

So wird beispielsweise in Meyer (1905) die *Exploration* beschrieben als die physische Untersuchung eines Kranken durch einen Arzt, während sie in Wikipedia mit *Anamnese* gleichgestellt wird, also der Erfragung von Informationen im Rahmen einer Erkrankung. Ein weiteres Beispiel findet sich in Herder (1854): Der Begriff *Proportionalität* wird beschrieben durch „Harmonie der Größenverhältnisse“ und „Proportionslehre der menschlichen Gestalt“, während für *Proportion* die Verhältnisgleichung in der Mathematik genannt wird; also was die Menschen heute eigentlich unter Proportionalität verstehen würden.

Diese Beispiele zeigen eine notwendige Modellierungsentscheidung für einen historischen Wissensgraphen auf: Sollen Konzepte, welche sich grundlegend verändert haben trotzdem miteinander aligniert werden und die Alignierung macht somit den semantischen Wandel sichtbar? Oder sollten nur Konzepte, welche tatsächlich semantische Äquivalente sind aligniert werden, da die Definitionen so unterschiedlich ausfallen können? Je nach Ziel, den der resultierende Graph verfolgt, kann diese Entscheidung unterschiedlich ausfallen.

Letztlich wirkt sich das historische Wissen auch auf den Aufbau der Taxonomie aus. Um eine vollständige Taxonomie automatisch zu generieren, kann auf bestehendes strukturiertes Wissen (wie Wikidata) zurückgegriffen werden. Allerdings besteht hierbei die Gefahr, historisches Wissen zu ignorieren oder zu überschreiben. In den Enzyklopädien wird beispielsweise der Begriff *Hexe* als „Unholdin“, „Weib“ oder „weissagende Frau“ beschrieben, während in Wikidata „Magier“ verwendet wird. Ein anderes Beispiel ist *Hierarchie*, welche in Wikidata als „Struktur“ oder „System“ eingeordnet wird, jedoch in den Enzyklopädien mit „Priesterherrschaft“, „Macht“ oder „alle Rechte der Römischen Päpste über die gesamte Christenheit“ beschrieben wird.

Zusammenfassung

Das Bilden eines historischen Wissensgraphen aus Enzyklopädien kombiniert zwei Forschungsnischen aus dem Forschungsfeld. Einerseits geht es hier um die Abbildung von lexikalisch-semantischem Wissen und nicht um nur um Entitätenwissen wie z.B. Wikidata, und andererseits stellt heterogener Fließtext die Datengrundlage für den Graphen dar. Beides sind Voraussetzungen, die sich selten in der aktuellen Forschung zur Erstellung von Wissensgraphen wiederfinden. Der Aufbau von EncycNet wird geleitet von praktischen Anforderungen an den Graphen, welche sich aus den vordefinierten Zielen ergeben. Dabei geht es um die Einschätzung, welche Informationen in den Enzyklopädien sich für eine Relationsextraktion anbieten, aber gleichzeitig soll möglichst viel Wissen mit dem Graphen abgedeckt werden. Zusätzlich, insbesondere um Inferenzen zu ermöglichen, muss der resultierende Graph eine möglichst große Informationsdichte in Tiefe (Taxonomie) und Breite (generische Relationen) aufweisen. Es ergibt sich daraus eine Bottom-up Strategie (schematisch zusammengefasst in Abbildung 2).

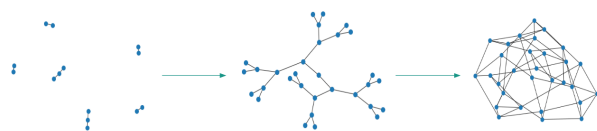


Abbildung 2: Schematische Darstellung der Erstellung des Graphen: gezielte Extraktion von Relationen aus den Enzyklopädien (links), Aufbau der Taxonomie (Mitte) und Verdichtung durch ungerichtete, generische Relationsextraktion (rechts).

Für EncycNet werden aktuell Alignierung sowie Relationsextraktion fertiggestellt. Da die größtmögliche Abdeckung für beide Aufgaben erzielt werden soll, werden die Methoden diesbezüglich kontinuierlich optimiert. Noch ausstehend ist die Evaluierung des extrahierten Wissens mit Golddaten. Final sollen dann über die Evaluierung die Relationen und die Alignierung mit Gewichten ausgestattet werden, welche die Konfidenz angeben. Im Frühjahr 2024 sollen dann die Daten in RDF* zur Verfügung gestellt werden.

Im Vordergrund dieses Beitrags sollten damit jene Entscheidungsfindungen stehen, die in Methoden-orientierten Beiträgen sonst meist nur am Rande erwähnt werden. Dabei wurde sich auf EncycNet bezogen, jedoch können die hier aufgeführten Ideen genauso für Projekte, die auf ähnliche Herausforderungen bei dem Aufbau eines Wissensgraphen stoßen, interessant sein.

Fußnoten

1. <https://encycnet.github.io/>
2. Eine ausführliche Übersicht über das Korpus ist auf <https://encycnet.github.io/corpus-overview.html> gegeben.
3. Die TEI-Daten können über Zenodo (<http://dx.doi.org/10.5281/zenodo.4039569>) heruntergeladen werden.

Bibliographie

Bradley, John und Harold Short. 2005. "Texts into databases: the Evolving Field of New-style Prosopography." *Literary and Linguistic Computing* 20 (1): 3-24.

Chaves-Fraga, David, Anastasia Dimou, Pieter Heyvaert, Freddy Priyatna und Juan Sequeda. Hrsg. 2021. "Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference (ESWC 2021)." In *CEUR Workshop Proceedings* 2873. <http://ceur-ws.org/Vol-2873/> (zugegriffen: 01. August 2022).

Chaves-Fraga, David, Anastasia Dimou, Pieter Heyvaert, Freddy Priyatna und Juan Sequeda. Hrsg. 2022. "Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2022) co-located with 19th Extended Semantic Web Conference (ESWC 2022)." In *CEUR Workshop Proceedings* 3141. <http://ceur-ws.org/Vol-3141/> (zugegriffen: 01. August 2022).

Cimiano, Philipp, John P. McCrae, und Paul Buitelaar. 2016. *Lexicon model for ontologies: community report, 10 May 2016*. <https://www.w3.org/2016/05/ontolex/> (zugegriffen: 07. Dezember 2022).

Clancy, Ryan, Ihab F. Ilyas und Jimmy Lin. 2019. "Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond." In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, Hong Kong, China.

Doerr, Martin. 2005. "The CIDOC CRM, an ontological approach to schema heterogeneity." *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Hagen, Thora, Fotis Jannidis und Andreas Witt. 2022. "Word sense alignment and disambiguation for historical encyclopedias." In *6th International Conference on Graphs and Networks in the Humanities*. urn:nbn:de:bsz:mh39-109834 (zugegriffen: 09. Dezember 2022). Vorveröffentlichung.

Hinzmann, Maria, Julia Röttgermann, Anne Klee, Moritz Steffes und Christof Schöch. 2022. "The French Enlightenment Novel as a Graph? Potentials and Challenges in the Construction of a Knowledge Network." In *6th International Conference on Graphs and Networks in the Humanities*. 10.5281/zenodo.5840088 (zugegriffen: 09. Dezember 2022). Vorveröffentlichung.

Jain, Nitisha, Alejandro Sierra-Múnera, Maria Lomaeva, Julius Streit, Simon Thormeyer, Philipp Schmidt und Ralf Krestel. 2022. "Generating Domain-Specific Knowledge Graphs: Challenges with Open Information Extraction." In *Proceedings of International Workshop on Knowledge Graph Generation from Text (Text2KG), co-located with the Extended Semantic Web Conference (ESWC 2022)*.

Kejriwal, Mayank. 2019. *Domain-specific knowledge graph construction*. New York: Springer International Publishing.

Mann, Mark, Filip Ilievski, Mohammad Rostami, Aashta und Basel Shbita. 2021. "Open Drug Knowledge Graph." In *Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2022) co-located with 19th Extended Semantic Web Conference (ESWC 2022)*. <http://ceur-ws.org/Vol-2873/paper10.pdf> (zugegriffen: 01. August 2022).

Perak, Benedikt. 2020. "Modeling Semantic Relations from a Dependency-Based Graph: A Corpus-Based Network Analysis of Croatian Parliamentary Debates." In *Graph Technologies in the Humanities - Proceedings 2020*. <https://ceur-ws.org/Vol-3110/paper9.pdf> (zugegriffen: 09. Dezember 2022).

Schröder, Markus, Christian Jilek und Andreas Dengel. 2021. "Mapping Spreadsheets to RDF: Supporting Excel in RML." In *Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2021) co-located with 19th Extended Semantic Web Conference (ESWC 2021)*.

Wu, Tianxing, Haofen Wang, Cheng Li, Guilin Qi, Xing Niu, Meng Wang, Lin Li und Chaomin Shi. 2020. "Knowledge graph construction from multiple online encyclopedias." In *World Wide Web* 23 (5): 2671-2698.

„Werktitel als Wissensraum“ – über die Potentiale von Werknormdaten für die Digitalen Geisteswissenschaften

Dietrich, Elisabeth

elisabeth.dietrich@klassik-stiftung.de

Herzogin Anna Amalia Bibliothek Weimar, Deutschland

Kolbe, Ines

ines.kolbe@dla-marbach.de

Deutsches Literaturarchiv Marbach, Deutschland

1. Normdaten für Werke: Ausgangslage und Nutzen

Werknormdaten bieten die Möglichkeit, zentrale werkbezogene Angaben, wie Autorschaft, Erscheinungsjahr, Erscheinungsort und Genre, normiert – sprich standardisiert – in einem digitalen Datensatz abzubilden und mittels eines unikal identifizierenden persistenten Identifiers zur Verfügung zu stellen. Normdaten bilden wichtige Bausteine beim explorativen Suchen und Finden wissenschaftlich valider Informationen, schaffen durch ihre Vernetzung mit anderen Daten und Entitäten ein Semantic Web, aus dem sich neue fachübergreifende Forschungsfragen generieren lassen. Obwohl der wissenschaftliche Nutzen hoch ist, werden Werknormdaten im bibliothekarischen Arbeitsalltag in der Regel nur vereinzelt und rudimentär erstellt. Systematisch und in hoher Qualität werden Werknormdaten meist nur in drittmittelfinanzierten Projekten für eine konkrete Medienform oder Epoche erfasst, z.B. für Druckgrafiken, Bühnenstücke oder Werke der Musik (zu Letzterem: Bicher & Wiermann 2018). Das Projekt »Werktitel als Wissensraum«¹ hat das Ziel, den Grundstock für ein zentrales elektronisches, dynamisches Werklexikon zur deutschen Literatur innerhalb der GND zu legen. Der Werkbegriff folgt dem FRBR-Modell, das die Grundlage für RDA bildet, dem internationalen Regelwerk für Bibliotheken und Archive.² Das Konzept der werkorientierten Erschließung ist nicht neu und nimmt seinen Anfang in den von Antonio Panizzi entwickelten, 1841 erschienenen »Rules for the Compilation of Catalogue«, denen „das Prinzip zugrunde [liegt], alle Katalogeintragungen für unterschiedliche Ausgaben (Auflagen, Übersetzungen) eines Werkes zusammenzuführen“ (Barnert, Dietrich, Kolbe und Schmidgall 2021, 140). Durch die Einführung des Regelwerks RDA (Resource Description and Access) in Deutschland wird

das zugrundeliegende FRBR-Modell mit den Ebenen Werk, Expression, Manifestation und Exemplar bei der Katalogisierung berücksichtigt. In Datenbanken können die miteinander in Beziehung stehenden Entitäten wie Werke, Übersetzungen (Expressionen), Ausgaben (Manifestationen) und Exemplare virtuell zusammengeführt werden. Die Idee mit Hilfe von Werknormdaten Informationen zu bündeln, wird auch in anderen Ländern umgesetzt, etwa in Finnland, Frankreich und den Vereinigten Staaten. In den jeweiligen Verzeichnissen der Nationalbibliotheken – in denen die Werknormdaten unterschiedlich umfangreich aufbereitet sind – finden sich neben den Übersetzungstiteln und Editionen teils auch Adaptionen des Grundwerks.³ Neben Bibliotheken können andere Einrichtungen ihre Bestände durch die Nutzung der Werknormdaten anreichern. Somit schafft unser von der DFG gefördertes Projekt entscheidende Voraussetzungen für eine materialübergreifende Vernetzung musealer, bibliothekarischer und archivarischer Sammlungen und verbessert die Recherchemöglichkeiten für Nutzende (vgl. Althage 2019).

2. Zur Arbeitsweise des Projekts

Vor Projektbeginn wurden die deutschsprachigen Werke aus »Kindlers Literatur-Lexikon«, Wilperts »Lexikon der Weltliteratur« und Frenzels »Daten deutscher Dichtung« als wichtigste Registrationsmedien der Nachkriegsgermanistik sowie vier neuere, nach der Jahrtausendwende erschienene Werklexika, die auch nichtkanonisierte Einzelwerke enthalten, sowie die Werknormdatenpools in Marbach und Weimar ausgewertet. Die in den 10 Datenquellen am häufigsten vertretenen deutschsprachigen Werke bilden die Grundmenge von 4.625 Werken, wovon 2.050 Werktitel mit Erscheinungsjahr von 1700 bis 1914 in der Herzogin Anna Amalia Bibliothek Weimar und 2.575 Titel erschienen von 1915 bis 2015 in der Bibliothek des Deutschen Literaturarchivs Marbach bearbeitet werden. Der Befund, dass die Menge der in den Lexika aufgeführten Werke für das 20./21. Jahrhundert größer ist als für das 18./19. Jahrhundert, ist an sich schon aussagekräftig. Insgesamt sind ca. 800 Autor:innen vertreten, darunter nur 65 Frauen. Die Auswahl der bekanntesten Werke erfolgte mit der Absicht, einen möglichst großen Nutzen für die Katalogisierung zu bieten, da die verbreitetsten Werke in verschiedenen Übersetzungen und Auflagen vorliegen.⁴ Wir empfehlen, in Folgeprojekten Werklexika mit anderen Schwerpunkten auszuwählen, z.B. »Die deutschsprachigen Schriftstellerinnen des 18. und 19. Jahrhunderts« von Elisabeth Friedrichs oder »Frauen Literatur Geschichte« von Hiltrud Gnüg und Renate Möhrmann.⁵

Jedes der 4.625 Werke wird in der Gemeinsamen Normdatei (GND) und in Wikidata ergänzend beziehungsweise neu erstellt. Die GND ist eine kooperativ gepflegte digitale Normdatenbank für Personen, Körperschaften, Konferenzen, Geografika, Sachbegriffe und Werke. Wikidata hat sich in den letzten Jahren zu einem Knotenpunkt für die Vernetzung von Wissen entwickelt. In einem zweiten Arbeitspaket werden für jedes Werk der Grundmenge

in Beziehung stehende Werke ermittelt. Mit Hilfe zahlreicher fachspezifischer Nachschlagewerke werden Werkbearbeitungen wie Vertonungen, Bühnenbearbeitungen, Verfilmungen sowie Vorlagen, Fassungen und Nachfolger identifiziert und wiederum als Werknormdaten erfasst. Im Laufe des Projekts wurden bereits 12.000 Werknormsätze in der GND bearbeitet oder neu erstellt. In einem dritten Arbeitspaket werden die Normdaten mit- samt der GND-Identifikationsnummern in bestehende Wikidata-Einträge eingepflegt oder in neue Einträge übernommen. Die so entstehenden Netzwerke erlauben Einblicke in die intellektuelle Produktion und Kollaboration und offenbaren literaturhistorische Entwicklungen, Trends, Debatten und Themenschwerpunkte (Märchen, Wertheriaden, Exilliteratur). Mit diesem Portfolio wird das Projekt seinem Anspruch einer offen zugänglichen und vernetzten literarischen Gedächtniskultur mithilfe von standardisierten Daten gerecht und trägt damit dem Tagungsmotto *Open Humanities, Open Culture* vollends Rechnung.

3. Auswertungsmöglichkeiten und Anwendungsszenarien der Daten

Bei der Recherche und Anreicherung der Werknormdaten in der GND wurden bereits einige Besonderheiten registriert. Eine erste Übersicht über die Verteilung der Werkformen je Zeitraum zeigt zum einen die Vielfalt der Werkformen bei den Beziehungswerken (Tabelle 1). Zum anderen lässt sich im Vergleich feststellen, dass Film- und Hörspielbearbeitungen für die Werke der neueren Literatur häufiger vertreten sind. Auch geschlechtsspezifische Prozesse lassen sich anhand des Datensets nachvollziehen: die schrittweise Eroberung weiterer Handlungsräume und Schreibpraktiken im 20. Jahrhundert lassen allmählich Frauen als Autorinnen hervortreten (vgl. Seifert 2021, 93). Überwiegen bei den Frauen die Genre Prosa, Lyrik, Drama und Autobiografie, zeigt sich bei den fachspezifischen Werken zu Staatskunde, Philosophie, Geschichte, Religion oder Medizin, dass die Wissenschaftsgeschichte und -literatur bis weit ins 20. Jahrhundert überwiegend männlich dominiert war.⁶ Systematische Auswertungen der Art und Häufigkeit der Werkbeziehungen sowie von Verfasserschaft je Zeitraum können interessante Aufschlüsse zur Literaturproduktion und -rezeption geben.

Tab. 1: Übersicht der Werkformen der neuen Werktitel mit prozentualer Verteilung pro Zeitraum, eigene Darstellung

Neue Werktitel: Werkformen	Werke 1700-1914	Werke 1915-2015
Werke der Literatur (Libretti, Manuskripte, Fachliteratur, Bühnenstücke, Lyrik)	37,2 %	26,2 %
Werke der Musik (Opern, Lieder, Musicals)	19,8 %	18 %
Filmwerke (Spiel- und Fernsehfilme, Serien)	20 %	21,7 %
Hörspiele	15 %	32,4 %
Werke der Bildenden Kunst (Gemälde, Grafik, Objekte, Installationen)	6,8 %	1 %
Ballette/ Tanztheater	0,6 %	0,17 %
Computerspiele (virtuell)	0,1 %	0,1 %
Spiele (konventionell, haptisch)	0,2 %	0,1 %

Der Schwerpunkt des Projekts liegt auf der Erfassung der Werknormdaten. Die Expertise bei der Auswertung der Daten mit DH-Methoden liegen bei Ihnen, den Wissenschaftler:innen der Digital Humanities. Wir sind gespannt auf Ihre Anregungen, welche Anforderungen an die Daten in der Community bestehen, damit diese für zukünftige Projekte berücksichtigt werden können. Der Frage nach der Verknüpfung der Werknormdaten mit verfügbaren Volltexten muss gemeinsam weiter nachgegangen werden. Die Auswertung von Volltexten in Bezug auf Werke als Entitäten sowie um Bezüge zwischen Werken zu erkennen, scheint vielversprechend. Die Daten könnten auch als Trainingsdaten für maschinelle Lernverfahren genutzt werden. Die digital zugängliche Datengrundlage aus dem Projekt lässt sich auf vielfältige Weise nachnutzen, etwa um Werkbeziehungen anhand kartografischer Anwendungen zu visualisieren und intellektuellen-Netzwerke in Europa abzubilden. Auch spielbasierte Anwendungen ließen sich anhand des Datensets erstellen, z.B. in Form eines Werke-Quartetts, in das bestimmte werkbasierte Kategorien eingefügt werden, etwa die Anzahl oder Vielfalt an Beziehungen oder der Publikumserfolg. Diese möglichen Anwendungsbeispiele zeigen, wie normierte Daten auch außerhalb bibliothekarischer Zusammenhänge von Nutzen sind und auf spielerische Weise Wissen vermitteln und neue fächerübergreifenden Forschungsprojekte im Bereich Open Humanities initiieren können.

4. Fallbeispiele: Möglichkeiten & Desiderata in Katalogen

Im Folgenden wird anhand von je einem Fallbeispiel aus den beiden Zeiträumen präsentiert, wie Werknormdatensätze die Funktion zentraler Sucheinstiege übernehmen können, die unterschiedliche bibliografische Informationen bündeln, vernetzen und jeweils in den Komplex der in Beziehung stehenden Werke hinein- führen. Die unterschiedlichen Visualisierungen dieser Datensets zeigen bestehende Möglichkeiten, zentrale Werkinformationen übersichtlich darzustellen und komfortable Sucheinstiege anzubieten.

Anhand des Werknormsatzes für Johann Wolfgang von Goethes Briefroman *Die Leiden des jungen Werthers* wird die Varianz von Bearbeitungen im Kontext der Werk- und Rezeptionsgeschichte veranschaulicht. Die gegenwärtigen Möglichkeiten einer Visualisierung des Daten-

sets werden hier mithilfe des GND-Explorers umgesetzt (Abb. 1). Anhand der Relationen-Ansicht werden die Vor- und Nachteile im Vergleich mit der Darstellung im „Faktenblatt“ sowie im Vergleich mit der konventionellen Ansicht im Katalog der deutschen Nationalbibliothek (DNB) erläutert. Gut zu unterscheiden sind im GND-Explorer die Bezüge zu (literarischen) Vorbildern, Adaptionen und Bearbeitungen sowie zu anderen Entitäten mittels einer farblichen Markierung (Autor, Geografika). Die jeweiligen Vorlagen für den Werther einerseits und seiner Vorbildfunktion andererseits sind über das Faktenblatt überschaubarer. Wie die Thematik des Werthers bildhaft verarbeitet wurde, also zu Kunstwerken anregte, ist in der Ansicht ungünstig übersetzt und der Zusammenhang zwischen Grund- und Beziehungswerk nur über das Faktenblatt nachvollziehbar.

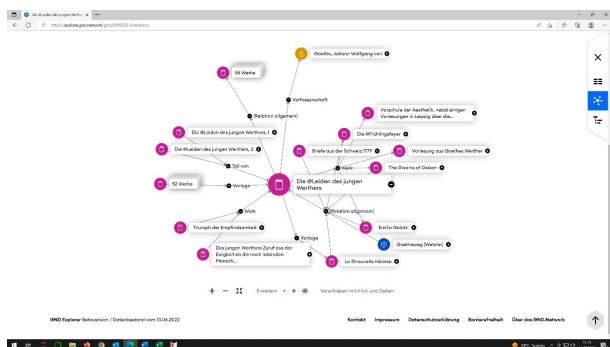


Abb. 1: „Die Leiden des jungen Werthers“, Quelle: GND-Explorer, Screenshot der Ansicht „Relationen“ Die @Leiden des jungen Werthers - Relationen - GND-Explorer(zugriffen: 19.07.2022)

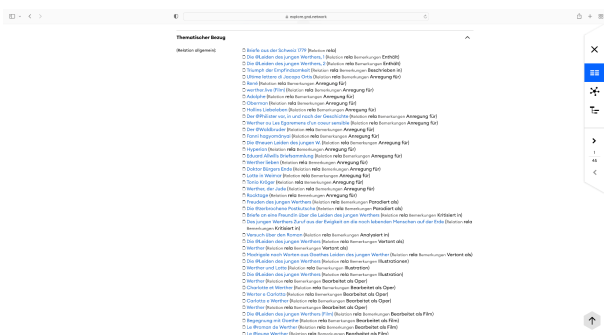


Abb. 2: „Die Leiden des jungen Werthers“, Quelle: GND-Explorer, Screenshot der Ansicht „Faktenblatt“ (Ausschnitt), Die @Leiden des jungen Werthers - Relationen - GND-Explorer (zugriffen: 19.07.2022)

Das zweite Beispiel ist Ilse Aichingers Roman *Die größere Hoffnung* von 1948. Hier wird der Werknormsatz im neuen Online-Katalog Kallias des DLA gezeigt (Abb. 2), dessen Beta-Version seit 2021 verfügbar ist. Unter der Detailansicht werden die mit dem Werk in Verbindung stehenden Bestände des DLA in vier Gruppen angeboten: unter „primäre Quellen“ findet man die Ausgaben des Romans und die Handschriften, unter „sekundäre Quellen“ die Literatur, die sich mit dem Roman beschäftigt, daneben werden Übersetzungen und Rezensionen zusammengestellt. Auf der rechten Seite befinden sich unterhalb der Abbildung eines Buchcovers und der Bezeichnung „Werkbeziehungen“ die im Werktitelprojekt

angelegten Werke. Folgt man einem Link, wird wiederum ein Werknormsatz, z.B. eines Hörspiels, angezeigt. Nicht zu allen diesen ermittelten Werken hat das DLA auch Bestände. Die Verknüpfung der Manifestationen mit Werknormdaten erfolgt bisher händisch bei der Katalogisierung oder in der Normdatenredaktion. Im Projekt ist die Entwicklung eines Tools geplant, das Manifestationen und Werke semi-automatisch verknüpft. Bis Ende 2022 wird zum Katalog ein Bereitstellungsdienst hinzugefügt, der es ermöglicht, auch sehr umfangreiche Datensets z.B. im csv-Format zu exportieren, um mit den Daten weiterzuarbeiten.

Abb. 3: „Die größere Hoffnung“, Quelle: Kallias – der Online-Katalog des Deutschen Literaturarchivs Marbach, Screenshot, <https://www.dla-marbach.de/find/opac/id/AK00140601> (zugriffen: 26.07.2022)

Wir freuen uns auf Ihre Fragen und Hinweise zur generellen Aufbereitung von Normdaten und zu unserem projektbezogenen Datenset. Wir möchten die in den Digital Humanities aktiven Wissenschaftler:innen über unseren Projektstand und die Potentiale werkbasierter Daten informieren und freuen uns über einen Austausch zu Auswertungen und Nachnutzungsoptionen der bereitgestellten Forschungsdaten.

Fußnoten

1. Von der DFG gefördert mit einer Laufzeit von 3 Jahren (2020-2023). Normdaten für Werke zur materialübergreifenden Vernetzung. Werktitel als Datenbasis für die Frage nach Verbreitungswegen von Literatur. - Archiv- und Forschungsbibliothek (klassik-stiftung.de), Werktitel als Wissensraum - DLA Marbach (dla-marbach.de) (zugegriffen: 25.07.2022)
2. Laut diesem Modell wird ein Werk „als eine intellektuelle oder künstlerische Schöpfung“ bezeichnet, d. h. es geht um den intellektuellen oder künstlerischen Inhalt (RDA, 5.1.2). Das Datenmodell der GND, in der wir die Werke erfassen, beruht auf FRBR und FRAD, diese wurden im IFLA Library Reference Model (IFLA LRM), siehe <https://www.iflastandards.info/lrm/lrmer.html> zusammengebracht. Zum Datenmodell der GND gibt es eine eigene Ontologie (https://d-nb.info/standards/elementset/gnd_20191015), die Klassen und Relationen definiert, mit welcher die GND-Daten in RDF beschrieben werden.
3. Die Finnische Nationalbibliothek bietet über einen Open Data Service (FENNICA) die Möglichkeit, werk-basierte Informationen zu finnischer Literatur einzusehen: Linked Data - National Library of Finland. Über die Französische Nationalbibliothek lassen sich ebenfalls werk-basierte Normdaten abrufen. Besonders eindrucksvoll präsentiert sich der Datenbestand der Library of Congress (LOC), welche mittels des BIBFRAME-Datenmodells die Daten aufbereitet. LC Linked Data Service: Authorities and Vocabularies | Library of Congress (loc.gov).
4. 4 "The Concept of a Work in WorldCat: An Application of FRBR" (2003) online unter: https://www.oclc.org/content/dam/research/publications/library/2003/lavoiie_frbr.pdf, hier Conclusion (Punkt 5): "Analysis suggests that concentrating on relatively large works, in particular those works whose content has been augmented, revised, or consists of collections of other works (a relatively small portion of the catalog) might be sufficient to capture the lion's share of benefits potentially available from implementing FRBR."
5. Friedrichs, Elisabeth 1991: Die deutschsprachigen Schriftstellerinnen des 18. und 19. Jahrhunderts : ein Lexikon. Stuttgart: Metzler; Gnüg, Hiltrud, Renate Möhrmann (Hrsg.) 1999: Frauen Literatur Geschichte. Schreibende Frauen vom Mittelalter bis zur Gegenwart. Stuttgart/Weimar: Metzler.
6. Diese Verteilung ergibt sich aus einer quantitativen Auswertung der 4.625 Werktitel aus dem Projekt.

Bibliographie

- Althage, Melanie. 2019. Normdaten – Knotenpunkte für den Wissensaustausch im Internet. Bericht zur Sitzung des Arbeitskreises Digital Humanities, WWU Münster, 1. Februar 2019. Normdaten – Knotenpunkte für den Wissensaustausch im Internet – CDH-Blog (uni-muenster.de) (zugegriffen: 25.07.2022).
- Barnert, Arno, Elisabeth Dietrich, Ines Kolbe und Karin Schmidgall. 2021. Vom Nutzen vernetzter Werke: Das Kooperationsprojekt »Werktitel als Wissensraum« des

Deutschen Literaturarchivs Marbach und der Herzogin Anna Amalia Bibliothek Weimar. In *ZfBB* 3, 68 : 138-151.

Baum, Constanze und Thomas Stäcker. 2015. Methoden – Theorien – Projekte. In: Grenzen und Möglichkeiten der Digital Humanities. In *Sonderband der Zeitschrift für digitale Geisteswissenschaften* 1. 10.17175/sb001_023 (zugegriffen: 21.07.2022).

Bicher, Katrin und Barbara Wiermann. 2018. Normdaten zu „Werken der Musik“ und ihr Potenzial für die digitale Musikwissenschaft. In *Preprints der Zeitschrift BIBLIOTHEK – Forschung und Praxis*. Normdaten zu „Werken der Musik“ und ihr Potenzial für die digitale Musikwissenschaft (hu-berlin.de) (zugegriffen: 25.07.2022).

Bischoff, Doerte und Susanne Komfort-Hein. 2019. Handbuch Literatur & Transnationalität. Berlin: De Gruyter.

Brown, Hillary und Gillian Dow. 2011. Readers, writers, salonnières: female networks in Europe, 1700 – 1900. Oxford, Bern (u.a.): Lang

Rippl, Gabriele und Simone Winko. 2013. Handbuch Kanon und Wertung: Theorien, Instanzen, Geschichte. Stuttgart: Verlag J. B. Metzler.

Seifert, Nicole. 2021. Frauen Literatur: abgewertet, vergessen, wiederentdeckt. Köln: Kiepenheuer & Witsch.

,Zu Rande kommen': Phänomen und Präsentation von Randnotizen am Beispiel der digitalen Ferdinand-Tönnies- Briefedition

Bamberg, Claudia

bamberg@uni-trier.de
Universität Trier, Deutschland

Dörk, Uwe

uwe.doerk@uni-due.de
Kulturwissenschaftliches Institut, Essen, Deutschland

Wierzock, Alexander

alexander.wierzock@hu-berlin.de
Kulturwissenschaftliches Institut, Essen, Deutschland

Trautmann, Tatjana

tatjana.trautmann@shlb.landsh.de
Schleswig-Holsteinische Landesbibliothek, Kiel

Burch, Thomas

burch@uni-trier.de
Universität Trier, Deutschland

Petkov, Radoslav

petkov@uni-trier.de
Universität Trier, Deutschland

Einleitung

In den Editionswissenschaften ist nicht unbemerkt geblieben, dass Randnotizen in Briefen (fortan: RN) keineswegs von marginaler Bedeutung sind. Vielmehr können sie aufgrund ihrer Länge, ihres inhaltlichen Gewichts und ihres rezeptionsästhetischen Reizes sogar die Hauptsache eines Briefs darstellen. Und das gilt vor allem für solche, die den Briefschreibenden selbst entstammen.¹ So wurde zu Theodor Fontanes „Kunst, auf Briefrändern zu schreiben“ etwa bemerkt, dass sie das Ergebnis „intensiver schriftstellerischer Arbeit“ seien, deren imposante „Architektur“ schwer zu entschlüsseln ist (Erler 1968: 318-319; Gabler 2020: 1239-1240). Damit stellen RN auch die Herstellung eines Transkripts und seiner grafischen Repräsentation vor eine große Herausforderung, die gewichtige editorische und im Fall von digitalen Editionen auch informationstechnologische Entscheidungen verlangt.

Trotzdem spielen RN in den *Digital Humanities* bzw. in digitalen Editionen bisher nur eine marginale Rolle: Die Suche nach Ansätzen, die editorisch und in Bezug auf ein geeignetes Datenmodell bereits bewährte Bau- und Fahrpläne an die Hand geben, erweist sich als eher vergeblich. Bisherige digitale Brief-Editionen geben RN meist konventionell wieder: Diplomatischen Editions-konventionen folgend werden sie topografisch, teils auch zusätzlich farblich abgesetzt präsentiert² oder an der zugehörigen Stelle hinzugefügt,³ bisweilen werden auch Arten von Randbeschriftungen unterschieden.⁴ Obgleich alle gesichteten Editionen ihre Transkriptionsregeln explizieren, begründen sie die Art ihres Umgangs mit RN nicht näher.

In den meisten Fällen ist ein solches Vorgehen ausreichend. Problematisch wird dieses erst, wenn RN mehrere Seiten umgreifen, unterschiedliche Sinneinheiten bilden, deren Abfolge gar schwer durchschaubar bzw. interpretationsbedürftig ist. Eine rein diplomatische Transkription führt dann zu fragmentierten Brieftexten, deren Sinneinheiten im Lesen erst mühsam rekonstruiert werden müssen. Besonders unbefriedigend ist eine solche Präsentationsweise für ‚User‘-Kontexte, deren Erkenntnisinteressen sich – wie in dem hier behandelten Fall von Ferdinand Tönnies – primär auf historische, politische, netzwerktheoretische, fach- und theoriegeschichtliche u.ä. Inhalte richten, aber nur selten poetische, literatur- und theorieästhetische Formen in den Blick nehmen. Für solche Fälle bedarf es anderer Lösungen. Wie ein solches hierfür spezifiziertes digitales Design aussehen kann, möchte dieser Vortrag zur Diskussion stellen. Er bezieht sich dabei auf das bereits entwickelte Konzept

einer digitalen Edition der Briefe des Soziologen Ferdinand Tönnies. Diese von der DFG geförderte Briefedition entsteht momentan am Kulturwissenschaftlichen Institut Essen, der Schleswig-Holsteinischen Landesbibliothek Kiel und am Trier Center for Digital Humanities (Kompetenzzentrum 2022).

Beobachtungen zu Randnotizen in Tönnies-Briefen

Die Praxis, briefkommunikativ über Bande zu spielen, tritt bei Tönnies nicht durchgängig, sondern nur episodisch in bestimmten, besonders intensiven Beziehungskonstellationen auf. Und genau dann spielen RN – darin mit Fontane vergleichbar – im brieflichen miteinander ‚zu Rande‘ wird. Kommen keine marginale, sondern eine zentrale Rolle. Wie zu zeigen sein wird, fungierten RN als Mittel einer Inklusionsstrategie, mit der die Intensität einer Freundschaft gesteigert und diese trotz Konflikten, divergierender Interessen oder Gefühlslagen aufrecht und produktiv gehalten

Ferdinand Tönnies: Gemeinschaft und Gesellschaft

Wer war Ferdinand Tönnies? Wissenschaftsgeschichtlich war er von besonderer Bedeutung, da er zu Beginn des 20. Jahrhunderts mit Max Weber, Georg Simmel und anderen die Soziologie als eine eigenständige Disziplin begründete. Sein 1887 erstmals erschienenes Werk „Gemeinschaft und Gesellschaft“ wurde nicht nur für die frühe Soziologie ein gesellschaftstheoretisches Standardwerk, sondern prägte auch die Kultur- und Bildungspolitik der Weimarer Republik. Das dort ausformulierte Konzept der „Gemeinschaft“, das er in Opposition zur strategischen Handlungssphäre der Gesellschaft konstruiert hatte, übte auf die politischen Diskurse der Republik einen großen Einfluss aus. Zugleich beeinflussten *Gemeinschafts-Konzeptionen* auch Tönnies' epistolare Praxis.

Besonders aufschlussreich ist hierbei die Rolle, die der Soziologe der Sprache attestierte. Er begriff sie nicht nur als Medium von Inhalten, sondern ebenso als das „wahre Organ“ von „Verständnis“ (Tönnies 2019: 144). Sprache lasse die Äußerung tiefer Gefühle zu. Gemüts-erregungen wie Schmerz, Lust, Furcht und Wunsch würden in Laute übersetzt, so dass Menschen aufgrund ihrer anthropologischen Befähigung zur „lebhaften [n] Sympathie“ zu einem wechselseitig „intimen Verständnis“ und zur gemeinschaftlichen „Übereinstimmung“ gelangen könnten (Tönnies 2009: 147). In dem von Tönnies besonders geschätzten Fall des intellektuellen Austauschs unter Freunden übersteigerte er das Übereinstimmen sogar zu einer „Art von unsichtbarer Ortschaft, eine[r] mystische[n] Stadt und Versammlung“, die er als geistige Gemeinschaft von anderen Vergemeinschaftungen abgrenzte (Tönnies 2019: 139, 137). Gerade die Briefe, in denen er intensiv RN gebrauchte, waren von dieser übersteigerten Gemeinschaftserwartung getragen. Um diesen bislang unbekannten Zusammenhang sichtbar zu

machen, musste auch für die digitale Edition eine spezifische Umsetzung gefunden werden.

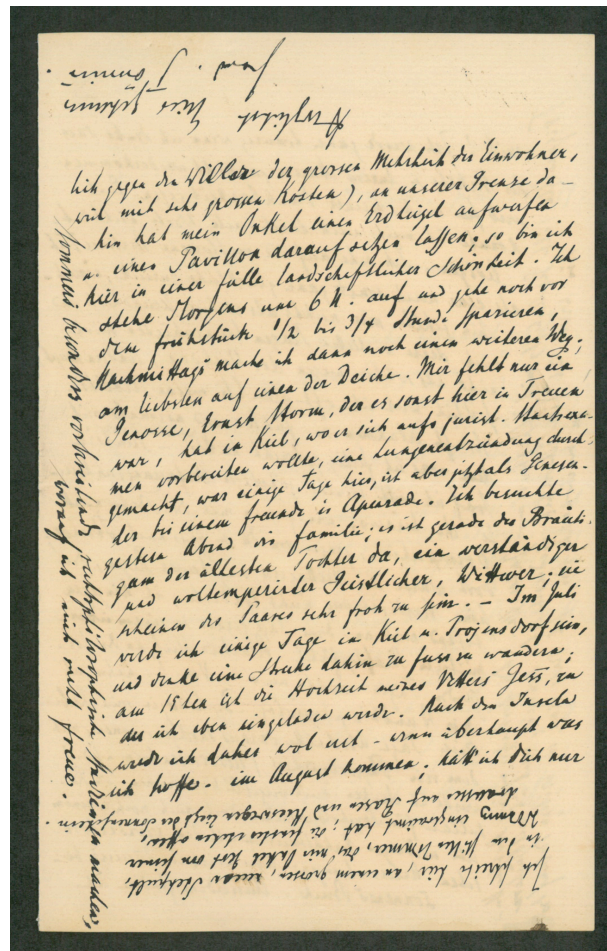
Tönnies als Briefschreiber

Tönnies wurde am 26. Juli 1855 in eine Welt der voll entfaltenen modernen Postinfrastruktur hineingeboren. Die Epochensignatur der Beschleunigung prozessierte gerade in der hochgradig brieflich geprägten Schriftkultur, wie sie für das Bürgertum des 19. und frühen 20. Jahrhunderts typisch war. In seiner „Kritik der öffentlichen Meinung“ von 1922 schrieb Tönnies: „In unendlichen Mengen schwirren heute Briefe [...] hin und her, am meisten innerhalb eines Landes, noch intensiver in engeren Gebieten, aber auch über die Grenzen von allen Orten, zu allen Orten des Erdballes“ (Tönnies 2002: 370). Das galt auch für Tönnies selbst, der im damaligen Wissenschaftsbetrieb bis zur Etablierung von Soziologie eine periphere Stellung einnahm und seine globalen wie lokalen Wissenschaftsbeziehungen primär über Briefe pflegte.

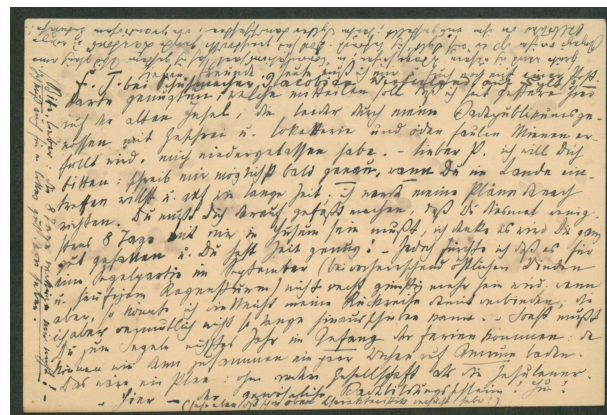
Die hohe Abhängigkeit vom Medium des Briefes schlug sich auch in einer permanenten Reflexion über Möglichkeiten und Grenzen dieser Kommunikationsform nieder, zumal Tönnies den Brief für ein unvollkommenes Surrogat der Kommunikation unter Anwesenden hielt. Trotzdem schrieb er fast täglich Briefe, da diese nicht nur Mittel zum Austausch mit Kollegen waren, sondern zugleich ein Medium zur Stiftung und Pflege sozialer Verbindungen im obigen Sinne.

Phänomen und Bedeutung von RN in Tönnies-Briefen

Das Aufkommen von RN lässt sich bei Tönnies ausschließlich in vertrauten Kommunikationssituationen beobachten. Nur nach einer Serie von Briefwechseln, nur nach wiederholten persönlichen Treffen in physischer Anwesenheit, nur nach oder einhergehend mit – diskret oder explizit – bekundeter Sympathie griff er auf dieses Stilmittel zurück. Je persönlicher und intensiver die Korrespondenz wurde, desto häufiger und komplexer setzte er RN ein. Kühlt die Beziehung wieder ab, wurden sie seltener und weniger komplex. Komplexität bezeichnet dabei das Maß an gleichzeitiger Rekursivität und Unübersichtlichkeit, wenn RN sich über mehrere Seiten hinweg erstrecken, in wechselnde Sinneinheiten und in unterschiedliche Arten untergliedern (z.B. fortlaufende Textabschnitte mit Apposition, Unterbrechung, Stern-Kommentar und Fußnotenentzitat).



Beispiel für einen Brief mit mehreren RN (Tönnies 1879: 3)



Beispiel für eine Postkarte mit mehreren RN (Tönnies 1881: 2)

Der Effekt für die Lesenden bestand zunächst darin, dass sie, bei gleichzeitig undeutlich werdender Schrift, eine Übersicht gewinnen, Reihenfolgen bilden und voneinander abgegrenzte Sinneinheiten und Verweiszusammenhänge rekonstruieren müssen. Systematisch zusammengefasst sind mit RN noch weitere kommunikativen Effekte verbunden:

A) Zunächst führt das Hinausschieben des Briefendes zu einer Verlängerung der Produktionssituation, wodurch

die briefliche Gemeinschaft mit der adressierten Person prolongiert wird.

B) Sodann bedingt der Zwang, die im Vergleich zum Kerntext i.d.R. schwerer lesbaren Randnotate zu entziffern und zu ordnen, eine längere Lesezeit und damit die Prolongierung der Rezeptionssituation.

C) RN setzen Impulse zur Anschlusskommunikation. Das gilt sowohl für den Schreibenden, der sich selbst durch RN zur Abfassung neuer RN oder zur Eröffnung neuer Briefseiten reizt, als auch für die Brieflesenden, die mit zusätzlichem Material für weitere Anschlusskommunikationen ‚gefüttert‘ werden. Diese Wirkung verstärkt sich durch eine spezifische Themenwahl. So streute Tönnies in RN bevorzugt Nachrichten aus dem geteilten sozialen Umfeld, so dass das, was ‚Tratsch‘ (Bergmann 1987: 198-202) genannt werden kann, zusätzlich zur Fortsetzung animierte.

D) Der Themenbereich Tratsch lässt das Bemühen erkennen, mittels diskret eingebrachter Informationen unterschiedliche Netzwerke miteinander zu verknüpfen und diese zu einem Kreis geistig Gleichgesinnter zusammenzuschließen (Simmel 1989: 237-257).

E) Häufig finden sich in RN jedoch auch Ergänzungen oder Präzisierungen literarischer oder terminologischer Art, so dass RN – insbesondere in wissenschaftlichen Kontexten – der literarischen Vernetzung dienen.

In der Summe des Zusammenwirkens von A, B, C und D ergibt sich nicht nur die Prolongierung der briefkommunikativen Gemeinschaft per Produktion und Rezeption, sondern auch eine verstetigte Kommunikation über episodische Einzelepisoden hinweg.

Erfassung und Darstellung von Randnotizen

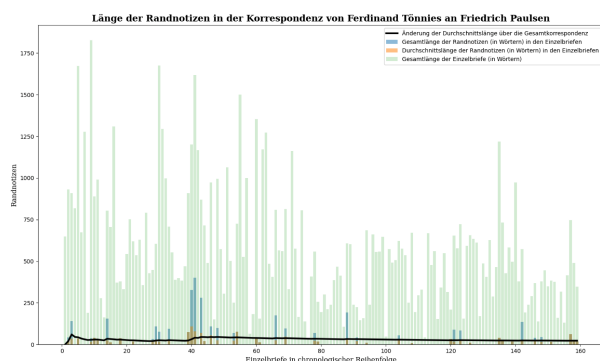
Aufgrund ihrer wichtigen Funktion musste für Randnotizen eine angemessene Präsentationsweise gefunden werden. Was aber heißt angemessen? Diese Frage lässt sich nicht allein aus der unterstellten Bedeutung von RN beantworten, sondern richtet sich zugleich an die Rezeptions- und Nutzungserwartung, die hochgradig durch die bisherige Leseerfahrung geprägt ist. Dabei ist zu beachten, dass in den Briefeditionen anderer zeitgenössischer soziologischer Fachklassiker wie Max Weber und Georg Simmel, an denen unser Editionsprojekt gemessen wird, Randnotizen keine oder fast keine Rolle spielen.⁵ Hinzu kommt eine Fachkultur, die an der Entwicklung von Theorien großes Interesse zeigt, nicht aber an sogenannten ‚philologischen Feinheiten‘. Der Brieftext hat daher primär das Kriterium der Lesbarkeit zu erfüllen. Sollte allerdings zugleich die Chance geboten werden, mithilfe des Transkripts das Faksimile entziffern zu lernen, dann wäre eine topografisch getreue Wiedergabe nötig.

Aus diesen divergierenden Ansprüchen an den Editionstext ergibt sich ein Zielkonflikt: Eine rein topografische Wiedergabe vermag den seitenübergreifenden Sinnzusammenhang von Randtexten nicht wiedergeben, so dass kein ‚lesbarer‘ Text entsteht. Eine Präsentationsweise, die Randnotizen in ihrem ‚Kontext‘ darstellt, eignet sich dagegen nicht als „Lesehilfe“. Die Lösungsmöglichkeit besteht somit darin, die digitale Briefedition mit beiden Möglichkeiten der Textwiedergabe auszustat-

ten und zugleich ein möglichst elegantes Wechseln der Präsentationsweise zu gewährleisten.

Die Dualität des Ansatzes muss daher mit einem geeigneten Datenmodell beschrieben und mit einer passenden Erfassungsmöglichkeit ausgestattet werden. Im Projekt geschieht dies während der Transkription eines Briefes in der virtuellen Forschungsumgebung des Softwareprogramms FuD (www.fud.uni-trie.de). Auch die Modellierung der RN erfolgt dann im Kontext der gesamten digitalen Briefedition innerhalb der dafür eingesetzten virtuellen Forschungsumgebung, die auf einem relationalen Datenbankmanagementsystem aufsetzt. Dabei erhalten die Randnotizen eindeutige Identifier, mittels derer die Verknüpfung von RN untereinander und innerhalb des Haupttextes abgebildet wird. Weitere Informationen, wie die topographische Positionierung der RN auf der Seite in Form von Metadaten (u.a. Angaben wie „Position: linker Rand; Drehung: 90°“ bzw. „Position: unterer Rand; Drehung: 180°“) werden auch erhoben. All diese Daten werden in unserem relationalen Entity-Relationship-Modell abgebildet und können somit in ihrer Gesamtkomplexität erfasst werden. Dadurch lassen sich die RN während der Transkription an den betreffenden Stellen einordnen; ebenso lässt sich die logische Lesereihenfolge konstruieren. Zusätzlich erlaubt das Analysemodul von FuD eine Annotation sämtlicher Textteile auf einer Metaebene und damit auch eine sich auf inhaltlicher Ebene bewegend Verknüpfung und Erschließung. Die konsequente Modellierung und Abbildung der Sachverhalte in einer relationalen Datenbank erlauben einen entsprechenden Export der Daten mithilfe einer geeigneten XML-Kodierung, für die im Laufe des Projektes eine Darstellung auf Basis der TEI-Guidelines (u.a. Modul „Linking, Segmentation and Alignment“) implementiert wird.

Über die Darstellung von Brieftexten und Randnotizen hinaus ermöglicht diese Art der Kodierung auch eine quantitative Analyse des Phänomens ‚RN‘ sowie eine gezielte Suche nach RN-Inhalten. So können einerseits für Fragen der Art „Bei welchen Korrespondenzpartnern verwendete Tönnies häufig RN?“, „Ändert sich die Häufigkeit von RN über den zeitlichen Verlauf einer Korrespondenz?“, „Wie verhält sich der Textumfang in den RN zur Gesamtlänge eines Briefes?“ usw. geeignete Visualisierungen erstellt werden. Andererseits lassen sich durch die Annotation der RN nach dem obigen Muster qualitative Fragen wie „In welchen Korrespondenzen treten vermehrt RN mit dem Thema ‚Tratsch‘ auf?“, „Wann werden welche ‚Nachrichten‘ bezogen auf die Gesamtkorrespondenz in den RN erwähnt?“ usw. beantworten. Auf Basis des Datenmodells können in der grafischen Benutzeroberfläche der Edition dann einerseits die einzelnen Randnotate eines aktuell gezeigten Digitalisats synoptisch oder die RN in ihrem übergreifenden Gesamtzusammenhang wiedergegeben werden. Andererseits können die Visualisierungskomponenten einzelne Befunde zu den vorgenannten Fragen in interaktiver Form anbieten, so dass die Benutzenden jederzeit von der Visualisierung Zugriff auf die Korrespondenzen haben. Durch eine modulare Anordnung der Visualisierungen lassen sich entsprechende Vergleichsmöglichkeiten schaffen. Eine einfache quantitative Analyse der RN innerhalb einer Korrespondenz zeigt die folgende Abbildung, in der die chronologische Verteilung der RN in 159 Briefen von Ferdinand Tönnies an Friedrich Paulsen abgebildet ist:



Man erkennt, dass Tönnies im Laufe der Kommunikation immer wieder RN verwendet, das Phänomen aber über die gesamte Dauer der Korrespondenz abnimmt.

Zusammenfassung

Wie gezeigt, spielen RN eine wichtige Rolle in der Art, wie Tönnies per Briefkommunikation bestimmte Netzwerke im Selbstverständnis nach *Gemeinschaft und Gesellschaft* strukturiert; als ein komplex genutztes Instrumentarium sind sie daher sowohl für die Art der Briefkommunikation als auch für die gelebte Praxis der von Tönnies entworfenen Theorie äußerst aufschlussreich. Aus diesem Grund musste für die Repräsentation der RN eine philologische, programmiertechnische und ästhetisch angemessene Lösung gefunden werden, die hier zur Diskussion gestellt wird.

Fußnoten

1. Im Unterschied zu solchen, die der Leserschaft wie etwa in Form des Glossierens, Hervorhebens, Verbesserns etc. zu verdanken und Gegenstand intensiver Rezeptionsästhetischer Forschung sind.
2. In der Briefedition August Wilhelm Schlegels werden etwa Randbeschriftungen im Fließtext farblich (schattiert) abgesetzt (Schlegel 2014-2021).
3. In der Briefedition Alfred Eschers (Jung 2022) werden mit Stern o. Ä. markierte Einfügungen "an der passenden Stelle des Brieftextes" integriert; Randnotizen oder nicht "eindeutig placierbare Ergänzungen" werden am Ende des Fließtextes wiedergegeben und per Mouseover im Faksimile mit einem gelben Rahmen sichtbar gemacht. Die Randbeschriftungen sind jedoch nie so komplex, dass sie mehrere Seiten umfassen. Siehe: <https://www.briefedition.alfred-escher.ch/briefe/B2616?action=search&view=single&odd=escher.od-d&view1=1#1.4.3.6.2.4>
4. So unterscheidet die Edition der Briefe Alexander von Humboldts etwa „Ergänzungen“ und „Anmerkungen“ farblich voneinander (Ette 2018-2022). Ergänzungen des Autors werden zudem ausgezeichnet und im Transkript bzw. der "Webansicht" an der als zugehörig interpretierten Stelle im Kerntext, farblich abgesetzt, eingefügt. Siehe: <https://edition-humboldt.de/richtlinien/ediarum.BASE/DE/text/ergaenzungen.html>

5. In der Briefedition der Simmel-Gesamtausgabe werden keine RN ausgewiesen. In der Max-Weber-Gesamtausgabe sind sporadisch RN enthalten. Sie werden im "textkritischen Apparat" (Weber 1994: 686, 851) ausgewiesen.

Bibliographie

- Bergmann, Jörg. 1987. *Klatsch. Zur Sozialform der diskreten Indiskretion*. Berlin: New York: De Gruyter.
- Erler, Gotthard. 1968. "Ich bin der Mann der langen Briefe" *Fontaneblätter* 1/7: 318-319.
- Ette, Ottmar (Hg.). 2018-2022. *Richtlinien der edition humboldt digital*, Berlin-Brandenburgische Akademie der Wissenschaften. <https://edition-humboldt.de/richtlinien/ediarum.BASE/DE/text/ergaenzungen.html> (zugegriffen: 2. August 2022).
- Gabler, Thorsten. 2020. "Theodor Fontanes Briefe" In *Handbuch Brief. Von der Frühen Neuzeit zur Gegenwart. Bd. 2: Historische Perspektiven - Netzwerke - Zeitgenossenschaften*, hg. von Marie Isabel Matthews-Schlingzig, Jörg Schuster, Gesa Steinbrink und Jochen Strobel, 1233-1244. Berlin; Boston: De Gruyter.
- Jung, Joseph (Hg.). 2022. *Digitale Briefedition Alfred Escher*, Relaunch Januar 2022, Zürich. <https://briefedition.alfred-escher.ch/briefe/B0503> (zugegriffen: 2. August 2022).
- Kompetenzzentrum – Trier Center for Digital Humanities. 2022. Ferdinand Tönnies-Briefe: Eine digitale Edition. <https://tcdh.uni-trier.de/de/projekt/ferdinand-toennies-briefe-eine-digitale-edition> (zugegriffen: 2. August 2022).
- Schlegel, August Wilhelm. 2014-2021. *Digitale Edition der Korrespondenz* [Version-01-22]. <https://august-wilhelm-schlegel.de>, hg. von Jochen Strobel und Claudia Bamberg, bearbeitet von Claudia Bamberg und Olivia Varwig in Zusammenarbeit mit Cornelia Bögel, Ruth Golschkin, Bianca Müller, Radoslav Petkov, Christian Senf und Friederike Wißmach (zugegriffen: 2. August 2022).
- Simmel, Georg. 1989. "Über soziale Differenzierung" In: *Georg Simmel. Aufsätze 1887 bis 1890. Über soziale Differenzierung. Die Probleme der Geschichtsphilosophie (1892)*, hg. von Heinz-Jürgen Dahme, 109-296. Frankfurt/M.: Suhrkamp.
- Tönnies, Ferdinand an Paulsen, Friedrich. 25.06.1879. In *Tönnies-Nachlass. Schleswig-Holsteinische Landesbibliothek*, Cb 54.51: Paulsen, 15.
- Tönnies, Ferdinand an Paulsen, Friedrich. 19.08.1881. In *Tönnies-Nachlass. Schleswig-Holsteinische Landesbibliothek*, Cb 54.51: Paulsen, 54.
- Tönnies, Ferdinand. 2002. "Kritik der öffentlichen Meinung" In *Ferdinand Tönnies. Gesamtausgabe Band 14. 1922*, hg. von Alexander Deichsel, Rolf Fechner und Rainer Waßner. Berlin; New York: De Gruyter.
- Tönnies, Ferdinand. 2009. "Philosophische Terminologie in psychologisch-soziologischer Ansicht" In *Ferdinand Tönnies. Gesamtausgabe Band 7. 1905-1906*, hg. von Arno Bammé und Rolf Fechner, 119-250. Berlin; New York: De Gruyter.
- Tönnies, Ferdinand. 2019. "Gemeinschaft und Gesellschaft" In *Ferdinand Tönnies. Gesamtausgabe Band 2. 1880-1935*, hg. von Bettina Clausen und Dieter Haselbach. Berlin; Boston: De Gruyter.

Weber, Max. 1994. "Briefe 1909-1910" In *Gesamtausgabe. Abteilung 2, Band 6*, hg. von M. Rainer Lepsius und Wolfgang J. Mommsen. Tübingen: J. C. B. Mohr (Paul Siebeck).

Doctoral Consortium

Constructing Multicultural Germany: Narratives on the Germany Men's National Football Team from 2006 to 2018

Kou-Herrema, Tianyi

koutiany@msu.edu

Michigan State University, USA; Universität Leipzig, Germany

Overview and Research Questions

My dissertation examines media narratives built around the multicultural German national football teams (both men and women) and argues that football not only helps pursue and formulate national identity, but also has become a battleground for both the promotion and contestation of German national identity. I aim to answer the following research questions: How was this multicultural team portrayed, utilized, and interpreted differently in the media between 2006 to 2018? What social forces created the demographically diverse national team and discourse surrounding them between this period? How has the rhetoric "multicultural national team" influenced ongoing debates over German national identity?

State of Research

This research builds on works drawn from the fields of sports studies, German studies, and digital humanities. Sports historian Kay Schiller (2015) warned readers not to confuse the rising acceptance of a multi-ethnic society with a positive endorsement of multiculturalism, which became a trend when studying football and national identity. Building on his argument, my research hopes to find out in which ways was the team perceived in the media as a representative and an endorsement of a supposed multicultural society. While Schiller recognized emerging discussions of multicultural society in debates over football nationalism, German studies scholars Stehle and Weber (2013) have focused on the phenomenon where media portrayed players differently based on their race rather than performance. However, most humanity scholars selected several articles and conducted close reading despite the number of materials they had at hand. With the help of computational methods, my work takes the approach of scalable reading (Müller) -- combining close and distant reading (Moretti, 2013),

and treats computational methods as an addition to rather than a replacement of close reading when analyzing cultural phenomena. I believe that only through detecting long-term trends and telling individual stories can one better answer the open-ended research questions formulated above.

Research Program

Since no databases exists containing news articles around the national teams, one of the central tasks of this dissertation is building a customized corpus consisting of news reports written and published between 2006 - 2018. The corpus consists of news reports collected from the LexisNexis database. This database contains enormous amount of data and is tricky to navigate. Therefore, I developed a strategy where I first read a sample collection of articles, documented and categorized key terms used in these articles, and eventually wrote three search strings to test which one gives me the best search results. Using the Precision at k metric, I determined the best search string that can be applied to LexisNexis' database. This innovative approach on extracting data from a rather complicated database could also be applied to other research in the future. As for now, I have completed data collection. To complete my dissertation, I need to finish data cleaning, and then conduct topic modeling and co-occurrence network analysis. Latent Dirichlet Allocation (LDA) is a widely used technique for topic modeling, which is the process of uncovering hidden topics in a collection of documents. Co-occurrence network analysis can help uncover hidden relationships and provide insights into the structure of my corpus. In this project, I use it to show the relationship between certain players and language used to describe them. These results could help me locate individual narratives that require closer examination. My analysis will then combine this qualitative approach with closely reading representative news articles. With this goal in mind, my work hopes to bring new perspectives to understand the complex role of the multicultural *Mannschaft* in the process of shaping contemporary German identity.

Schedule

- December 2022: Cleaning and testing corpus; conducting initial topic modeling analysis
- January - early March 2023: Conducting co-occurrence network analysis; compiling articles that need to be read closely; presenting at the 2023 DHd conference and receiving feedback
- April & May 2023: Writing the distant reading section of my dissertation, including descriptions of methods, process, and results; compiling articles that need to be read closely
- June - September 2023: Completing the qualitative section where I conduct analysis on news articles that touch on football players involved in political debates
- October - February 2023: Writing the three main chapters of my dissertation

- March and April 2024 – Finishing the Epilogue chapter and revising all chapters

sprünglich als Philologie mit dem Interesse an Sprache und Text, weniger mit deren Veräußerung in Bildern.

Bibliographie

Moretti, Franco. 2013. *Distant Reading*. 1st edition. Verso.

Müller, Martin. n.d. "Scalable Reading." Accessed August 3, 2022. <https://scalablereading.northwestern.edu/>.

Schiller, Kay. 2015. "Siegen Für Deutschland? Patriotism, Nationalism and the German National Football Team, 1954-2014." *Historical Social Research / Historische Sozialforschung* 40 (4 (154)): 176–96.

Stehle, Maria, and Beverly M. Weber. 2013. "German Soccer, the 2010 World Cup, and Multicultural Belonging." *German Studies Review* 36 (1): 103-124.

Diagramme edieren – zur kritischen Repräsentation visueller Narrative

Sutor, Nadine

sutor@uni-wuppertal.de

Bergische Universität Wuppertal, Deutschland

Einleitung und Problemstellung

Zu allen Zeiten haben sich Menschen ein Bild von der Welt gemacht und festgehalten, wie sie diese verstanden und interpretiert haben. Seit ihren skizzenhaften Anfängen ist bis heute eine Vielzahl von schematischen Bildern entstanden. "Praktiken visueller Welterzeugung" (Reudenbach 2011, Vorbemerkung) in Form von Zeichnungen lassen sich bereits in der Antike beobachten und haben sich bis heute als Mittel zur Konstruktion von Ordnungsvorstellungen bewährt. Anschaulichkeit als grundlegende Kategorie für das Verständnis von der Welt manifestiert sich auch im Diagramm. Das wissenschaftliche Interesse an der Diagrammatologie ist in den letzten Jahrzehnten stark gestiegen. In den digitalen geistes- und sozialwissenschaftlichen Fächern wurde die Darstellung abstrakter Daten und Zusammenhänge in graphisch-visuell erfassbarer Form immer stärker zu einer wichtigen Quelle bei der Generierung von Wissen (Lancaster, Schaal 2016, 5). Es wirkt der sogenannte „visual turn“, der sich abwendet von einer rein sprachlichen Wissensvermittlung und den Fokus stattdessen auf bildhafte Narrative legt: Die textuelle Ebene wird ergänzt durch die visuelle Dimension. Bisher ist die Editionswissenschaft eher unreflektiert mit der Frage umgegangen, wie man Diagramme als bildhafte Darstellungen kritisch wiedergeben kann, da es bis dato keine editorische Theorie der Diagramme gibt. Die Editorik versteht sich ur-

Quellen

Das Thema der Promotion ist grundsätzlich transdisziplinär angelegt. Zwar handelt es sich bei den Fallstudien um früh- bzw. hochmittelalterliche Texte, allerdings spielen die philologische und die historische Dimension nur eine Nebenrolle. Stärker wird der Blick auf kulturwissenschaftliche und medientheoretische oder gar kunsthistorische Fragen zu richten sein. Letztlich geht es um editorische Fragen, die verschiedene Disziplinen betreffen. Zwei Quellen sollen für eine Analyse unter editionswissenschaftlichen Gesichtspunkten untersucht werden. Die in zahlreichen Handschriften durch das Mittelalter überlieferte Kosmologie *De natura rerum* des Isidor von Sevilla (560-636) und das bekannteste Werk des Petrus von Poitiers (1125/1130-1205) *Compendium historiae in genealogia christi*. Sie markieren einerseits in dem die Antike tradierenden Frühmittelalter und andererseits in dem hier für Innovation stehenden Hochmittelalter das Entstehen der Diagrammatik im engeren, heutigen Sinne.

Isidor von Sevilla behandelt naturkundliche Themen. Er nutzt seine Diagramme als Erklärung von mathematisch-physikalischen Konzepten, die inhaltlich logisch und schlüssig, jedoch zu komplex sind, als dass man sie textuell in Form eines Narrativs beschreiben könnte. Die diagrammatischen Darstellungen sollen diese Prozesse bildhaft darstellen und so ihr Verständnis legitimieren. Das „Compendium“ von Petrus von Poitiers ist wegen seiner Rezeption für die in den folgenden Jahrhunderten entstandenen graphischen Visualisierungen von Geschichte von großer Bedeutung. Er nutzt mehrere Diagrammformen, um die biblische Erzählung mit Erläuterungen zu versehen, bzw. durch bildhafte Darstellungen verständlicher zu machen. Seine Diagramme machen etwas, das verbal beschrieben wird als visuelle Struktur sichtbar und zeigen damit, dass die textliche Beschreibung und das damit gemeinte jeweils unterschiedlich interpretiert werden kann. Er wählt das Diagramm als eine Form der Wissensvermittlung, die über Sprache hinausgeht.

Forschungsfragen und Methode

Mit Blick auf die einleitend formulierte Problemstellung können zwei zentrale Forschungsfragen aufgezeigt werden:

1. Editorische Herausforderung: Was ist aus der klassischen Textkritik auf die „Editorik der Diagramme“ übertragbar? Ziel der Dissertation ist keine Edition beider Werke, sondern die Entwicklung einer „Diagrammkritik“ und ein dafür aufgestelltes Regelwerk, welches explizit auf die beiden vorgestellten Quellen angewendet werden soll.
- 2.

Re-medialisierung und Re-Codierung von Diagrammen: Gegenstand des praktischen Teils ist die Entwicklung von Formen einer kritischen Wiedergabe diagrammatischer Darstellungen. Das Konzept der Repräsentation als Skala geht mit der Frage einher, wie man ein Diagramm für heute „sprechend“ und verständlich machen kann. Beginnend bei einem quellennahen Abbild über eine fortschreitende Abstraktion, Normierung und Idealisierung zu immer mehr "Nutzer*innennähe."

Mit SVG als Verfahren der digitalen Editorik ist die Methode zu benennen, die im Praxisteil der Promotion Anwendung finden soll. Für die digitale Bildrepräsentation soll unter Hinzunahme von SVG als XML-basierter Technologie die unter 2. formulierte Frage diskutiert werden, inwieweit eine editorisch naheliegende oder eine auf Ästhetik abzielende Mimetik durch eine systematisierende Wiedergabe ergänzt werden kann. Die Realisierung unterschiedlicher Abstraktionsstufen gibt einerseits Aufschluss über Entstehungskontexte des Diagramms, über Schreiberspezifika oder offenbart stemmatologische Nachbarschaften und Abfolgen. Sie ermöglicht andererseits die Produktion abstrahierender und idealisierender Abbilder und konfrontiert die Quelle mit einer gegenwartsbezogenen Perspektive: Was wäre eine zeitgemäße Form der Wiedergabe?

Bezug zu Themen aus den Digital Humanities

Das Thema dieser Arbeit bietet Anknüpfungspunkte zu weiteren, durchaus diskussionswürdigen Themen in den DH: Wie können Kulturartefakte codiert werden? Wie werden sie re-medialisiert? Wie können wir Relationen mentaler Denkstrukturen und medialen Ausdrucksformen systematischer aufdecken? Wie beeinflussen Technologien und Medien, die uns zur Verfügung stehen, wie wir unsere Welt sehen und mit ihr umgehen?

Bibliographie

Anderson, Benjamin. 2022. "Between Diagramm and Image On Yuval's Harp." In *The Diagram as Paradigm. Cross-Cultural Approaches*, hg. von Jeffrey F. Hamburger, David J. Roxburgh und Linda Safran. 93-113. Cambridge: Harvard University Press.

Assmann, Jan. 2012. "Schriftbildlichkeit. Etymographie und Ikonographie." In *Schriftbildlichkeit. Wahrnehmbarkeit, Materialität und Operationalität von Notationen*, hg. von Sybille Krämer, Eva Cancik-Kirschbaum und Rainer Totzke, 139-149. Berlin: Akademie Verlag.

Bauer, Ernst. 2010. "Grundzüge der Diagrammatik." In *Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld*, hg. von Matthias Bauer und Christoph Ernst, 17-109. Bielefeld: transcript.

Brandstetter, Gabriele. 2012. "Schriftbilder des Tanzes. Zwischen Notation, Diagramm und Ornament." In *Schriftbildlichkeit. Wahrnehmbarkeit, Materialität und*

Operativität von Notationen, hg. von Sybille Krämer, Eva Cancik-Kirschbaum und Rainer Totzke, 61-79. Berlin: Akademie Verlag.

Eastwood, Bruce S. 2001. "The Diagram of the Four Elements in the Oldest Manuscript of Isidore's 'De natura rerum'." *Studi Medievali* 42: 547-564.

Ernst, Christoph. 2021. "Ikonizität, Schema und Diagramm." In *Diagramme zwischen Metapher und Explikation. Studien zur Medien- und Filmästhetik der Diagrammatik*, hg. von Christoph Ernst. Präsenz und implizites Wissen 5: 153-164.

Frank, Ingo. 2017. "Diagrammatische Denkerwerkzeuge in den Digital Humanities – Ansatz zur zeichentheoretischen Grundlegung." In *Semiotik als Theorie der Digitalen Geisteswissenschaften*, hg. von Martin Siefkes und Roland Posner. Zeitschrift für Semiotik (1-2) 39: 51-83.

Giardino, Valeria und Gabriel Greenberg. 2015. "Varieties of Iconicity." *Review of Philosophy and Psychology* 6: 1-25.

Krämer, Sybille. 2014. "Zur Grammatik der Diagrammatik. Eine Annäherung an die Grundlagen des Diagrammgebrauchs." In *Diagramm und Narration*, hg. von Hartmut Bleumer. Zeitschrift für Literaturwissenschaft und Linguistik (176) 44: 11-28.

>Krämer, Sybille. 2005. "Operationsraum Schrift: Über einen Perspektivwechsel in der Betrachtung von Schrift." In *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine*, hg. von Gernot Grube, Werner Kogge und Sybille Krämer, 23-61. München: Wilhelm Fink Verlag.

Krämer, Sybille. 2012. "Punkt, Strich, Fläche. Von der Schriftbildlichkeit zur Diagrammatik." In *Schriftbildlichkeit. Wahrnehmbarkeit, Materialität und Operativität von Notationen*, hg. von Sybille Krämer, Eva Cancik-Kirschbaum und Rainer Totzke, 79-101. Berlin: Akademie Verlag.

Lancaster, Kelly und Schaal, Gary S. (2016). "Ein Bild sagt mehr als 1000 Worte? Visualisierungen in den Digital Humanities." In *Digital Classics Online*, hg. von Roxana Kath, Michaela Rücker, Reinhild Scholl, Charlotte Schubert, (2, 3): 5-22.

Manolova, Divna. 2022. "Space, Place, Diagramm. Cleomedes and the Visual Program of Munich, Bayerische Staatsbibliothek, Cod.gr. 482." In *The Diagram as Paradigm. Cross-Cultural Approaches*, hg. von Jeffrey F. Hamburger, David J. Roxburgh und Linda Safran. 149-167. Cambridge: Harvard University Press.

Mersch, Dieter. 2012. "Schrift/Bild – Zeichnung/Graph – Linie/Markierung. Bildepisteme und Strukturen des ikonischen 'Als'." In *Schriftbildlichkeit. Wahrnehmbarkeit, Materialität und Operativität von Notationen*, hg. von Sybille Krämer, Eva Cancik-Kirschbaum und Rainer Totzke, 305-329. Berlin: Akademie Verlag.

Müller, Kathrin. 2008. "Visuelle Weltaneignung: astronomische und kosmologische Diagramme in Handschriften des Mittelalters." PhD diss., Universität Hamburg.

Raible, Wolfgang. 2012. "Bildschriftlichkeit." In *Schriftbildlichkeit. Wahrnehmbarkeit, Materialität und Operativität von Notationen*, hg. von Sybille Krämer, Eva Cancik-Kirschbaum und Rainer Totzke, 201-217. Berlin: Akademie Verlag.

Reudenbach, Bruno. 2011. "Ein Weltbild im Diagramm – Ein Diagramm als Weltbild. Das Mikrokosmos-Makrokosmos-Schema des Isidor von Sevilla." In *Atlas der Weltbilder*, hg. von Christoph Marschikies, Ingeborg Reichle,

Jochen Brüning und Peter Deuffhard, 33-40. Berlin: Akademie Verlag.

Smets, Alexis und Christoph Lüthy. 2009. "Words, Lines, Diagrams, Images: Towards a History of Scientific Imagery". *Early Science and Medicine* 14: 398-439.

Stetter, Christian. 2005. "Bild, Diagramm, Schrift." In *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine*, hg. von Gernot Grube, Werner Kogge und Sybille Krämer, 115-137. München: Wilhelm Fink Verlag.

Treude, Linda und Sascha Freyberg. 2012. "Diagrammatik und Wissensorganisation." *LIBREAS. Library Ideas* 21.

Wallis, Faith. 2015. "What a Medieval Diagram Shows: A Case Study of 'Computus'." *Studies in Iconography* 36: 1-40.

Wallis, Faith und Calvin B. Kendall. 2016. *Isidore of Seville. On the Nature of Things*. Translated Texts for Historians 66.

W.J.T. Mitchell. 1987. *Iconology. Image, Text, Ideology*. Chicago: University of Chicago Press.

Worm, Andrea. 2020. *Geschichte und Weltordnung. Graphische Modelle von Zeit und Raum in Universalchroniken vor 1500*. Berlin: Deutscher Verlag für Kunstwissenschaft.

Worm, Andrea. 2018. "Medium und Materialität: Petrus von Poitiers' Compendium historiae in genealogia Christi in Rolle und Codex." In *Codex und Material*, hg. von Patrizia Carmassi und Gia Toussaint. Wolfenbütteler Mittelalter-Studien 34: 39-64, Wiesbaden: Harrassowitz Verlag.

Worm, Andrea. 2013. "Visualising the Order of History: Hugh of Saint Victor's Chronicon and Peter of Poitiers' Compendium Historiae." In *Romanesque and the Past: Retrospection in the Art and Architecture of Romanesque Europe*, hg. von Richard Plant und John McNeill, 243-264. London: Cambridge University Press.

Erweiterbare, interaktive Softwareplattform für die Anwendung von Sprachtechnologie in großen Textkorpora zur Unterstützung von Such- und Analyseworkflows in den Digital Humanities

Petersen-Frey, Fynn

fynn.petersen-frey@uni-hamburg.de
Universität Hamburg, Deutschland

Motivation und Einleitung

In Wissenschaftsdisziplinen wie den digitalen Geistes- und Sozialwissenschaften „Digital Humanities“ (DH), besteht ein großes Interesse daran, umfangreiche Mengen von Text, z.B. aus historischen Zeitschriften (Purschwitz 2018, 109-142), sozialen Medien (Stier u. a. 2017, 1365-1388) oder Zeitungen, hinsichtlich verschiedener Fragestellungen auszuwerten. Dabei ist oft eine Kombination von qualitativen Analyseschritten mit quantitativen Auswertungen gefragt (Stulpe und Lemke 2016, 17-61).

Eine ausschließlich manuelle Bearbeitung ist angesichts des Umfangs und des daraus resultierenden Aufwands oftmals ausgeschlossen. In diesen Fällen ermöglicht es eine (teil-)automatisierte computerlinguistische Verarbeitung dennoch Analyse und Auswertung durchzuführen.

Ziel dieses Dissertationsvorhabens ist es daher geeignete Methoden zu entwickeln und in eine moderne Sprachtechnologieplattform zu integrieren, die es Wissenschaftlern aus den DH ermöglicht, selbstständig große Textkorpora mit Hilfe (teil-)automatischer Verarbeitungsschritte aus der CL vorzubereiten und hinsichtlich vielfältiger Fragestellungen auszuwerten. Dies umfasst lexikalische wie semantische Suchfunktionen, intuitive Annotationsfunktionen, AI- bzw. Human-in-the-Loop Funktionalitäten zur (teil-)automatischen Skalierung von Annotationen auf große Korpora, über klassische syntaktische Anreicherungen hinausgehende inhaltliche NLP-Methoden wie Koreferenzausflösung und Zitaterkennung sowie qualitative und quantitative Analysemöglichkeiten.

Forschungsstand

Verwandte Software aus der Computerlinguistik wie *brat* (Stenetorp u. a. 2012, 102-107), *WebAnno* (Eckart de Castilho u. a. 2016, 76-84), *INCEpTION* (Klie u. a. 2018, 5-9) oder *TextAnnotator* (Abrami u. a. 2020, 891-900) sind vorrangig für linguistische Annotationen gedacht, um annotierte Korpora zu erstellen; weniger jedoch für die inhaltliche Bearbeitung einer DH-Forschungsfrage mittels Suche, Annotation, Aggregation und Analyse.

Für spezifische Recherchezwecke existieren computerlinguistische Anwendungen wie *SiNLP* (Crossley u. a. 2014, 511-534) zur Diskursanalyse, *ALCIDE* (Moretti u. a. 2016, 100-112) zur Analyse von historischem und politischem Diskurs, *new/s/leak* (Wiedemann u. a. 2018, 313-322) zum Entdecken berichtenswerter Geschehnisse in großen Textkorpora sowie *LawStats* (Ruppert u. a. 2018, 212-222) zur Suche und Analyse von Revisionen des Bundesgerichtshofs. Diese Anwendungen sind jedoch weniger geeignet andere Fragestellungen zu bearbeiten.

In den Sozialwissenschaften gängige qualitative Analysetools wie *MaxQDA* oder *atlas.ti* verfügen über Annotations- und qualitative Analysefunktionen, sind aber proprietär, erlauben kaum Kollaboration und bieten keine modernen NLP-Methoden.

Recogito (Simon u. a. 2017, 111-132) und *CATMA* (Gius u. a., 2022) sind intuitive Annotationsanwendungen für

die DH mit Kollaborationsfunktionen. *Recogito* setzt einen Fokus auf Orte und Integration in eine Karte mittels automatischer Erkennung von Entitäten. *CATMA* bietet konfigurierbare Annotationsschemata und vielfältige Analysefunktionen. Beide verfügen jedoch nur über eine bescheidene bzw. keine Suche, wenig Unterstützung für große Korpora und keine Möglichkeit zur Skalierung manueller Annotationen.

WebLicht (Hinrichs u. a. 2010, 25–29) integriert enorm viele NLP-Module, die zu individuellen Pipelines zusammengefügt werden können. *Nopaque* (Universität Bielefeld 2022) unterstützt historische Dokumente mittels OCR und syntaktische Analysen anhand Keyword-In-Context-Suche auf Basis gängiger NLP-Modelle und individueller NLP-Pipelines. Beiden fehlen jedoch Annotationsmöglichkeiten, Kollaborationsfunktionen, Dokumentensuche sowie Möglichkeiten zur quantitativen Analyse großer Korpora.

ILCM (Niekler u. a. 2018, 1313–1319) ist eine Textmining-Umgebung für die datengetriebene Forschung auf der großen Textmengen mittels statistischer Analysen. Es fehlen jedoch moderne NLP-Modelle sowie interaktive Funktionalitäten wie Annotation oder Suche.

Forschungsvorhaben und Forschungsfragen

Das Forschungsvorhaben steht unter der übergreifenden Forschungsfrage: Wie kann eine digitale Arbeitsumgebung entwickelt werden, welche undogmatisch die Anwendung von DH-Methoden durch moderne NLP-Methoden für Suche, Annotation, Skalierung, Aggregation und Analyse auf großen Korpora unterstützt?

Das Ziel ist es geeignete Methoden für eine Sprachtechnologieplattform zu entwerfen, die intuitiv aus den DH genutzt werden kann, um manuelle Arbeit mit Textdokumenten automatisch auf große Korpora zu skalieren. Dies umfasst u. a. eine semantische Suche um ähnliche Aussagen zu einer bestimmten Textpassage im gesamten Korpus zu finden, die Möglichkeit gefundenen Textpassagen qualitativ manuell zu analysieren oder automatisch zu aggregieren für quantitative Auswertungen. Im Zusammenspiel von Entity-Linking, Koreferenzauflösung, Zitaterkennung und Auffinden ähnlicher Textpassagen soll eine Skalierung manueller Annotation bzw. Kodierung von Textspannen auf den gesamten Korpus ermöglicht werden, indem Annotationen einzelner Textstellen auf passende Stellen gesamten Korpus angewandt wird.

Wie können dem aktuellen Stand der Forschung entsprechende computerlinguistische Methoden undogmatisch und intuitiv für die Bearbeitung verschiedener DH-Fragestellungen nutzbar gemacht werden?

Zum aktuellen Stand der Forschung zählen kontextualisierte Embeddings wie *BERT* (Devlin u. a. 2019, 4171–4186), die jedem Wort und (Ab-)Satz eine Bedeutung in Abhängigkeit des Kontexts anhand der Position in einem hochdimensionalen Vektorraum zuordnen. Wie lässt sich auf dieser Basis eine semantische Ähnlichkeitssuche zum Auffinden verwandter Aussagen mit variierender Länge entwickeln? Dabei ist zu klären, welche

Lösungen für die rechenintensiven Operationen bei der Ähnlichkeitssuche mit großen Korpora skalieren.

Um diese NLP-Methoden nutzbar zu machen, werden sie in eine webbasierte Benutzeroberfläche integriert, die das Durchsuchen, Annotieren, Skalieren, Aggregieren und Auswerten großer Korpora mittels der zuvor beschriebenen computerlinguistischen Funktionalitäten unterstützt. Ferner wird geeignete Softwarearchitektur entworfen, welche die Integration bestehender und zukünftiger CL-Softwarebibliotheken ermöglicht.

Bibliographie

Abrami, Giuseppe, Manuel Stoeckel und Alexander Mehler. 2020. "TextAnnotator: A UIMA Based Tool for the Simultaneous and Collaborative Annotation of Texts". English. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, S. 891–900. isbn: 979-10-95546-34-4. url: <https://www.aclweb.org/anthology/2020.lrec-1.112>.

Crossley, Scott A., Laura K. Allen, Kristopher Kyle und Danielle S. McNamara. 2014. "Analyzing Discourse Processing Using a Simple Natural Language Processing Tool". In: *Discourse Processes* 51.5-6, S. 511–534. doi: 10.1080/0163853X.2014.910723.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. 2019. "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, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, S. 4171–4186. doi: 10.18653/v1/N19-1423. url: <https://www.aclweb.org/anthology/N19-1423>.

Eckart de Castilho, Richard, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank und Chris Biemann. 2016. "A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures". In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee, S. 76–84. url: <https://www.aclweb.org/anthology/W16-4011>.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, und Jan Horstmann. 2022. "CATMA". Zenodo, url: <https://doi.org/10.5281/zenodo.6419805>.

Hinrichs, Erhard W., Marie Hinrichs and Thomas Zastrow. 2010. "WebLicht: Web-Based LRT Services for German". In: *Proceedings of the ACL 2010 System Demonstrations*. S. 25–29. url: <http://www.aclweb.org/anthology/P10-4005>

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho und Iryna Gurevych. 2018. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, S. 5–9. url: <http://tubiblio.ulb.tu-darmstadt.de/106270/>.

Moretti, Giovanni, Rachele Sprugnoli, Stefano Menini und Sara Tonelli. 2016. "ALCIDE: Extracting and visualising content from large document collections to support humanities studies". In: *Knowledge-Based Systems 111*, S. 100–112. issn: 0950-7051. doi: <https://doi.org/10.1016/j.knosys.2016.08.003>. url: <http://www.sciencedirect.com/science/article/pii/S0950705116302635>.

Niekler, Andreas, Arnim Bleier, Christian Kahmann, Lisa Posch, Gregor Wiedemann, Kenan Erdogan, Gerhard Heyer und Markus Strohmaier. 2018. "ILCM - A Virtual Research Infrastructure for Large-Scale Qualitative Data". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), S. 1313–1319 url: <https://www.aclweb.org/anthology/L18-1209>.

Purschwitz, A. (2018). "Netzwerke des Wissens – Thematische und personelle Relationen innerhalb der halleschen Zeitungen und Zeitschriften der Aufklärungsepoche (1688–1818)". In: *Journal of Historical Network Research 2* (1), S. 109–142.

Ruppert, Eugen, Dirk Hartung, Phillip Sittig, Tjorben Gschwander, Lennart Rönneburg, Tobias Killing und Chris Biemann. 2018. LawStats – Large-Scale German Court Decision Evaluation Using Web Service Classifiers. In: *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018*, Hamburg, Germany, August 27–30, 2018, Proceedings. Springer-Verlag, Berlin, Heidelberg, 212–222. doi: 10.1007/978-3-319-99740-7_14

Simon, Rainer, Elton Barker, Leif Isaksen und Pau De Soto Caameres. 2017. "Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2". In: *Journal of Map & Geography Libraries*, 13:1, S. 111–132. doi: 10.1080/15420353.2017.1307303

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou und Jun'ichi Tsujii. 2012. "brat: a Web-based Tool for NLP-Assisted Text Annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, S. 102–107. url: <https://www.aclweb.org/anthology/E12-2021>.

Stier, Sebastian, Lisa Posch, Arnim Bleier und Markus Strohmaier. 2017. "When populists become popular: comparing Facebook use by the right-wing movement Pegida and German political parties". In: *Information, Communication & Society 20*, S. 1365–1388. doi: 10.1080/1369118X.2017.1328519.

Stulpe, Alexander und Matthias Lemke. 2016. "Blended Reading". In: *Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Hrsg. von Matthias Lemke und Gregor Wiedemann. Wiesbaden: Springer Fachmedien Wiesbaden, S. 17–61. isbn: 978-3-658-07224-7. doi: 10.1007/978-3-658-07224-7_2. url: https://doi.org/10.1007/978-3-658-07224-7_2.

Wiedemann, Gregor, Seid Muhie Yimam und Chris Biemann. 2018. "New/s/leak 2.0 – Multilingual Information Extraction and Visualization for Investigative Journalism". In: *Social Informatics*. Hrsg. von Stefan Staab, Olessia Koltsova und Dmitry I. Ignatov.

Cham: Springer International Publishing, S. 313–322. isbn: 978-3-030-01159-8.

Universität Bielefeld. 2022. "nopaque". <https://nopaque.uni-bielefeld.de> (zugegriffen 12. Dezember 2022)

Feministische Filmgeschichte als Linked Open Data: Ein Thesaurus für das Women Film Pioneers Project (WFPP)

Junginger, Pauline

pauline.junginger@uni-marburg.de

Philipps-Universität Marburg, Deutschland

Das *Women Film Pioneers Project* (WFPP) ist eine etablierte Online-Ressource für die Forschung zu Frauen im Frühen Kino (Gaines, Vatsal und Dall'Asta, o. J.). Der Fokus des WFPP liegt auf dem Erzählen individueller Geschichten von Filmpionierinnen und der Sichtbarmachung blinder Flecken, um die Aufmerksamkeit auf das zu lenken, was selbst in der feministischen Filmtheorie lange unvorstellbar schien: Die große Rolle die Frauen zu Beginn der Filmgeschichte gespielt haben (Dang 2020). Anfang der 1990er Jahre zunächst als klassisches Buchprojekt konzipiert, wurde das WFPP 2013 als digitale Plattform veröffentlicht, die vorrangig drei Ziele verfolgt: die Verfügbarmachung von Forschungsergebnissen, die Anregung weiterer Forschung und das Hinterfragen klassischer filmhistorischer Narrative, in denen die Beiträge von Frauen unerwähnt bleiben. Obwohl strukturierten Metadaten eine zentrale Bedeutung bei der Sichtbarmachung, Zugänglichkeit und Nachnutzung von Forschungsdaten und digitalen Publikationen zukommt (Baca 2016; Flanders und Jannidis 2018), arbeitet das WFPP bisher nur sehr eingeschränkt mit Metadaten.

¹ Das Ziel meiner Dissertation ist deshalb die Entwicklung eines Thesaurus und dazugehöriger Annotationsrichtlinien für die semantische Verschlagwortung der Profile über Filmpionierinnen. Die zentralen Forschungsfragen des Dissertationsprojekts lauten dabei: Wie kann ein Thesaurus für die feministische Filmgeschichtsschreibung am Beispiel des WFPP entwickelt und angewendet werden? Wie kann die Forschung zu Frauen im Frühen Kino mit Hilfe strukturierter Metadaten sichtbarer und zugänglicher gestaltet werden? Welche Geschichte kann durch das WFPP (nicht) erzählt werden?

Für die Entwicklung eines Thesaurus für das WFPP sind neben der feministischen Filmgeschichtsschreibung verschiedene Forschungsfelder relevant, die sich mit der Zugänglichkeit von Wissen, digitalen Publikationen, Kategorisierungspraktiken und Metadaten beschäftigen. Zentraler Bezugspunkt sind dabei die digitalen Geisteswissenschaften, hier insbesondere Forschung zur di-

gitalen Modellierung und Verfügbarmachung von textbasierten Ressourcen (Flanders und Jannidis 2018; Blaney et al. 2021). In diesem Zusammenhang beschreibt Murtha Baca wie elementar Metadaten für die Gestaltung digitaler Ressourcen sind (Baca 2016), was ebenfalls an die anwendungsbezogene Forschung im Bereich des Forschungsdatenmanagements anschließt (Dierkes 2021; Iglezakis et al. 2021). Um Metadaten im Sinne der FAIR-Prinzipien interoperabel zu gestalten und mit anderen Daten verlinkbar zu machen, bietet sich ihre Bereitstellung in Linked Open Data an (Schmidt, Thiery und Trognitz 2022). Diesbezüglich hat sich das *Simple Knowledge Organization System* (SKOS) als ein Standard etabliert, um Thesauri maschinenlesbar zu formalisieren (Zaytseva 2020).

Die methodische Umsetzung des Dissertationsprojekts erfolgt in drei Schritten. Für die Modellierung der Themen- und Begriffsfelder des Thesaurus wurden zunächst neun Kategorien identifiziert, die für filmhistoriografische Untersuchungen besonders relevant sind: Personen, Werke, Tätigkeiten, Orte, Techniken, Institutionen, Filmgenre, Themen und Publikationen. Im Zuge einer manuellen Textanalyse und -annotation mit CATMA werden die knapp 300 Profile mittels der neun Kategorien verschlagwortet. Dabei werden die gewählten Kategorien evaluiert, jeweils um Begriffsfelder ergänzt und in Form eines Thesaurus organisiert. Im nächsten Schritt erfolgt die Formalisierung des Thesaurus anhand von SKOS. Abschließend wird der Thesaurus exemplarisch an einer Auswahl von Profilen erprobt und zusätzlich um Annotationsrichtlinien ergänzt. Im Rahmen von Workshops mit Autor*innen, Nutzer*innen und zentralen Akteur*innen des WFPP werden der Thesaurus und die Annotationsrichtlinien evaluiert und Empfehlungen für die Weiterentwicklung formuliert.

Die Arbeit mit SKOS wird von einer kritischen Reflexion darüber begleitet, welche Form der Datenmodellierung durch die Anwendung dieses Standards ermöglicht und was dadurch verhindert wird (Flanders und Jannidis 2018, 11). Dabei geht es sowohl um eine Auseinandersetzung mit den epistemischen Bedingungen von SKOS (Drucker 2011), als auch um ein Nachdenken über die spezifischen Anforderungen feministischer Filmgeschichtsschreibung an die Gestaltung von Kategorien.² Wie kann es gelingen, spezifisch feministische Werte in die Gestaltung von Kategorien einfließen zu lassen, wenn wir davon ausgehen, dass es sich dabei nicht um neutrale Praktiken handelt, sondern um Prozesse, in die politische Haltungen und Werte eingeschrieben sind (Bowker und Star 1999; Drabinski 2013; D'Ignazio und Klein 2020)? Im Zusammenhang mit feministischer Filmgeschichtsschreibung geht es konkret auch um die Frage, wie die zahlreichen Leerstellen und Wissenslücken, die es aufgrund verloren gegangener Quellen und der fehlenden Dokumentation „feminisierter Arbeit“ (Hill 2016, 5) in Bezug auf die frühe Filmindustrie gibt, in daten-basierten Ansätzen dargestellt werden können (Dang 2020).

Der Vortrag präsentiert eine erste Version des Thesaurus und diskutiert zentrale Erkenntnisse aus der Annotation der Sammlung im Hinblick auf die Frage, inwiefern durch die formale Beschreibung der Profile die Struktur der Wissensproduktion im *Women Film Pioneers Project* sichtbar gemacht und somit auch daraufhin

untersucht werden kann, wie das WFPP feministische Filmgeschichte schreibt.

Fußnoten

1. Bei den Profilen werden Schlagwörter hinsichtlich der Tätigkeiten der Pionierinnen sowie der Länder, in denen sie tätig waren, vergeben.
2. In ihrem CfP für die Zeitschrift *Digital Humanities Quarterly* (DHQ), weisen Dominik Gerstorfer, Evelyn Gius und Janina Jacke darauf hin, dass die systematische Reflexion über Kategoriensysteme bisher eine Leerstelle in den Digital Humanities darstellt.

Bibliographie

- Baca, Murtha. 2016. „Introduction to Metadata.“ Los Angeles: Getty Research Institute. <http://www.getty.edu/publications/intrometadata> (zugegriffen am 27.07.2022).
- Blaney, Jonathan, Jane Winters, Sarah Milligan und Martin Steer. 2021. *Doing digital history*. Manchester: Manchester University Press.
- Bowker, Geoffrey C. und Susan Leigh Star. 1999. *Sorting Things Out: Classification and its Consequences*. Cambridge: MIT Press.
- Dang, Sarah-Mai. 2020. „Unknowable Facts and Digital Databases: Reflections on the Women Film Pioneers Project and Women in Film History.“ *Digital Humanities Quarterly* 14 (4). <http://www.digitalhumanities.org/dhq/vol/14/4/000528/000528.html> (zugegriffen am 01.08.2022).
- D'Ignazio, Catherine und Lauren F. Klein. 2020. *Data Feminism*. Cambridge: MIT Press.
- Dierkes, Jens. 2021. „Planung, Beschreibung und Dokumentation von Forschungsdaten.“ In *Praxishandbuch Forschungsdatenmanagement*, hg. von Markus Putnings, Heike Neuroth und Janna Neumann, 303-325, Berlin/Boston: De Gruyter Saur. <https://doi.org/10.1515/9783110657807>.
- Drabinski, Emily. 2013. „Queering the Catalog: Queer Theory and the Politics of Correction.“ *The Library Quarterly* 83 (2): 94-111. <https://doi.org/10.1086/669547>.
- Drucker, Johanna. 2011. „Humanities Approaches to Graphical Display.“ *Digital Humanities Quarterly* 5 (1). <http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html> (zugegriffen am 15.07.2022).
- Flanders, Julia und Fotis Jannidis, hg. 2018. *The Shape of Data in the Digital Humanities: Modeling Texts and Text-based Resources*. London: Routledge. <https://doi.org/10.4324/9781315552941>.
- Gaines, Jane, Radha Vatsal und Monica Dall'Asta. o. J. *Women Film Pioneers Project*. New York: Columbia University Libraries. <https://wfpp.columbia.edu/> (zugegriffen am 27.07.2022).
- Hill, Erin. 2016. *Never Done: A History of Women's Work in Media Production*. New Brunswick: Rutgers University Press.
- Iglezakis, Dorothea, Marc Fuhrmans, Susanne Arndt, Évariste Demandt, Stephan Hachinger, Daniela Hausen, Giacomo Lanza, Johannes Lipp, Rainer Stotzka

und Džulia Terzijska. 2021. „Interoperabilität von Metadaten innerhalb der NFDI: Konsortienübergreifender Metadaten-Workshop am 2./3. Juli 2020.“ In *Bausteine Forschungsdatenmanagement* 2: 124–35. <https://doi.org/10.17192/bfdm.2021.2.8313>.

Schmidt, Sophie C., Florian Thiery und Martina Trognitz. 2022. „Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata.“ *Digital* 2 (3): 333–64. <https://doi.org/10.3390/digital2030019>.

Listen in historischen Zeitungen: Herausforderungen und Potenziale der digitalen Analyse einer vernachlässigten Textsorte

Rastinger, Nina C.

ninaclaudia.rastinger@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage (Österreichische Akademie der Wissenschaften) / Universität Wien

Listen begegnen uns im Alltagsleben in vielfältiger Form und ihre Erstellung und Nutzung kann als fundamentale Kulturtechnik erachtet werden (Adelmann 2021, 26). Gleichzeitig ist es ebendiese Fundamentalität von Listen, die sie für wissenschaftliche Untersuchungen lange Zeit irrelevant erscheinen hat lassen: „Listen werden zumeist unterschätzt, weil sie uns so einfach und selbstverständlich vorkommen“, halten Schaffrick und Werber (2017, 303) fest und fassen damit den Forschungsstand zusammen: Wenn auch einzelne Erscheinungsformen von Listen – nämlich jene in der Literatur (u.a. Mainberger 2003, Belknap 2004, Barton et al. 2022) und modernen digitalen Kontexten (u.a. Esposito 2017, Bubenhofer 2020) – zunehmend in den Fokus der Forschung rücken, ist die Textform von den meisten Disziplinen, wie der (historischen) Linguistik, trotz ihrer hohen Frequenz bislang nur in Ausnahmefällen thematisiert worden (z.B. Doležalová 2009, Waldspühl 2019).

Dieses ‚Übersehen‘ im doppelten Sinne betrifft insbesondere die Erforschung frühneuzeitlicher Presseprodukte: Obwohl sich in einer Vielzahl historischer Zeitungen (periodisch publizierte) Listen finden, wie die „Anzeige der hier angekommenen Personen“ in der *Münchener Zeitung*, die „Lista aller Getauften“ im *Wien[n]erischen Diarium* oder das „Verzeichnis der Verstorbenen“ in der *Preßburger Zeitung*, wurden Texte dieser Art im Gegensatz zu Nachrichtenartikeln (z.B. Pfefferkorn, Riecke und Schuster 2017) oder Inseraten (z.B. Bendel 1998) bislang weder theoretisch systematisiert noch korpusbasiert auf ihre textuellen Eigenschaften hin

untersucht – und dies obwohl sie für frühneuzeitliche Zeitungsherausgeber einen zentralen Bestandteil ihrer Produkte darstellen. So kündigt beispielsweise Johann Baptist Schönwetter die diversen Listen des *Diariums* bereits im Titelkopf von dessen erster Ausgabe als einen „besondern Anhang / Daß auch alle diejenige Personen / welche wöchentlich allhier gestorben / hingegen was von Vornehmen geböhren / dann copuliret worden / ferner anhero und von dannen verreisert / darinnen befindlich“ (WD 08.08.1703, 1) an und auch Johann Michael Landerer verspricht im Avertissement zur *Preßburger Zeitung*, sie „wird allezeit die Verstorbenen richtig anzeigen“ (PZ 14.07.1764, 5). Dass Listen in historischen Zeitungen derart prominente Erwähnung finden, deutet auf ihren besonderen Wert für das damalige Lesepublikum hin – und damit auf ihre hohe Forschungsrelevanz. Hinzu kommt, dass gerade sogenannte ‚kleine Texte‘ sich oftmals als semiotisch hochkomplex erweisen und von „kommunikativer (formaler wie inhaltlicher) Prägnanz“ gekennzeichnet sind (Klug 2021, 219).

Mit ebendiesem Potenzial von Listen (und listenartigen¹ Texten) setzt sich die Dissertation auseinander, indem sie diese innerhalb von Zeitungen zwischen 1600 und 1850 empirisch verfolgt. Praktisch umgesetzt wird dies über Methoden einer multimodalen Korpuslinguistik sowie der Digital Humanities: Indem etwa in bereits bestehenden digitalen Zeitungskorpora (z.B. ANNO, DiFMOE, DIGITARIUM, DTA, impresso, Teßmann digital) nach spezifischen sprachlichen Ausdrücken im Volltext (z.B. „List(ela)“, „Vert?z(ale)ichni(s)ß“) sowie strukturellen Annotationen (z.B. TEI-Elemente <list>, <item>) gesucht werden kann, lässt sich über ein effizientes Distant Reading eine erste ‚Liste von Listen‘ erstellen, welche die Zeitungstextsorte für diverse Disziplinen überschaubar und damit nutzbar machen soll.

Überdies wird das identifizierte Material – im Sinne eines Close bzw. Scalable Readings (Mueller 2020) – mithilfe digitaler Textanalyse-Tools, wie Voyant Tools oder CATMA, sowohl zeitungsspezifisch als auch -übergreifend auf rekurrente textuelle Muster verschiedener Ebenen befragt:

1. Listentypus/-paradigma (z.B. Sterbeliste, Ankunftsliste, Preisliste, Inventar)
2. Selektionsprinzipien (z.B. Exklusion spezifischer Stände, Inklusion ‚leerer‘ Items)
3. Ordnungsprinzipien (z.B. chronologisch, hierarchisch, alphabetisch, geographisch)
4. Einsatz typographischer Ressourcen (z.B. Einrückungen, Aufzählungszeichen, Zwischenüberschriften)
5. Sprachliche Muster (z.B. zunehmende Abkürzungsdichte, Parallelen/Unterschiede zwischen selbem Listentypus in verschiedenen Zeitungen)
6. Pragmatische Textfunktion (z.B. Dokumentation, Handlungsaufforderung)

Zudem sollen durch Case Studies zu ausgewählten Textzeugen insbesondere die spezifischen Herausforderungen und Potenziale der historischen Zeitungstextsorte ‚Liste‘ für die Digital Humanities herausgearbeitet werden. Einen ersten Schritt in diese Richtung stellt das von der Stadt Wien geförderte Projekt „Zu Gast in Wien – digitale Ansätze zur (semi-)automatischen Auswertung der Ankunftslisten des *Wien[n]erischen Diariums*“ (PI: Nina C. Rastinger, 2022–2023) dar. Hierin wer-

den für die zwischen 1703 und 1725 zweimal wöchentlich in der historischen *Wiener Zeitung* erschienenen Listen zur „Ankunfft derer Hoch- und niederen Stands=Personen“ zuerst über Transkribus hochqualitative Volltexte erstellt und diese dann auf sprachliche Muster hin analysiert, welche wiederum als Basis für eine (semi-)automatische Named Entity Recognition und Visualisierung der Daten auf historischen Stadtplänen Wiens dienen können:

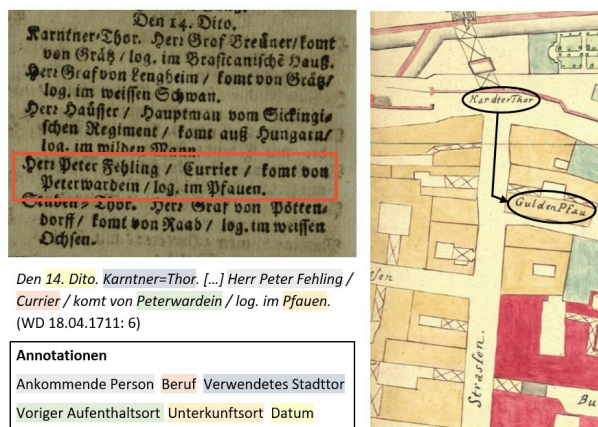


Abbildung 1: Faksimile, annotierter Volltext und Karten-Darstellung eines exemplarischen Ankunftslisteneintrags des Wien[er]ischen Diariums

Erste Ergebnisse dieser Fallstudie demonstrieren, dass Listen in frühneuzeitlichen Zeitungen aufgrund ihrer makrotypographischen Gestaltung (u.a. vermehrte Zwischenüberschriften, Einrückungen) zwar eine Herausforderung für automatische Layoutanalysen bilden, ihnen durch ihre starke Strukturiertheit, ihren diachron konsistenten Aufbau und ihre hohe semantische Dichte aber ein besonderes Potenzial für (semi-)automatische Informationsextraktionsprozesse innewohnt – wodurch ihre Volltextdigitalisierung und korpuslinguistische Untersuchung letztlich auch auf dieser Ebene einen hohen Erkenntnisgewinn verspricht.

Fußnoten

1. Man denke etwa an verwandte Formen wie Tabellen, die es erst korpusbasiert abzugrenzen gilt.

Bibliographie

- Adelmann, Ralf. 2021. *Listen und Rankings. Über Taxonomien des Populären*. Bielefeld: transcript.
- Barton, Roman Alexander, Julia Böckling, Sarah Link und Anne Runggemeier, Hrsg. 2022. *Forms of List-Making: Epistemic, Literary, and Visual Enumeration*. Cham: Springer Nature.
- Belknap, Robert E. 2004. *The List: The Uses and Pleasures of Cataloguing*. New Haven: Yale University Press.

Bendel, Sylvia. 1998. *Werbeanzeigen von 1622–1798. Entstehung und Entwicklung einer Textsorte*. Tübingen: Niemeyer.

Berlin-Brandenburgische Akademie der Wissenschaften, Hrsg. *Deutsches Textarchiv (DTA)*. <https://www.deutschestextarchiv.de> (zugegriffen: 01.08.2022).

Bubenhöfer, Noah. 2020. *Visuelle Linguistik: Zur Genese, Funktion und Kategorisierung von Diagrammen in der Sprachwissenschaft*. Berlin, Boston: De Gruyter.

Digitales Forum Mittel- und Osteuropa e.V., Hrsg. *Digitales Forum Mittel- und Osteuropa (DiFMOE)*. <https://www.difmoe.eu> (zugegriffen: 01.08.2022).

Doležalová, Lucie, Hrsg. 2009. *The charm of a list: From the Sumerians to computerised data processing*. Newcastle upon Tyne: Cambridge Scholars.

Esposito, Elena. 2017. "Organizing without Understanding. Lists in Ancient and in Digital Cultures." *LiLi, Zeitschrift für Literaturwissenschaft und Linguistik* 47(3): 351–359.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh und Jan Horstmann, Hrsg. *CATMA 6 (Version 6.5)*. <https://app.catma.de> (zugegriffen: 01.08.2022).

Haß-Zumkehr, Ulrike. 1998. "Wie glaubwürdige Nachrichten versichert haben": *Formulierungstraditionen in Zeitungsnachrichten des 17. bis 20. Jahrhunderts*. Tübingen: Narr.

Impresso . *Media Monitoring of the Past*. Unterstützt vom Schweizerischen Nationalfonds (CR-SII5_173719). <https://impresso-project.ch/app> (zugegriffen: 01.08.2022).

Klug, Nina-Maria. 2021. "Kleine Texte des Alltags: Was uns z.B. Zigarettenschachteln alles sagen und zeigen können." In *Kleine Texte*, hg. von Steffen Pappert und Kersten Sven Roth, 191–225. Berlin: Peter Lang.

Landesbibliothek Dr. Friedrich Teßmann, Hrsg. *Teßmann digital*. <https://digital.tessmann.it> (zugegriffen: 01.08.2022).

Mainberger, Sabine (2003): *Die Kunst des Aufzählens: Elemente zu einer Poetik des Enumerativen*. Berlin: De Gruyter.

Mueller, Martin. 2020. Scalable Reading. <https://scalablereading.northwestern.edu/> (zugegriffen: 01.08.2022).

Österreichische Nationalbibliothek, Hrsg. *AustriaN Newspapers Online (ANNO)*. <https://anno.onb.ac.at> (zugegriffen: 01.08.2022).

Pfefferkorn, Oliver, Jörg Riecke und Britt-Marie Schuster, Hrsg. 2017. *Die Zeitung als Medium in der neueren Sprachgeschichte. Korpora – Analyse – Wirkung*. Berlin, Boston: De Gruyter.

READ-COOP, Hrsg. *Transkribus*. <https://readcoop.eu/de/transkribus> (zugegriffen: 01.08.2022).

Resch, Claudia und Dario Kampkaspar, Hrsg. *Wienerisches DIGITARIUM*. <https://digitarium.acdh.oeaw.ac.at> (zugegriffen: 01.08.2022).

Sinclair, Stéfan und Geoffrey Rockwell, Hrsg. *Voyant Tools*. <https://voyant-tools.org> (zugegriffen: 01.08.2022).

Waldspühl, M. (2019): Die Liste als Ordnungsmedium im mittelalterlichen Libri vitae. *LiLi, Zeitschrift für Literaturwissenschaft und Linguistik* 49(2): 197–218.

Mixed Methods in der Genozidforschung

Schirmer, Miriam

miriam.schirmer@tum.de

Technische Universität München, Universität Regensburg, Deutschland

Eine Analyse von Schilderungen traumatischer Erfahrungen in ZeugInnenaussagen vor internationalen Völkermordtribunalen

Gerichtstranskripte internationaler Völkermordtribunale gelten als verlässliche Quelle, um verschiedene Aspekte von Völkermord und der Rolle von Überlebenden im Prozess zu beleuchten. Bisher wird die Fülle an öffentlich zugänglichen Gerichtstranskripten als historische Quelle jedoch kaum genutzt; es liegen nur wenige Studien vor, die entsprechende Transkripte direkt einbeziehen (z. B. Mullins 2009; Perrin 2016). Zudem konzentrieren sich die vorhandenen Studien meist ausschließlich auf eine recht enge Auswahl von Gerichtsprotokollen und verwenden entweder einen qualitativen oder einen quantitativen Ansatz (Brönnimann u.a. 2013; King und Meernik 2017). Um jedoch große Mengen von Textdaten aus Gerichtstranskripten zu verarbeiten und sie systematisch zu analysieren, kann eine Kombination aus qualitativen und quantitativen (einschließlich computergestützten) Methoden zu ganzheitlicheren Ergebnissen und damit zu einer vollständigeren Erfassung des Forschungsgegenstandes führen.

Forschungsfrage

Ziel des Dissertationsprojekts ist es daher herauszufinden, wie Mixed Methods zur Analyse großer Mengen von Transkripten von Völkermord-Tribunalen beitragen können. Der Schwerpunkt liegt darauf, wie ZeugInnen als „key part of any trial“ (Extraordinary Chambers in the Courts of Cambodia 2019, 1) traumatische Erfahrungen vor Gericht beschreiben. Durch ihre Aussage stellen ZeugInnen nicht nur wichtiges Beweismaterial zur Verfügung, sie erzählen auch von ihrem persönlichen Schicksal und ihrer individuellen Überlebensgeschichte. Trotz der emotionalen Herausforderungen, die solche Schilderungen mit sich bringen können, berichten Aussagende jedoch auch von positiven Aspekten, wie Dankbarkeit gegenüber der juristischen Aufarbeitung oder einer positiven Bedeutung für den persönlichen Bewältigungsprozess (Henry 2009, 118; Strasser u.a. 2016, 161-165).

Um sich der Rolle von ZeugInnen internationaler Völkermordtribunale möglichst vielseitig anzunähern, wird im Rahmen der Dissertation eine Bandbreite verschiedener Methoden des Natural Language Processing (NLP) angewendet. Der Einbezug qualitativer Methoden durch einen Mixed-Methods-Ansatz stellt weiterhin sicher, dass der Kontext der ZeugInnenenaussagen berücksichtigt wird (Creswell und Plano Clark 2007).

Struktur und Inhalt

Als kumulatives Projekt gliedert sich die Dissertation in mehrere Einzelarbeiten, die sich mit unterschiedlichen Aspekten von ZeugInnenenaussagen befassen und verschiedene methodische Ansätze verfolgen. Zunächst

wurde in einem bereits erschienenen Paper als wichtige Grundlage das *Genocide Transcript Corpus* (GTC) erstellt, das Textdaten der drei größten ad-hoc Völkermordtribunale (Rote-Khmer-Tribunal, Internationaler Strafgerichtshof für das ehemalige Jugoslawien, Internationaler Strafgerichtshof für Ruanda) umfasst (Schirmer, Kruschwitz, Donabauer 2022). Ein zweites Paper (aktuell im Reviewverfahren) gibt in einem dreistufigen Mixed-Methods-Design mit NLP-Klassifikation, Sentiment-Analysen und qualitativer Inhaltsanalyse einen Überblick über verschiedene Methoden zur Analyse traumatischer Inhalte in ZeugInnenenaussagen. Auf diesen beiden Papern aufbauend befasst sich der dritte Teil der Dissertation mit dem Training eines NLP-Algorithmus, der Textabschnitte mit Berichten traumatischer Erfahrungen automatisiert erkennt. Mit Hilfe eines qualitativen Ansatzes soll zudem herausgefunden werden, wie sich diese Passagen von anderen ZeugInnenenaussagen unterscheiden. Im vierten Teil der Arbeit wird Topic Modeling (Blei, Ng, Jordan 2003) angewendet, um herauszufinden, welche thematischen Muster in einzelnen ZeugInnenenaussagen zu finden sind. Dabei wird kritisch auf die Interpretierbarkeit von Topic Models eingegangen und es werden Möglichkeiten aufgezeigt, wie durch auf Topic Modeling aufbauende statistische Analysen neue Erkenntnisse gewonnen werden können. Die gefundenen Topics werden qualitativ überprüft. In einem abschließenden Paper werden die Ergebnisse zusammengeführt und auf verschiedene Völkermordtribunale angewandt, um Unterschiede zu berücksichtigen und durch einen breiteren Blickwinkel allgemeinere Schlussfolgerungen zu ermöglichen.

Die skizzierten Paper zeigen die Bandbreite des Anwendungsbereichs von Mixed Methods in der Völkermordforschung auf und legen den Grundstein für zukünftige Studien dieser Art. Indem traditionelle Methoden der historischen Quellenanalyse und automatisierte, computergestützte Prozesse aus dem NLP-Bereich kombiniert werden, sollen digitale Ansätze der Geschichts- und Völkermordforschung aufgezeigt und weitere Studien in diesem Bereich ermutigt werden.

Relevanz

Die Relevanz dieses Projekts wird anhand von drei Aspekten deutlich: Einerseits stellt die Dissertation mit ihrem Mixed-Methods-Ansatz auf methodischer Ebene einen komplett neuen Zugang in der Genozidforschung dar, der es ermöglicht, ZeugInnenenaussagen systematisch und in ihrer Fülle zu analysieren. Dabei machen es NLP-Methoden erstmals möglich, Muster in Aussagen zu entdecken, die aufgrund der Fülle des Materials sonst nicht sichtbar geworden wären. Zweitens wird dieses Projekt dazu beitragen, Gerichtsprotokolle als verlässliche Informationsquelle in der Genozidforschung weiter zu etablieren, für das Feld der *Digital History* zu öffnen und damit diesen Forschungsbereich weiterzuentwickeln. Schließlich hat die Dissertation auch gesellschaftliche Relevanz, indem die Ergebnisse der durchgeführten Studien die Situation von ZeugInnen vor Gericht umfassend beleuchten und insbesondere psychologische Herausforderungen thematisieren. Die Ergebnisse werden verschiedenen NGOs zur Verfügung gestellt, wobei gemeinsam erarbeitet werden soll, wie die Erkenntnisse in die Öffentlichkeits- und Bildungsarbeit mit einbezogen werden können (Kooperation mit dem Auschwitz

Institute for the Prevention of Genocide and Mass Atrocities und Genocide Alert e.V.).

Die Finalisierung der Dissertation wird im Frühjahr 2024 angestrebt.

Bibliographie

Blei, David M., Andrew Y. Ng und Michael I. Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3: 993-1022.

Brönnimann, Rebecca, Jane Herlihy, Julia Müller und Ulrike Ehlert. 2013. "Do testimonies of traumatic events differ depending on the interviewer?" *The European Journal of Psychology Applied to Legal Context*, 5.1: 97-121.

Creswell, John W. und Vicki L. Plano Clark. 2007. *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.

Extraordinary Chambers of the Courts of Cambodia. 2019. *ECCC. At a Glance*. Extraordinary Chambers in the Courts of Cambodia.

Henry, Nicola. 2009. "Witness to Rape: The Limits and Potential of International War Crimes Trials for Victims of Wartime Sexual Violence." *International Journal of Transitional Justice*, 3.1: 114-134.

King, Kimi Lynn und James David Meernik. 2017. *The Witness Experience: Testimony at the ICTY and its Impact*. New York, NY: Cambridge University Press.

Mullins, Christopher W. 2009. "He Would Kill Me With His Penis: Genocidal Rape in Rwanda as a State Crime." *Critical Criminology*, 17.1: 15-33.

Perrin, Kristen. 2016. "Memory at the International Criminal Tribunal for the Former Yugoslavia (ICTY): Discussions on Remembering and Forgetting Within Victim Testimonies." *East European Politics and Societies*, 30.2: 270-287.

Schirmer, Miriam, Udo Kruschwitz und Gregor Donabauer. 2022. "A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts." In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 4504-4512.

Strasser, Judith, Julian Poluda, Chhim Sothea und Phuong Pham. 2016. "Justice and Healing at the Khmer Rouge Tribunal: The Psychological Impact of Civil Party Participation." In *Cambodia's hidden scars: Trauma psychology and the Extraordinary Chambers in the Courts of Cambodia*, hg. von Beth Van Schaack, Daryn Reicherter, and Youk Chhang, 190-212. Phnom Penh: Documentation Center of Cambodia.

Stilometrie in der Diplomatie: Ein neues Forschungsfeld?

Geißel, Pia

geissel@uni-wuppertal.de

Bergische Universität Wuppertal, Deutschland

Seit rund 20 Jahren steigt die online verfügbare Menge an digitalen Daten in unterschiedlichster Form und Qualität. Dies erleichtert den Geisteswissenschaftlerinnen den Zugang zu neuen Materialien und senkt die Hemmschwelle sich mit Big Data zu beschäftigen und der „methodological monoculture of close reading“ (Karsdorp et al. 2021, *preface*) zu entkommen. Dadurch müssen sie sich aber auch vermehrt mit den technischen, sprich, computergestützten Anwendungen und Programmiersprachen beschäftigen. Zudem müssen Daten meist noch gesammelt, von überflüssiger Information bereinigt und Ergebnisse visuell aufbereitet werden. Alle diese neuen Schritte sind je nach Datengrundlage aufwändig, da Techniken erst erlernt und korrekt angewendet werden müssen. Als zusätzlicher Faktor werden für Berechnungen großer Datenmengen fachfremde Methoden aus der Mathematik oder Informatik entlehnt. Die Anwendung dieser arithmetischen Methoden werden jedoch oftmals nicht ausreichend hinterfragt. Zusätzlich fehlt zur Überprüfung der Methoden auch die empirische Evidenz vor allem dann, wenn das Untersuchungsmaterial keine faktischen Hinweise liefern kann.

Auch in der Geschichtswissenschaft wenden sich, angeregt durch niedrigschwelligen Zugang zu online verfügbaren Texten, Geisteswissenschaftler:innen neuen Forschungsfragen zu. So eröffnet beispielsweise die Stilometrie neue Möglichkeiten, die Anonymität eines mittelalterlichen Textzeugens aufzuheben und neue Thesen bezüglich Überlieferung und Organisation in Schreibstuben und Kanzleien aufzustellen. Dabei bewegen sich jedoch erstens die Forschungsfragen häufig um eine konkrete Identifizierung eines Stiles, einer Kanzlei oder einer Schreibschule und weniger über Makrosignale wie übergeordnete geographische, kulturelle oder sprachliche Dimensionen. Zweitens vernachlässigt die Arbeit am konkreten Korpus auch die Auseinandersetzung mit der Auswahl der geeigneten mathematischen Verfahren, die hinter den computergestützten Berechnungen stehen. Eine Auseinandersetzung auf der Makroebene über die Effektivität von beispielsweise einem Distanzmaß-Verfahren wie Burrows's Delta auf die konkrete Textgattung der Urkunden findet bis heute noch nicht statt. Zwar gibt es Messungen über den Wirkungsgrad des Verfahrens für Lyrik und Prosa in lateinischer Sprache und kurzer oder auch fehlerhafter Texte¹, dennoch finden darin die spezifischen Eigenschaften von Urkunden nicht ausreichend Berücksichtigung: Individuelle, orthographische Signale, formelhafte Sprache und lückenhafte Überlieferungen sind nur einige der spezifischen Faktoren, welche noch nicht ausreichend für Burrows's Delta untersucht wurden (vgl. Eder 2013, 2015).

Betrachtet man das weitere Forschungsfeld zum Thema Autorschaftsattributions werden aktuell vermehrt andere statistische Ansätze in Erwägung gezogen, die nicht zwingend auf Textfeatures wie Wortlängen oder inhaltlichen Merkmale wie Worthäufigkeiten basieren. Eine Fokussierung auf syntaktische oder semantische Zusammenhänge beispielsweise könnte bei der Untersuchung der Signale in Urkunden ebenso Distinktionen herausheben. Eine Loslösung vom Vector-Space-Model und Burrows Delta hin zu Topics und Neuronalen Netzen wurde zwar bisher auf dem Typus Urkunden noch nicht im großen Rahmen angewandt, dennoch könnte man dadurch potentiell das Manko der gerin-

gen Textlängen und der formelhaften Sprache umgehen. Neuere Ansätze haben sich zudem das Ziel gesetzt, durch eine Kombination mehrerer methodischer und textimmanenter Ansätze Merkmale herauszuarbeiten. Diese umschließen dann folglich nicht mehr nur die klassische Stilometrie, sondern auch die oben genannten syntaktischen und semantischen Features. Ob diese neuen Ansätze bei Urkunden Wirkung zeigen, soll ein Ziel dieser Dissertation werden.

Ein grundsätzliches Problem bezüglich der Urkundentexte ist allerdings, dass diese häufig nicht von ihren originalen Textzeugen, sondern nur aus digitalisierten Editionen entnommen sind, die nicht dem eigentlichen Überlieferungstext entsprechen müssen. Wie sehr vertraut man Texten aus älteren Editionen vor 1945, in denen die Lachmannsche Editionstechnik angewendet wurde, mit der die *emendatio* angeblicher Fremdeingriffe unbedarft angewendet und zudem schlecht oder gänzlich undokumentiert in den Druck gegeben wurde (vgl. Plachta 2006, S. 29)?

Aus diesen Überlegungen heraus ist es naheliegend, die stilometrischen Verfahren einmal aus der Makroperspektive zu untersuchen: Anstatt sich mit einer These zu beschäftigen, die aus dem konkreten Material, vielleicht sogar aus dem close reading-Prozess selbst, entstanden ist, sollten die mathematischen Methoden an einer großen und diversen Menge an Urkundenmaterial untersucht werden. Eine mannigfaltige Auswahl bieten dazu mehrere Urkundenportale, wie *monasterium.net*, *Cartae Europae Medii Aevi*, *Codice diplomatico della Lombardia medievale* oder *Telma*.² Dabei spielen zunächst weder die Urkundentypen, noch die zeitliche Dimension, in der die Urkunden entstanden sind, eine Rolle. Die Annahme ist eher, dass verschiedene Verfahren unterschiedlich starke Distinktionen der Urkunden unterstreichen. Ihre jeweilige Sensibilität für bestimmte Eigenschaften des Textmaterials gilt es herauszuarbeiten und methodisch zu begründen. So lassen sich Stärken und Schwächen der Methoden analysieren und gegenüberstellen, nicht im Sinne eines Rankings („bessere vs schlechtere Methode“), sondern im Hinblick auf ihre Eignung für spezifische Korpora und Fragestellungen. Ein solcher Ansatz erlaubt es Forschenden nicht nur, eine fundiertere Entscheidung über die anzuwendende Methode zu treffen, sondern im Idealfall sogar, diese gezielt auf den eigenen Anwendungsfall abzustimmen und zu verfeinern.

Fußnoten

1. Stichwort fehlerhaftes OCR.
2. <https://www.monasterium.net/mom/home>, <http://telma-chartes.irht.cnrs.fr/>, <https://cema.lamop.fr/>, <https://lombardiabeniculturali.it/cdlm/edizioni/>.

Bibliographie

Argamon, Shlomo. 2008. „Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations.“ In *Literary and Linguistic Computing*, 23,2. S. 131-147. Oxford: University Press.

Burrows, John. 2002. „Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship.“ In *Literary and Linguistic Computing*, 17,3. S. 267-287. Oxford: University Press.

Burrows, John. 2003. „Questions of authorship: attribution and beyond: a lecture delivered on the occasion of the Roberto Busa Award ACH-ALLC 2001, New York.“ In *Computers and the Humanities*, 37,1. S. 5-32. Dordrecht: Kluwer.

Eder, Maciej. 2012. „Computational stylistics and Biblical translation: How reliable can a dendrogram be?“ In *The Translator and the Computer*, hg. von Piotrowski/Grabowski, S. 155-170. Wrocław: Verlag der Hochschule für Philologie in Wrocław.

Eder, Maciej. 2013. „Mind your Corpus: Systematic errors in authorship attribution.“ In *Literary and Linguistic Computing*, 28,4. S. 603-614. Oxford: University Press.

Eder, Maciej. 2015. „Does size matter? Authorship attribution, small samples, big problem.“ In *Digital Scholarship in the Humanities*, 30,2. S. 167-182. Oxford: Oxford Academic.

Eder, Maciej und Jan Rybicki. 2015. „Go Set A Watchman while we Kill the Mockingbird In Cold Blood.“ <https://computationalstylistics.github.io/blog/harper-lee/> (zugeschrieben: 03. August 2022).

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch und Thorsten Vitt. 2017. „Understanding and explaining Delta measures for authorship attribution.“ In *Digital Scholarship in the Humanities* 32, Suppl. 2. S. ii4- ii16 10.1093/llc/fqx023.

Jockers, Matthew und Daniela M. Witten. 2010. „A comparative study of machine learning methods for authorship attribution.“ In *Literary and Linguistic Computing*, 25,2. S. 215-23. Oxford: University Press.

Juola, Patrick und Stephen Ramsay. 2017. „Six Septembers: Mathematics for the Humanist.“ Zea E-Books Collection, 55. Lincoln: Zea Books.

Karsdorp, Folger, Mike Kestemont und Allen Riddel. 2021. „Humanities Data Analysis: Case Studies with Python.“ Princeton: Princeton University Press.

Koppel, Moshe, Jonathan Schler und Shlomo Argamon. 2009. „Computational Methods in Authorship Attribution.“ In *Journal of the American Society for Information Science and Technology*, 60,1. S. 9-26. New York: Wiley.

Koppel, Moshe und Yaron Winter. 2014. „Determining If Two Documents Are Written by the Same Author.“ In *Journal of the American Society for Information Science and Technology*, 65,1. S. 178-187. New York: Wiley.

Kou, Gang, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, und Fawaz E. Alsaadi. 2020. „Evaluation of Feature Selection Methods for Text Classification with Small Datasets Using Multiple Criteria Decision-Making Methods.“ *Applied Soft Computing* 86. 10.1016/j.asoc.2019.105836.

Modupe, Abiodun, Turgay Celik, Vukosi Marivate, und Oludayo O. Olugbara. 2022. „Post-Authorship Attribution Using Regularized Deep Neural Network.“ *Applied Sciences* 12 (15). 10.3390/app12157518.

Plachta, Bodo. 2006. *Editionswissenschaft eine Einführung in Methode und Praxis der Edition neuerer Texte*. 2. Aufl. Stuttgart: Reclam.

Plakias, Spyridon, und Efsthathios Stamatatos. 2008. „Tensor Space Models for Authorship Identification.“ In *Artificial Intelligence: Theories, Models and Applications*,

herausgegeben von John Darzentas, George A. Vouros, Spyros Vosinakis, und Argyris Arnellos, 239–49. Lecture Notes in Computer Science. Berlin: Springer.

Potha, Nektaria und Efstathios Stamatatos. 2017. „An Improved Impostors Method for Authorship Verification.“ <http://www.icsd.aegean.gr/lecturers/stamatatos/papers/CLEF-Potha-2017.pdf> (zugegriffen: 03. August 2022).

Vogeler, Georg. 2006. „Vom Nutz und Frommen digitaler Urkundeneditionen.“ In *Archiv für Diplomatik*, 52. S. 449–466. Wien: Böhlau.

Vogeler, Georg. o.J. „Die Text Encoding Initiative (TEI) als Werkzeug des Urkundeneditors – Erfahrungen und Desiderate.“ https://rep.adw-goe.de/bitstream/handle/11858/00-001S-0000-0023-9A13-A/6_Vogeler.pdf?sequence=70. (zugegriffen: 03. August 2022).

Wu, Haiyan, Zhiqiang Zhang, und Qingfeng Wu. 2021. „Exploring Syntactic and Semantic Features for Authorship Attribution“. *Applied Soft Computing* 111. 10.1016/j.asoc.2021.107815.

Theorising the Aesthetic Properties of Reading in a Digital Social Reading (DSR) Environment: Exploring DSR Practices in India

Ghosh, Sharanya

ghosh.5@iitj.ac.in

Indian Institute of Technology, Jodhpur, Indien

Digital Social Reading, a term proposed by Rebora et al. (2021), is described as "a wide variety of practices related to the activity of reading and using digital technologies and platforms... to share with other people, thoughts and impressions about texts" (Pianzola, forthcoming). This study will dive deep into the world of Indian Digital Social Reading practices, thus filling the gap in current global discourse. Aesthetic experience, central to reading as an activity, forms the basis of aesthetic judgement of a literary text. Therefore, it is essential that our understanding of how reading happens also includes some empirical analysis of its aesthetic properties - the foundational "units" of aesthetic judgement. Combining these two strands, this study aims to conceptualise a theoretical framework of aesthetic properties (AP) of reading fiction in the context of Indian Digital Social Readers by employing an embedded mixed-method approach.

Literature:

While most scholarship in DSR is global north-centric, Pianzola et al. (2020) offer a more global perspective in their study of Wattpad as a literary resource. Pianzola's upcoming book is the only textbook-like work organising the different DSR studies in a historical timeline and ta-

xonomies; DSR's role in cognitive, aesthetic, and educational aspects of reading, and its pedagogical properties. The democratising agency of DSR (Sedo 2011; Kellner 2016); statistical interpretation of data on gender biases, gender and genre fixation, authorship and gender identity etc., (Thelwall and Kousha 2017); the impact of tweaking reviews for book sales (Nan Hu and colleagues 2012) - are some of the most prominent works. Holur et al. (2021) reflected upon the manners of reading novels that readers engage in, making a significant contribution to "infinite vocabulary networks". Koolen, Neugarten, and Boot (2022) identified impact categories in English book reviews.

The Oxford Handbook of Cognitive Literary Studies (2015) discusses intersections between the literary and cognitive, reflecting on imagination, brain imaging, and its neural networks; imagery as the key element of aesthetic experience; the delicate balance between internal and external cognition giving rise to an intense aesthetic response etc. The Oxford Handbook of Aesthetics (2005) is a comprehensive work on philosophical ideas of aesthetics and its properties; arts and their relationship with aesthetics; notions of interpretation etc. In a self-reflexive essay, Matravers and Levinson (2005) defend Levinson's ontology of aesthetic properties while also questioning issues like aesthetic autonomy. More empirical works, such as by Larson et al. (2007), measured the aesthetics of reading based on typographic alterations and textual optimisation that affected the reader's cognitive functions and frowning while reading. DeClerq (2002) observed that existing definitions of aesthetic properties pertain mostly to visual objects, leaving enough scope for inquiry into the other forms of aesthetic expressions, such as music and literature.

Hypotheses and research questions:

Based on the hypotheses that digital space and digital technologies redefine the social reading experience and that the reading experience as a whole can be best understood through an analysis of the aesthetic properties of reading, this project attempts to answer the following questions:

1. What parameters determine the aesthetic properties that govern a reader's aesthetic judgement of fiction? (the underlying assumption here is that aesthetic properties differ for fiction and non-fiction)
 1. How are these parameters of AP affected by the affordances created by the digital?
2. Is it possible to formulate a framework of aesthetic properties of reading fiction from DSR user data?

Parameters of AP can be explained better through their dimensions, categorised according to their epistemological qualities, such as cultural, affective, semantic, and communal.

Methodology:

Owing to the complex and abstract nature of the concepts, I propose an embedded mixed methodology. As the main crux of the study, subjective data will be gathered through semi-structured interviews of Indian digital social readers. Codes generated from the transcripts (using NVivo) will help identify the parameters of AP, which will then be used in the supplementary survey involving bilingual adult Indian readers. This data should help reinforce the proposed framework. Cluster sampling will

be used for this purpose. R will be used for data collation and final analysis. The project does not concentrate on a fixed corpus given the generic nature of the questions it asks and India's multilingual readership. Data collection begins in mid-January 2023, following the completion of basic theorising.

Limitations:

- Does not consider non-fiction works
- survey method is not the most reliable of all quantitative methods
- difficulty justifying abstract concepts through empirical approach.

Bibliographie

- Allington, Daniel . 2016. 'Power to the reader' or 'degradation of literary taste'? Professional critics and Amazon customers as reviewers of *The Inheritance of Loss* . *Language and Literature: International Journal of Stylistics*, 25(3): 254–278. Accessed April 17, 2022, <https://doi.org/10.1177/0963947016652789> .
- Bourrier, Karen, and Thelwall, Mike . 2020. The Social Lives of Books: Reading Victorian Literature on Goodreads. *Journal of Cultural Analytics*, 5(1). Accessed April 10, 2022, <https://doi.org/10.22148/001c.12049> .
- Creswell, John W., and Clark, Vicky, L. Plano. 2010. Choosing a mixed method design. In *Designing and conducting mixed methods research*: 58–88. SAGE. Accessed October 21, 2022, https://www.sagepub.com/sites/default/files/upm-binaries/10982_Chapter_4.pdf .
- De Clercq, Rafael . 2002. The Concept of an Aesthetic Property. *The Journal of Aesthetics and Art Criticism*, 60(2): 167–76. Accessed November 10, 2022, <http://www.jstor.org/stable/1520014> .
- De Clercq , Rafael . 2008. The Structure of Aesthetic Properties. *Philosophy Compass* , 3 (5): 894–909. Accessed November 10, 2022, <https://doi.org/10.1111/j.1747-9991.2008.00165.x> .
- Driscoll, Beth, and DeNel, Sedo R. 2018. Faraway, So Close: Seeing the Intimacy in Goodreads Reviews. *Qualitative Inquiry* , 25(3): 248–259. Accessed April 17, 2022, <https://doi.org/10.1177/1077800418801375> .
- Hajibayova, Lala . 2019. Investigation of Goodreads' reviews: Kakutanied, deceived or simply honest?. *Journal of Documentation* , 75(3): 612–626. Accessed April 10, 2022, <https://doi.org/10.1108/jd-07-2018-0104> .
- Holur, Pavan, Shahsavari, Shadi, Ebrahimzadeh, Ehsan, Tangherlini, Timothy R., and Roychowdhury, Vwani . 2021. Modeling Social Readers: Novel Tools for Addressing Reception from Online Book Reviews. London, UK; Royal Society Open Science. Accessed April 10, 2022, <https://arxiv.org/pdf/2105.01150.pdf> .
- Hu, Nan, Bose, Indranil, Koh, Noi Koh S., and Liu, Ling . 2012. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems* , 52(3): 674–684. Accessed May 5, 2022, <https://doi.org/10.1016/j.dss.2011.11.002> .
- Hu, Nan, Pavlou, Paul, A., and Zhang, Jie . 2017. On Self-Selection Biases in Online Product Reviews. *MIS Quarterly* , 41(2), 449–471. Accessed June 12, 2022, <https://doi.org/10.25300/MISQ/2017/41.2.06> .
- Kellner, Douglas . 2001. Critical Pedagogy, Cultural Studies, and Radical Democracy at the Turn of the Millennium: Reflections on the Work of Henry Giroux. *Cultural Studies* ↔
- Critical Methodologies , 1(2): 220–239. Accessed April 10, 2022, <https://doi.org/10.1177/153270860100100205> .
- Koolen, Marijn, Neugarten, Julia, and Boot, Peter . 2022. 'This book makes me happy and sad and I love it': A Rule-based Model for Extracting Reading Impact from English Book Reviews. In *Conference Reader of the 1st Annual Conference of Computational Literary Studies* . Accessed May 5, 2022, https://jcls.io/media/journals/12/CCLS2022_Conference-Reader_2022-05-16v2.pdf .
- Kousha, Kayvan, Thelwall, Mike, and Abdoli, Mahshid . 2017. Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology* , 68(8). Accessed June 12, 2022, <https://doi.org/10.1002/asi.23805> .
- Larson, Kevin, Hazlett, Richard L., Chaparro, Barbara, S., and Picard, Rosalind, W . Measuring the aesthetics of reading. *People and Computers XX – Engage* . 41–56. Accessed June 20, 2022, https://doi.org/10.1007/978-1-84628-664-3_4 .
- Levinson, Jerrold . 2005. The Domain of Aesthetics. *The Oxford Handbook of Aesthetics* . England: Oxford University Press.
- Matravers, Derek, and Levinson, Jerrold . 2005. Aesthetic Properties. *Proceedings of the Aristotelian Society, Supplementary Volumes* , 79: 191–227. Accessed April 17, 2022, <http://www.jstor.org/stable/4106940> .
- Pianzola, Federico . 2021. Digital Social Reading: Sharing Fiction in the 21st Century (Forthcoming). MIT Press. Accessed April 10, 2022, <https://wip.mitpress.mit.edu/digital-social-reading> .
- Pianzola, Federico, Rebora, Simone, and 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 margin. *PLOS One* , Accessed April 10, 2022, <https://doi.org/10.1371/journal.pone.0226708> .
- Rebora, Simone, Boot, Peter, Pianzola, Federico, Gasser, Brigitte, Herrmann, J Berenike, Kraxenberger, Maria, Kuijpers, Moniek M, Lauer, Gerhard, Lendvai, Piroška, Messerli, Thomas C, and Sorrentino, Pasqualina . 2021. Digital humanities and digital social reading. *Digital Scholarship in the Humanities* , 36(2). Accessed April 17, 2022, <https://doi.org/10.1093/llc/fqab020> .
- Robson, Jon . 2018. Is Perception the Canonical Route to Aesthetic Judgment?. *Australasian Journal of Philosophy* , 96 (4): 1–12. Accessed May 5, 2022, <https://doi.org/10.1080/00048402.2017.1389964> .
- Rowberry, Simon, P . 2016. Commonplacing the public domain: Reading the classics socially on the Kindle. *Language and Literature: International Journal of Stylistics* , 25(3): 211–225. Accessed May 5, 2022, <https://doi.org/10.1177/0963947016652782> .
- Swann, Joan, and Allington, Daniel . 2009. Reading groups and the language of Literary texts: A case study in social reading. *Language and Literature: International Journal of Stylistics* , 18(3): 247–264. Accessed June 12, 2022, <https://doi.org/10.1177/096394700910585> .

Thelwall, Mike, and Kousha, Kayvan . 2016. Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology* , 68(4): 972-983. Accessed April 17, 2022, <https://doi.org/10.1002/asi.23733> .

Thelwall, Mike, and Bourrier, Karen . 2019. The reading background of Goodreads book club members: a female fiction canon? *Journal of Documentation* , 75(5): 1139-1161. Accessed May 5, 2022, <https://doi.org/10.1108/jd-10-2018-0172> .

Zunshine, Lisa . 2015. *The Oxford Handbook of Cognitive Literary Studies* . England: Oxford University Press.

Von Wissensdingen und Werkräumen. Graph-basierte Modellierung von Denk- und Arbeitsspuren in Nachlässen

Stahn, Lena-Luise

lstahn@uni-wuppertal.de

Bergische Universität Wuppertal, Deutschland

Fragestellung und theoretischer Rahmen

Die aus der Medientheorie stammende Vorstellung von den überlieferten Dokumenten, den Artefakten als "Wissensdingen",¹ "denen zeitgenössisch zugeschrieben wurde, ein inhärentes Wissen über ihre eigene Natur zu enthalten sowie dieses erschließbar und vermittelbar machen zu können" (Müller 2020, 17), und des durch ihre Kontextualisierung sichtbar gemachten "Wissensraums" (Rheinberger 1992) soll auf Nachlässe und ihre Aufarbeitung in Form von Digitalen Editionen übertragen werden.² Das "artefaktische" (Sahle 2017, 239) Dokument bildet das "Wissensding", es ist Träger des Wissens, des Textes im Nachlass. Es wurde von der Autorin erschaffen oder von ihr in ihre Arbeit einbezogen. Mit jedem Objekt verband sie ein Gedanke, eine Idee. Es repräsentiert ein "Stück" ihres Denkens und Arbeitens. Der Nachlass bildet *als Kontext des Denk- und Arbeitsprozesses* den "Wissensraum", der das (Text)Werk als Ganzes verkörpert.

Um das "Autorinnen-Werk" zu erfassen und die Dynamik des Schaffensprozesses³ aufdecken zu können, entwickelt die Dissertation eine Ontologie, die die Nachlassbestandteile als typologisch, konzeptionell, inhaltlich und

strukturell relationierte Elemente beschreibt (Zangerl & Pollin 2020, 125; Spadini & Tomasi 2021, 1f.). Auf Modellerebene wird damit ermöglicht, die Arbeitspraktiken, die Prozesse des Denkens und Schreibens, den "Laborcharakter"⁴ des Werks nachvollziehbar zu machen. Dabei liegt die Frage zugrunde, inwieweit Ansätze und Methoden der Wissensmodellierung (Davis et al. 1993; Flanders & Jannidis 2015) dazu geeignet sind, den Nachlass als kontextualisierten "Wissensraum" zu öffnen, in dem sich Arbeitspraxis und Werk der Autorin begreifen lassen.

Positionierung in den DH: SDE zwischen Text und Daten

Obwohl sie mit ihrer stetigen Veränderbarkeit und Aktualität das Potential haben, als "Protokoll des Forschungsprozesses" (Sahle 2010, 27) (kursiv übernommen) zu gelten, bleibt der editorische Blick auch Digitaler Editionen (SDE) der statischen, dokument-basierten Sichtweise verhaftet (Van Zundert 2016, 83-106). Zur Erschließung von Nachlässen existieren für jeden Dokumenttyp eigene meist der bibliothekarisch-archivarischen Domäne entstammende Regeln,⁵ was wenig Raum lässt für Ansätze der Intertextualität (Broich & Schulte-Middelich 1985; Spadini & Tomasi 2021), der Idee, alles sei ein einziger "Text, der sich selbst permanent zitiert" (Neuhaus 2014, 236). Ein übergreifendes, semantisch differenzierendes Modell, in das alle Dokumente eingebettet wären, und damit eine Möglichkeit, ein adäquates Bild der Arbeitspraktiken zu schaffen, die zu ihrer Entstehung innerhalb des Werkkontexts geführt haben, fehlt.

Die "Transmedialisierung" verlangt nach Sahle 2017 von der SDE, ihren Schwerpunkt von der medialen Präsentation des Materials auf das "als Daten gefasste akkumulierte Wissen und das ihnen zugrundeliegende Modell als Explikation der editorischen Methode"⁶ zu verlagern. Diese Datafizierung (Hyvönen 2020) versucht, eine Antwort auf die relevanter werdende Frage nach der Langzeitverfügbarkeit und Nachhaltigkeit (Fritze 2019) von SDE zu geben: Angesichts der "Begrenzungen von TEI" (Sahle 2017, 247) und der Entwicklung immer neuer, lokaler, eigener Lösungen steigt die Dringlichkeit, die Daten über den Projektzeitraum hinaus "lebendig" und für menschliche und maschinelle Anwendungen nutzbar und interoperabel zu halten und damit einem Informationsverlust entgegenzuwirken (Daquino & Tomasi 2015, 1f.).

Dennoch verbleiben in der praktischen Umsetzung die projektspezifisch angepassten Modelle und (TEI/XM-L-)Daten häufig innerhalb der (geschlossenen) Projektdatenbank, eine übergreifend interpretierbare, graph-basierte Semantik fehlt, das Konzept des Knowledge Graphs (Rehbein 2017, 165) findet kaum Verwendung (Spadini & Tomasi 2021, 1). Diese Problematik zeigt sich paradigmatisch für den Bereich der Nachlasserschließung.⁷ Derzeit befinden sich mehrere Ansätze in der Entwicklung,⁸ es fehlt jedoch an Anwendungsbeispielen,⁹ während das Nebeneinander vieler neu entwickelter Modelle einen zusätzlichen Aufwand der Konsolidierung erfordert.

Praktischer Ansatz

Ziel des Projekts ist es, durch die Entwicklung einer domänenspezifischen Ontologie den Nachlass Niklas Luhmanns (1927-1998)¹⁰ in einem graph-basierten Datenmodell abzubilden (Allemang & Hendler 2011). Dazu werden existierende Modellierungsansätze aus den Bereichen Museum, Archiv und Bibliothek evaluiert, um die in den Nachlassdokumenten ermittelten Entitäten und Relationen adäquat zu beschreiben. Das im Projekt bereits genutzte FRBRer-Modell (Madison et al. 2009) wird auf seine Passfähigkeit überprüft und ergänzt bzw. durch Alternativen ersetzt. Im Fokus steht FRBRoo/LRM als eine an CIDOC-CRM¹¹ angepasste FRBR-Version,¹² daneben werden gängige Metadatenschemata wie DCTerms, SKOS, PRISM oder die SPAR Ontologies (Lüschow 2020, 82; Tomasi 2012) sowie spezifisch für bestimmte Fragestellungen entwickelte Ontologien,¹³ insbesondere außerhalb des Bibliothekskontextes,¹⁴ untersucht.

Das Vorhaben befindet sich noch in der Startphase. Ein erstes Mapping zu FRBRoo/LRM zeigt sich als prinzipiell machbar, die andauernde Entwicklung des Werkmodells im Projekt erfordert jedoch eine kontinuierliche Überprüfung des gewählten Ansatzes.

Kontextualisierung

Der Erkenntnisgewinn des Vorhabens liegt im Bereich der Datenmodelle und Ontologien für die Nachlasserschließung: Die Arbeit am Nachlass Luhmanns steht stellvertretend für wissens- und werktheoretisch basierte Untersuchungen, insbesondere bedingt durch den Zettelkasten als "Werk" und Arbeitsinstrument, dessen Spuren sich durch die weiteren Arbeiten Luhmanns ziehen und den Prozess der Wissensanreicherung nachvollziehbar machen. Der Nachlass in seinem Gesamtkontext ist damit gut geeignet, die trotz bekannter Einschränkungen (s.o.) weithin genutzten Modellierungsansätze aus den bibliothekarisch-archivarischen bzw. editionswissenschaftlichen Bereichen auf die Fragestellung nach Werkcharakter und Arbeitsprozess hin zu überprüfen. Gleichzeitig ist im Verlauf der Modellierung zu erwarten, das Verständnis von Werk und Arbeitsprozess vertiefen und Erkenntnisse auch auf wissenschaftstheoretischer Basis gewinnen zu können.¹⁵

Mit dem resultierenden, konzeptionell und technisch implementierten Knowledge Graph lässt sich das Verhältnis der Dimensionen von "Werk" auf Basis des Nachlasses abbilden. Dieser proof of concept zeigt eine Möglichkeit, der SDE eine Ebene in Form formal explizierter Information¹⁶ (Vogeler 2021, 79) zu geben, die mithilfe informationswissenschaftlicher Methoden Fragen editionswissenschaftlichen, werk-konstitutiven Charakters neu zu betrachten hilft, perspektivisch mit Auswirkungen auch auf die editorische Praxis.

Fußnoten

1. Eine erste Auslegung des Begriffs, wie er in dieser Arbeit verwendet werden soll, lehnt sich an Rheinbergers Definition des "epistemischen Dings" an (Rheinberger 1992). Wie in (Thaut 2012, 7) erläutert, verwendet er den Begriff synonym zu "Erkenntnisding" und dem in dieser Arbeit hauptsächlich genutzten "Wissensding". Müller bietet eine gute Übersicht zur Entwicklung des Begriffs (Müller 2020, 17).
2. Zur Begründung, weshalb diese Übertragung gerechtfertigt ist, vgl. Sahles Äußerung zum Geltungsbereich Digitaler Editionen: "Grundsätzlich ist aber alles edierbar, was einer kritischen Aufbereitung bedarf. Theoretisch also z. B. auch Bildwerke für die kunsthistorische Forschung oder ganz allgemein museale physische Objekte der Kulturgeschichte." (Sahle 2017, 239).
3. "Während diese genetischen Zusammenhänge in den Künstlerhäusern von Moreau und Rodin unmittelbar ins Auge springen, ist es noch heute Gegenstand der Diskussion, wie diese Dynamik im Fall der Textgenese in eine Edition zu transportieren ist – Thomas Mann hat dies anschaulich einen "Bohrungsprozeß" genannt." (Plachta 2016, 31). In diesem Zusammenhang wird im Folgenden auch von Arbeitspraktiken gesprochen, im Sinn der "Praktiken der Wissensvermittlung und -generierung", wie sie von Müller (Müller 2020, 18) definiert werden.
4. Plachta 2016, 33.
5. Während RNA und RNAB den Begriff des Werks allein als Sammelbezeichnung für bestimmte Dokumentarten und zur Abgrenzung von anderen Archivbestandteilen verwenden, formuliert FRBR ihn zwar als theoretische Entität, allerdings nur bezogen auf in Bibliotheken vorkommende und damit veröffentlichte Schriften. Vgl. <https://kalliope-verbund.info/de/standards/regelwerke.html> (abgerufen am 27.05.22) und <https://www.ifla.org/de/references/best-practice-for-national-bibliographic-agencies-in-a-digital-age/resource-description-and-standards/bibliographic-control/functional-requirements-the-frbr-family-of-models/functional-requirements-for-bibliographic-records-frbr/> (abgerufen am 27.05.22). Auch neuere Ansätze entstammen entweder ebenfalls dem bibliothekarisch-archivarischen Bereich, etwa Records in Context (vgl. <https://www.ica.org/en/records-in-contexts-conceptual-model>, abgerufen am 12.12.22), Library Reference Model (Llanes-Padrón & Pastor-Sánchez 2017, Riva et al. 2018) oder RDA (Tillet 2011), und weisen einen entsprechenden Fokus auf oder sind auf eine Domäne zugeschnitten, in der der Werk-Begriff keine Rolle spielt (hier sei insbesondere auf die inzwischen zahlreich existierenden Erweiterungen des CIDOC-CRM hingewiesen, vgl. z.B. <https://ontome.net/project>, abgerufen am 12.12.22).
6. "Man kann deshalb sagen, dass der eigentliche Wandel in der Editorik nicht so sehr in einem Wechsel der Medien liegt, als in ihrer Transmedialisierung: Bei Editionen geht es nicht nur um ihre mediale Erscheinung, sondern vor allem um das als Daten gefasste akkumulierte Wissen und das ihnen zugrundeliegende Modell als Explikation der editorischen Methode." (Sahle 2017, 241).

7. "Klassische archivarische und bibliothekarische Erschließungsmaßnahmen bilden die beste Grundlage für das Auffinden und Erforschen von Nachlassdokumenten. Sie haben aber ihre Grenzen, wenn es darum geht, die komplexen Verbindungen von Einzeldokumenten und Nachlassteilen untereinander und nach außen aufzuzeigen und das (Kontext-)Wissen der Bearbeiter*innen zu den Originalen zu formalisieren." (Zangerl & Pollin 2020, 125).
8. Insbesondere die "Records in Context"-Ontologie ist für das Vorhaben von Bedeutung, vgl. Llanes-Padrón Dunia & Pastor-Sánchez 2017, 387–405.
9. Die einzige, der Autorin derzeit bekannte, Untersuchung in diesem Zusammenhang widmet sich der Entwicklung einer domänenspezifischen Nachlass-Ontologie im Rahmen des Projekts "Stefan Zweig digital" (Zangerl & Pollin 2020). Diese Untersuchung kann dem Vorhaben aber nur bis zu dem Punkt als Orientierung dienen, ab dem eine explizite Modellierung der Domäne vorgenommen wird. Es muss differenziert werden zwischen der Domäne des Schriftsteller- und des wissenschaftlichen (Soziologen-)Nachlasses.
10. Das Langzeitprojekt "Niklas Luhmann – Theorie als Passion. Wissenschaftliche Erschließung und Edition des Nachlasses" (Laufzeit 2015–2030, gefördert durch die Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste) beschäftigt sich mit der Erschließung des wissenschaftlichen Nachlasses Luhmanns. Webpräsenz und Edition sind zu finden unter <https://niklas-luhmann-archiv.de/>.
11. Die aktuelle Version ist zu finden unter <https://cidoc-crm.org/>.
12. "The idea that both the library and museum communities might benefit from harmonising the two models" <https://cidoc-crm.org/frbroo/short-intro-frbroo>; <https://www.cidoc-crm.org/frbroo/Issue/ID-360-Irmoo>.
13. Bspw. Citation Typing Ontology (CiTO), Intertextual Relationships Ontology for literary studies (INTRO) oder CRMtex (Spadini & Tomasi 2021, 1f.). Frühe Beispiele für Editionen mit einem semantischen Ansatz sind die "Semantic Scholarly Digital Edition" des Projekts "Paolo Bufalini's Notebook" (<http://projects.dharc.unibo.it/bufalini-notebook/>), die Edition der Korrespondenz Jakob Burckhardts (<https://burckhardtsources.org/>) und Old Bailey Online (<https://www.oldbaileyonline.org/>).
14. Coyle 2022.
15. Angelehnt an den u. a. von Eide beschriebenen "purpose of learning new things through the modeling activity", vgl. Eide 2014, 5.
16. Erwähnt sei hier auch die Sahle'sche Auslegung von "Text als Inhalt", vgl. Sahle 2013 III, 45ff.

Bibliographie

- Allemang, Dean, und James Hendler. 2011. *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier.
- Broich, Ulrich, und Bernd Schulte-Middelich. 1985. "Intertextualität: Formen Funktionen anglistische Fallstudien." *Konzepte der Sprach- und Literaturwissenschaft* 35. Tübingen: Niemeyer.
- Coyle, Karen. 2022. "Works, Expressions, Manifestations, Items: An Ontology". *Code4Lib Journal* 53.
- Daquino, Marilena, und Francesca Tomasi. 2015. "Historical Context Ontology (HiCO): a conceptual model for describing context information of cultural heritage objects". In *Research Conference on Metadata and Semantic Research*, 424–436. Cham: Springer.
- Davis, Randall, Shrobe, Howard, und Szolovits, Peter. 1993. "What is a knowledge representation?", *AI magazine* 14 (1): 17–17.
- Eide, Øyvind. 2018. "Ontologies, Data Modeling, and TEI". *Journal of the Text Encoding Initiative* 8 (2013). <https://doi.org/10.4000/jtei.1191>.
- Flanders, Julia, und Fotis Jannidis. 2015. "Knowledge organization and data modeling in the humanities". Konferenzbeitrag. urn:nbn:de:bvb:20-opus-111270.
- Fritze, Christiane. 2019. "Wohin mit der digitalen Edition? Ein Beitrag aus der Perspektive der Österreichischen Nationalbibliothek." *Bibliothek Forschung und Praxis* 43 (3): 432–40.
- Madison, Olivia, John Byrum, Jr., Suzanne Jouguelet, Dorothy McGarry, Nancy Williamson, und Maria Witt. 2009. "Functional requirements for bibliographic records final report." September 1997, as amended and corrected through February 2009. International Federation of Library Associations and Institutions.
- Hyvönen, Eero. 2020. "Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery". *Semantic Web* 11 (1): 187–93.
- Llanes-Padm, Dunia, und Juan-Antonio Pastor-Sánchez. 2017. "Records in Contexts: The Road of Archives to Semantic Interoperability". *Program* 51 (4): 387–405. 10.1108/PROG-03-2017-0021.
- Lüschow, Andreas. 2020. "Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane". In *DHd 2020 Spielume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstrakte*, 80–84.
- Müller, Miriam. 2020. "Der sammelnde Professor: Wissensdinge an Universitäten des Alten Reichs im 18. Jahrhundert." *Wissenskulturen. Reihe I, Wissensgeschichte, Bd. 1*. Stuttgart: Franz Steiner Verlag.
- Neuhaus, Stefan. 2014. *Grundriss der Literaturwissenschaft*. 4., überarb. und erw. Aufl. Tübingen: Francke.
- Plachta, Bodo. 2016. "Arbeitsweisen und Editionsstrategien. Eine Annäherung aus historischer Perspektive". In *Textgenese und digitales Edieren. Wolfgang Koeppens "Jugend" im Kontext der Editionsphilologie*, hg. von Katharina Krüger, 19–37. Berlin [ua]: De Gruyter.
- Rehbein, Malte. 2017. "Ontologien". In *Digital Humanities*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein, 162–176. Stuttgart: J.B. Metzler.
- Rheinberger, Hans-Jörg. 1992. "Das 'Epistemische Ding' und seine technischen Bedingungen". In *Experiment-Differenz-Schrift*, hg. von Hans-Jörg Rheinberger und Wim J. Van der Steen, 67–86. Marburg: Basiliken-Press.
- Riva, Pat, Patrick Le Boeuf, und Maja Žumer. 2018. "IFLA library reference model: A conceptual model for bibliographic information".
- Sahle, Patrick. 2010. "Zwischen Mediengebundenheit und Transmedialisierung". *Editio* 24 (2010): 23–36. 10.1515/9783110223163.0.23.
- Sahle, Patrick. 2013. "Digitale Editionsformen: Zum Umgang mit der Überlieferung unter den Bedingungen des

Medienwandels". *Schriften des Instituts für Dokumentologie und Editorik* 7-9. Norderstedt: BoD Books on Demand.

Sahle, Patrick. 2017. "Digitale Edition". In *Digital Humanities*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein, 234-249. Stuttgart: J.B. Metzler.

Spadini, Elena, und Francesca Tomasi. 2021. "Introduction". In *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, hg. von Elena Spadini, Francesca Tomasi und Georg Vogeler, 1-6. Norderstedt: BoD Books on Demand.

Thaut, Lioba. 2010. *Sammeln am Deutschen Hygiene-Museum Dresden 1990 bis 2010: Klassifikation, Kontingenz und Wissensproduktion*. Oldenburg: Institut für Materielle Kultur.

Tillett, Barbara B., 2011. "Keeping libraries relevant in the Semantic Web with resource description and access (RDA)". *Serials* 24(3), 266-272: <http://doi.org/10.1629/24266>.

Tomasi, Francesca. 2012 "Digital editions between embedded markup and external representation. A case study: Vespasiano da Bisticci's Letters". In *Dall'Informatica umanistica alle culture digitali*, hg. von Fabio Ciotti und Gianfranco Crupi, 201-219. Casa Editrice Università La Sapienza.

Van Zundert, Joris. 2016. "Barely Beyond the Book?" In *Digital Scholarly Editing: Theories and Practices*, hg. von Matthew James Driscoll und Elena Pierazzo, 83-106. Open Book Publishers.

Vogeler, Georg. 2021. "Standing-off Trees and Graphs: On the Affordance of Technologies for the Assertive Edition". In *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, hg. von Elena Spadini, Francesca Tomasi und Georg Vogeler, 73-94. Norderstedt: BoD Books on Demand.

Zangerl, Lina Maria, und Christopher Pollin. 2020. "Der Nachlass als Netzwerk: Zur Entwicklung einer Nachlass-Ontologie am Beispiel des Projekts 'Stefan Zweig digital'" In *digital humanities austria 2018: empowering researchers*, hg. von Katharina Zeppezauer-Wachauer et al., 123-127. Austrian Academy of Sciences Press.

Posterpräsentationen

Analyse, Produktion, Reflexion: Nachnutzungsszenarien für Forschungsdaten am Beispiel der Daten des Projekts *Dehmel digital*

Bläß, Sandra

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

Flüh, Marie

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

Nantke, Julia

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

Reul, Christian

christian.reul@uni-wuerzburg.de
Universität Würzburg, Deutschland

Das Ziel wissenschaftlicher Editionen besteht seit jeher in der Nachnutzung durch die (wissenschaftliche) Community. Unter den Vorzeichen von Open Data und Open Science ändern sich allerdings die Möglichkeiten der Bereitstellung und Nachnutzung des erschlossenen Materials. Gleichzeitig haben Wissenschaftler:innen, die mit Methoden der Digital Humanities arbeiten, andere Anforderungen und Bedarfe an bereitgestellte Daten z.B. im Hinblick auf den Umfang der Korpora und die spezifischen Datentypen. Ziel unseres Beitrags ist es, anhand der unterschiedlichen, von uns im digitalen Editions- und Forschungsprojekt *Dehmel digital* (Nantke 2022) produzierten Datentypen systematisch Szenarien der digitalen Nachnutzung durchzuspielen und anhand von Beispielen zu präsentieren. Die Darstellung ist projektbezogen und nicht erschöpfend in Bezug auf alle denkbaren Datentypen und Nutzungsmöglichkeiten. Dennoch sollen die dargestellten Szenarien in ihrer Bandbreite exemplarisch auch für andere Projektkontexte fungieren können.

Wir beziehen uns auf folgende Datentypen: 1) Metadaten von Briefen unterschiedlicher Schreibender aus dem Korrespondenznetz von Ida und Richard Dehmel, 2) digitale Bilder der Dokumente, 3) maschinenlesbarer Text der Briefe, 4) Annotationen von Entitäten sowie 5) algorithmische Modelle.

In Abhängigkeit vom jeweiligen Datentyp ergeben sich unterschiedliche Nachnutzungsszenarien. Diese reichen von dem aus editorischer Sicht klassischen Szenario der

Nutzung der bereitgestellten Daten in (literatur)wissenschaftlichen Analysen über die Nutzung zur Produktion eigener Korpora und Modelle, die dann wiederum Gegenstand der Nachnutzung werden können, bis hin zur Algorithmen-gestützten Reflexion der konzeptuellen Grundlagen einer solchen Datensammlung.

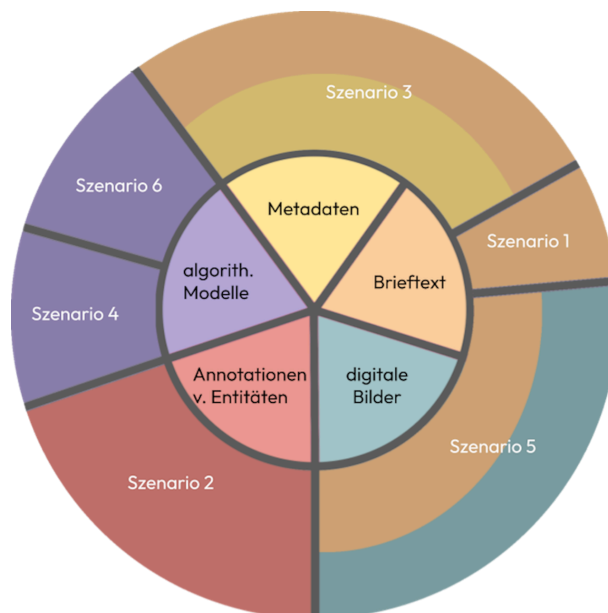


Abb.1: Datentypen und Nachnutzungsszenarien im Überblick

Szenario 1: Analyse von Briefinhalten auf der Basis von Datentyp 3

Briefe stellen relevante Quellen für die Rekonstruktion historischer Diskurse dar (Baillot 2011). Diese in einem Gesamtüberblick und nicht nur in Einzelbeispielen zu erfassen, ist mittels Close Reading-Verfahren kaum zu bewältigen. Ein zentrales computergestütztes Nachnutzungsszenario für unsere Daten sind daher Analysen mittels Distant Reading-Verfahren. Auf der Basis der erzeugten Transkripte kann u.a. eine automatisierte Exploration der zentralen Briefinhalte über Topic Modeling umgesetzt werden (Andorfer 2017; Henny-Kramer/Neuber 2023). Ergänzend hierzu lassen sich z.B. stimmungsmäßige Gewichtungen in den Briefen durch Sentiment Analysis ermitteln und zu den Topics ins Verhältnis setzen.

Szenario 2: Korrespondenznetze sichtbar machen auf der Basis von Datentyp 4

In den im Projekt *Dehmel digital* produzierten Daten sind Entitäten (Personen, Orte, Institutionen, Werke) annotiert. Netzwerkanalysen auf der Basis dieser Annotationen bieten die Möglichkeit, Dynamiken innerhalb der vernetzten Kommunikationspraxis offenzulegen und genauere Einblicke in personelle Kontakte, organisatorische Strukturen und räumliche Bewegungen zu erlangen (Nantke/Bläß/Flüh 2022).

Szenario 3: Vernetzung mit anderen Briefeditionen auf Basis von Datentyp 1 und 3

Die von uns erzeugten Briefmetadaten können über die Plattform *correspSearch* abgerufen werden. Dadurch können Nachnutzende unsere Daten im Rahmen individueller Suchanfragen in Kombination mit den Daten anderer Briefeditionen nutzen.

Szenario 4: Texte erschließen auf der Basis von Datentyp 5

Neben den erschlossenen Dokumenten stellen wir auch die im Projekt von uns trainierten HTR- und NER-Modelle zur Nachnutzung zur Verfügung. Auf Basis dieses Datentyps können weitere Dokumente, die nicht Teil des Projekts sind, erschlossen und somit neue Daten für die weitere Nachnutzung produziert werden. Dies gilt zum einen für handschriftliche Dokumente der Schreibenden, für die wir HTR-Modelle trainiert haben (z.B. Stefan Zweig, Detlev v. Liliencron, Julie Wolfthorn). Zum anderen können die Named Entity-Classifiers für die Erschließung weiterer deutschsprachiger Briefe aus einem ähnlichen Zeitraum genutzt werden. Es besteht auch die Möglichkeit, auf der Basis unserer Trainingsdaten spezifische Modelle für andere Anwendungsfälle nachzutrainieren (Flüh/Lemke 2022).

Szenario 5: gemischte HTR-Modelle trainieren auf Basis von Datentyp 2 und 3

Die in *Dehmel digital* teilautomatisiert generierten, qualitativ hochwertigen Transkripte zahlreicher unterschiedlicher Schreibender bieten in Kombination mit den zugehörigen Bilddigitalisaten den idealen Ausgangspunkt für das Training sog. 'gemischter Modelle', mit deren Hilfe sich deutlich mehr unterschiedliche Handschriften aus dem Zeitraum um 1900 transkribieren lassen.

Szenario 6: Reflexion der theoretischen Fundierung von Datensammlungen auf Basis von Datentyp 5

Unsere NER-Classifiers wurden auf den Dokumententyp 'Brief um 1900' trainiert. Eine experimentelle Anwendung z.B. des Orte-Classifiers auf ein Korpus mit Texten eines deutlich abweichenden Dokumententyps (z.B. fiktionale Texte) kann insbesondere in Kombination mit einem auf den Dokumententyp zugeschnittenen Classifier dazu beitragen, die theoretisch-konzeptuelle Fundierung offenzulegen, welche in die Modellierung des Classifiers eingegangen ist, indem die Ergebnisse der Classifier vergleichend betrachtet werden (vgl. dazu die Fallstudie von Flüh/Schumacher/Nantke im Erscheinen).

Bibliographie

Andorfer, Peter. 2017.: "Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich". *Zeitschrift für digitale Geisteswissenschaften* 2017. https://doi.org/10.17175/2017_002.

Baillot, Anne. 2011. *Netzwerke des Wissens. Das intellektuelle Berlin um 1800*. Berliner Wissenschafts-Verlag.

Flüh, Marie, Mareike Schumacher und Julia Nantke. Im Erscheinen. "Place and Space in Literature. Named Entity Recognition as a Possibility for Spatial Modelling in Computational Literary Studies". In: *Geography Meets Digital Humanities*, hg. von Finn Dammann und Dominik Kremer. Bielefeld: transcript.

Flüh, Marie, und Marc Lemke. 2022. "An Experimental Attempt to Use Transfer Learning for Named Entity Recognition in Letters from the 19th and 20th Century". *DH2022*, 2022.

Henny-Krahmer, Ulrike und Frederike Neuber. 2023. "Topic Modeling in Digital Scholarly Editions". *Machine Learning and Data Mining for Digital Scholarly Editions*, hg. von Bernhard Geiger u. a., Bd. 18, Books on Demand.

Nantke, Julia. 2022. *Dehmel digital*. hg. von ders. unter Mitarbeit von Sandra Bläß und Marie Flüh. <https://dehmel-digital.de> [zugegriffen: 21. Juli 2022].

Nantke, Julia, Sandra Bläß, Marie Flüh und David Maus. 2022. "Best of Both Worlds. Zur Kombination algorithmischer und manueller Verfahren bei der Erschließung großer Handschriftenkorpora". *DHd 2022. Konferenzabstracts*, 2022, <https://doi.org/10.5281/zenodo.6328113>.

A Quantitative Analysis of Digital Scholarly Editions

Kurzmeier, Michael

mkurzmeier@ucc.ie

University College Cork, Irland, Republik

O'Sullivan, James

james.osullivan@ucc.ie

University College Cork, Irland, Republik

Murphy, Órla

o.murphy@ucc.ie

University College Cork, Irland, Republik

Pidd, Michael

m.pidd@sheffield.ac.uk

Digital Humanities Institute, University of Sheffield

Wessels, Bridgette

bridgette.wessels@glasgow.ac.uk

University of Glasgow

Digital scholarly editions are key resources for arts and humanities research, and predate in various forms the concepts of digital humanities or humanities computing (Sula and Hill 2019). While individual projects are remembered for their contribution to the field, few comprehensive data sources exist to show the development of the field. This poster is both an analysis of the sources from which to write a history of digital scholarly editing, and an overview of the state and development of the field using quantitative methods.

Digital editions are positioned between drawing from archived material, and being an archive themselves (Dillen 2019, 266). In addition to that, digital editions also are web resources in need of archiving, lest they fall subject to link rot and very soon disappear from the web either for the lack of a persistent identifier or lack of maintenance. For digital editions past and present, two main data sources are available. Patrick Sahle lists around 700 editions in a curated catalog (Sahle, n.d.), while the Catalogue of Digital Editions features about 320 digital editions in a database (Franzini 2012). Both sources have different criteria for inclusion, overlap in content and differ in gra-

nularity, yet these are the sources from which a history of digital scholarly editions will mostly draw. Analysis of these sources will present them in their scope, aim and usability for research, while highlighting underrepresented areas of data collection on digital scholarly editions.

Adding to the collection of material to describe the history and development of digital editions, the second phase of C21 Editions project engaged in semi-structured interviews with an extensive group of 50 experts and stakeholders from a range of relevant disciplines, including digital scholarly editing, digital publishing, archiving and preservation, interface design, and creative practice. As the field of digital scholarly editions is by now old enough to span entire academic careers, these interviews represent a wealth of insight into the developments of the field. The interviews have been coded and allow for automated analysis of this novel primary source. This poster will present the preliminary results of analyzing the coded interviews with a special emphasis on digital publishing, open data and open access and research infrastructures.

A quantitative analysis of the data sources combined with the coded interviews will then provide data-driven insight into the development of digital scholarly editions since the 1970s. The analysis will in a first step focus on the amount of projects and their average duration over time to produce an overview of the field. In a second step, long-term cycles such as the adaptation of TEI-XML and open access standards will be analyzed. Preservation and availability of all editions listed in both data sources will show the loss rate affecting digital scholarly editions and lead back to a discussion of the current state and history of the field based on the work currently being undertaken within the C21 Editions project.

Acknowledgements

This research is part of C21 Editions: Scholarly Editing and Publishing in the Digital Age, a three-year international collaboration jointly funded by the Arts & Humanities Research Council (AH/W001489/1) and Irish Research Council (IRC/W001489/1).

Bibliographie

Dillen, Wout. 2019. 'On Edited Archives and Archived Editions'. *International Journal of Digital Humanities* 1 (2): 263–77. <https://doi.org/10.1007/s42803-019-00018-4>.

Franzini, Greta. 2012. 'Gfranzini/Digeds_Cat: First Release'. Zenodo. <https://doi.org/10.5281/ZENODO.1161425>.

Sahle, Patrick. n.d. 'A Catalog of Digital Scholarly Editions - by Title, Complete List, a-z (714 Items)'. https://v3.digitale-edition.de/vlet_a-z.html.

Sula, Chris Alen, and Heather V Hill. 2019. 'The Early History of Digital Humanities: An Analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004)'. *Digital Scholarship in the Humanities*, November, fqz072. <https://doi.org/10.1093/llc/fqz072>.

ARS – Architecture Research Stage

Dürfeld, Michael

michael.duerfeld@tu-berlin.de
TU Berlin, Deutschland

Stein, Christian

christian.stein@hu-berlin.de
HU Berlin, Deutschland

List, Ferdinand

f.list@udk-berlin.de
UdK Berlin, Deutschland

Rahman, Zead

zead.rahman@tu-berlin.de
TU Berlin, Deutschland

Dias, Renata

realdias@gmail.com
UdK Berlin, Deutschland

Thran, Niklas

n.thran@udk-berlin.de
UdK Berlin, Deutschland

Marschner, Michèle

michele.marschner@uni-potsdam.de
Uni Potsdam, Deutschland

Die Ansprüche an Wissenschaftskommunikation verändern sich derzeit deutlich: Das Bedürfnis Forschender, ihre Ergebnisse offen zu teilen, sich aktiv mit anderen Forschenden zu vernetzen und selbstorganisiert zusammen zu arbeiten, ist Teil einer umfassenden institutionellen Transformation, die ihre Infrastrukturen noch entwickeln oder weiterentwickeln muss. Digitale Technologien eröffnen hier neue Optionen, die einer im Wandel begriffenen Wissenskultur entgegenkommen und neuen Formen der Konnektivität und des Arbeitens entsprechen können.

Mit der Architecture Research Stage - ARS (www.architectureresearchstage.de) wird eine solche Vernetzungsplattform für die Architekturforschung entwickelt und erprobt. Die Architekturforschung steht dabei vor einer besonderen Herausforderung, da sie in vielen unterschiedlichen und unverbundenen Fachdisziplinen entwickelt wird. Zudem findet diese Forschung nicht allein im akademischen Bereich statt, sondern auch im außer-akademischen Bereich, in den Architektur- und Ingenieurbüros und in der Industrie. Das Potential der Architekturforschung liegt jedoch – wie bei der Architekturpraxis auch – gerade in der Synthese, die das Wissen aus diesen unterschiedlichen Disziplinen zusam-

menführt und vereint. ARS richtet sich deshalb explizit an alle Beteiligten der Architekturforschung: Ob sie im akademischen oder außer-akademischen Bereich forschen, alleine oder im Verbund, ob sie selbst Forschungsergebnisse produzieren oder diese herausgeben, ausstellen, diskutieren oder managen, als Profis, Studierende oder interessierte Laien.

Eine passende Infrastruktur für die spezifischen Bedürfnisse der Forschenden in der Architektur fehlt bisher. Die bestehenden institutionellen Repositorien der Hochschulen oder die überinstitutionellen Repositorien aggregieren und archivieren vornehmlich Forschungsergebnisse und -daten, bieten aber keine überzeugenden Funktionen für das Vernetzen an. Es sind die sogenannten akademischen Netzwerkplattformen wie Academia und ResearchGate, die auf diesen Bedarf reagieren und dort ihre Marktnische gefunden haben. Sie entsprechen jedoch weder den Open-Access-Standards, noch folgen sie den FAIR-Data Prinzipien oder garantieren eine langfristige Speicherung der Daten. Will die Forschungsgemeinschaft sich nicht von wenigen privatwirtschaftlich orientierten Plattformen oder Softwaresystemen abhängig machen, muss sie aus sich selbst heraus disziplinspezifische Vernetzungs-Plattformen entwickeln.

ARS bietet für eine kollaborative Architekturforschung ein übergreifendes Modell und eine aktiv vernetzende Infrastruktur. Erstmals werden nicht nur Ergebnisse und Daten, sondern auch Kontexte der Architekturforschung auf einer webbasierten Plattform gemeinsam generiert, vernetzt und sichtbar gemacht. Auf Grundlage eines Akteursmodells für interdisziplinäre Zusammenarbeit können die jeweiligen Entstehungskontexte durch die Forschenden eigenständig beschrieben, bewertet und mit Alternativen, Wünschen und Bedarfen ergänzt werden. Mit Bezug auf die Akteurs-Netzwerktheorie von Bruno Latour (Latour und Woolgar 1979) und die Netzwerktheorie von Harrison White (White 2008) werden dabei menschliche und nicht-menschliche Einflussfaktoren auf den Forschungsprozess unter dem Akteurs-Begriff versammelt. Aus vorhergehenden Forschungen zu interdisziplinären Forschungskontexten (Dürfeld et al. 2021) konnten dabei elf unterschiedliche Akteursklassen als relevant identifiziert werden: Personen, Organisationen, Themen, Methoden, Quellen, Ereignisse, Aufgaben, Werkzeuge, Orte, Zeiten und Gelder. Diese Akteur:innen sind miteinander durch spezifische, semantisch ausformulierte Bindungen verbunden, wodurch sich komplexe Forschungszusammenhänge einheitlich und vergleichbar modellieren, analysieren und planen lassen.

Zentrales Element von ARS sind die *plusPublikationen*, die Forschungsergebnisse mit deren konkretem Entstehungskontext verbinden. Die von den Forschenden erstellten *plusPublikationen* berichten so immer auch etwas über ihre eigene Entstehung: Welche Werkzeuge und Methoden wurden verwendet, mit welchen Personen wurde diskutiert, welche Organisationen unterstützten die Forschung, welche Quellen wurden verwendet, welche Ereignisse haben den Forschungsprozess beeinflusst, an welchen Orten wurde gearbeitet, welche Aufgaben mussten erledigt werden, wieviel Zeit und Geld stand zur Verfügung? Mittels Graphentechnologie auf der Basis der Open-Source-Graphdatenbank Neo4j werden die Forschungsergebnisse und Forschungskontexte auto-

matisch vernetzt. Über ein bildbasiertes, variables Interface wird das Akteursnetzwerk der Architekturforschung dargestellt und durchsuchbar gemacht.

Der Mehrwert von ARS liegt in fünf Bereichen: (1) Vergrößerung der Sichtbarkeit eigener Forschungen und Vernetzung mit Architekturforschenden durch Präsentation auf einer innovativen Forschungsplattform (2) Passgenaues Auffinden von Kooperationspartner:innen für neue Projekte, (3) Veröffentlichen und Finden von detaillierten, bewerteten Forschungsdaten, (4) Reflektion des eigenen Forschungsverhaltens durch das Erkennen von Wiederholungen, Lücken und Mustern und (5) Partizipation an einer aktiven Architekturforschungs-Community.

Hauptaufgaben des von der DFG für drei Jahre (2021-2024) geförderten experimentellen Pilotprojektes sind Aufbau, Erprobung und Evaluierung von ARS. Das Projekt wird als ein Verbundprojekt der TU Berlin und der UdK Berlin, unter der Beteiligung von Kooperationspartner:innen aus dem akademischen und außer-akademischen Bereich in Berlin-Brandenburg durchgeführt.

Bibliographie

Dürfeld, Michael, Anika Schultz, Christian Stein (et al.). 2021. "Kollaborative Architekturforschung als Programm einer Architekturwissenschaft". In *Architekturwissenschaft. Vom Suffix zur Agenda*, hg. von Juan Almarza Anwandter, Jan Bovelet, Michael Dürfeld (et al.), 210-233. Berlin: Universitätsverlag der TU Berlin.

Latour, Bruno und Steve Woolgar. 1979. *Laboratory Life: The Construction of Scientific Facts*. Beverly Hills: Sage Publications.

White, Harrison C. 2008. *Identity and Control: How Social Formations Emerge*. Princeton: University Press.

Barockpoetik als Wikibase: Eine Datenbank zu konfessionsgeschichtlichen Aspekten in deutschen Barockpoetiken

Haider, Thomas Nikolaus

thomas.haider@uni-goettingen.de
Universität Göttingen, Deutschland

Schennach, Stephanie

stephanie.schennach@uni-goettingen.de
Universität Göttingen, Deutschland

Thelen, Julius

julius.thelen@uni-goettingen.de
Universität Göttingen, Deutschland

Wesche, Jörg

joerg.wesche@uni-goettingen.de
Universität Göttingen, Deutschland

Motivation

Dieses Poster stellt eine Datenbank vor, die sich auf die thesaurierende Erschließung sämtlicher Konfessionsaspekte in den deutschen Poesielehrbüchern der Barockzeit richtet. Das barocke Poetikparadigma, das sich zeitlich von Opitz' *Buch von der Poeterey* (1624) bis zu Gottscheds *Critischer Dichtkunst* (1730) erstreckt (Wesche 2004, 164), wird damit erstmals systematisch im Hinblick auf Fragen der Konfessionalität beforscht, die gerade für das 17. und 18. Jahrhundert von zentraler historischer Bedeutung sind. Das Projekt hat sich zum Ziel gesetzt, die konfessionsgeschichtlichen Inhalte von insgesamt 54 historisch einschlägigen Poetiken offen zugänglich und durchsuchbar zu machen. Damit fällt es unter das Konferenzthema ‚Open Data‘.

Die Datenbank entsteht im Rahmen des Teilprojekts „Uneindeutige Barockdichtung. Poetische und konfessionelle Ambiguität in Schlesien als kulturdynamische Faktoren einer neuen deutschen Dichtkunst (1620 bis 1742)“ der DFG-Forschungsgruppe „Ambiguität und Unterscheidung. Historisch-kulturelle Dynamiken“, das Prof. Wesche gegenwärtig an der Universität Göttingen leitet. Die Wikibase ist zu finden unter <http://barockpoetik.de>

Datenmodellierung

Im Fokus steht, die Modellierung der Daten so vorzunehmen, dass die Inhalte dynamisch abrufbar sind. Dies wurde umgesetzt durch eine dedizierte Wikibase, die unsere spezifischen Inhalte im Stil von Wikidata darstellt und per Volltextsuche und einem SPARQL Query Interface zugänglich macht. Die Daten sind dabei als ‚Items‘ und ‚Properties‘ organisiert, wobei Items als Knoten und Properties als Kanten in einem Graph verstanden werden können, was es uns erlaubt beliebige Beziehungen im Graphen darzustellen und zu suchen.

Zentrale Elemente der Datenbank sind Autoren und ihre Werke, welche als Items gespeichert wurden. Siehe etwa das Item:Q29 (<http://barockpoetik.de/media-wiki/index.php/Item:Q29>), welches den Eintrag für den Autor Martin Opitz festlegt, und dabei diverse Statements definiert, die für die Autoren relevant sind (wie etwa das verfasste Werk, seinen Vor- und Nachnamen, das Geburts- und Todesjahr) und auf einen Eintrag in der GND verweist (Deutsche Nationalbibliothek). Autoren, die in der Wikipedia geführt sind, werden ebenfalls mit einem Link dorthin ausgestattet.

Das assoziierte Werk von Opitz, *Das Buch von der Deutschen Poeterey*, ist hingegen unter Item:Q82 abgelegt (<http://barockpoetik.de/media-wiki/index.php/Item:Q82>). Neben relevanten Metadaten (Publikationsjahr und -ort, Autor, Digitalisat), enthalten Einträge zu Werken Informationen zu bibliographischen Angaben, zur Sekundärliteratur sowie die werkeigenen Kapitelüberschriften. Hinzu kommen als Kernstück der Arbeit Textstellen, die konfessionsgeschichtlich relevant sind. Die Verschlagwortung folgt hier innerhalb der im Projekt entwickelten Systematik ‚Dichtung/Theologie‘, ‚Inspiration‘, ‚Sprachgenealogien‘, ‚Themen/Gattungen‘, ‚Autoritäten‘, ‚Widmungen/Adressaten‘ und ‚Exempelpolitik‘.

Einzelne Textstellen sind dabei als Items angelegt, um sie mehrfach verschlagwortet zu können (wobei eine Textstelle unter verschiedenen Schlagworten geführt werden kann). Zudem besitzt jede Textstelle eine Angabe zu ihrer Fundstelle, um sie im Digitalisat ausfindig zu machen.

Die Volltextsuche erlaubt es, nach bestimmten Inhalten zu suchen. Etwa ergibt eine Suche nach dem Wort ‚Mensch‘ alle Textstellen, in denen dieses Wort vorkommt, oder eine Suche nach ‚[aaq]‘ liefert alle Textstellen mit einer Auszeichnung für die Schriftart ‚Antiqua‘, welche in frühneuzeitlichen Drucken konventionell für fremdsprachliches Material verwendet wird. Diese Textstellen können wiederum durch eine ‚Inverse Suche‘ dem jeweiligen Werk zugeordnet werden.

Der SPARQL Endpoint erlaubt es uns, beliebige Daten zu extrahieren, wie etwa eine Übersicht der Autoren mit ihren Werken, oder zum Beispiel alle Werke, die ‚Exempelpolitik‘ enthalten, mit ihren Autoren, und den entsprechenden Textstellen, sortiert (oder gefiltert) nach Publikationsjahr.

Fazit

Die Datenbank leistet im Bereich der Geschichte der Poetik als einem zentralen Forschungsgebiet der germanistischen Literaturwissenschaft einen substantiellen Beitrag zur Exploration aktueller Methoden der Linked Open Data (Chiarcos et al., 2022; Sturgeon, 2022) und der Nutzung von Wikidata in den Digital Humanities (Zhao, 2022) in einem (frühneuzeit-)historischen Forschungsgebiet, indem es nicht nur um die Digitalisierung und (forschungs-)öffentliche Bereitstellung von Textdaten geht, sondern diese zugleich mit einem auf Fragen der Konfessionalität gerichteten Erkenntnisinteresse digital aufbereitet und empirisch ausgewertet werden. Übergeordnetes Forschungsziel ist es dabei, von der Universität Göttingen aus ein erweiterbares Portal „Barockpoetik digital“ zu etablieren, in dem die Forschungsdaten zum Thema zentral gebündelt und verfügbar gehalten werden. Als konkrete Erweiterungsperspektive erschließt das Team derzeit im Rahmen des DFG-Schwerpunktprogramms „Übersetzungskulturen der Frühen Neuzeit“ sämtliche Aspekte, die im Poetikkorpus translationsgeschichtlich relevant sind.

Bibliographie

Chiarcos, Christian, Ionov, Maxim, Fäth, Christian. 2022. "Linked Open Dictionaries (2015-2022): Achievements, Experiences and Challenges with respect to LOD Technology in Linguistics and the Philologies". In: Book of Abstracts. International Digital Humanities Conference, Tokyo (online).

Sturgeon, Donald. 2022. "Towards a crowdsourced linked open knowledge base of East Asian historical sources". In: Book of Abstracts. International Digital Humanities Conference, Tokyo (online).

Wesche, Jörg. 2004. "Literarische Diversität: Abweichungen, Lizenzen und Spielräume in der deutschen Poesie und Poetik der Barockzeit", Tübingen

Zhao, Fudie, 2022. "How to Critically Utilise Wikidata – A Systematic Review of Wikidata in DH Projects". In: Book of Abstracts. International Digital Humanities Conference, Tokyo (online).

Beginen in Köln: Von der Textdatenbank zur zeitgemäßen digitalen Auszeichnung und Analyse

Bigalke, Jan

Jan.Bigalke@uni-koeln.de

Universität zu Köln, Cologne Center for eHumanities

Blumtritt, Jonathan

jonathan.blumtritt@uni-koeln.de

Universität zu Köln, Cologne Center for eHumanities

Gengnagel, Tessa

tessa.gengnagel@uni-koeln.de

Universität zu Köln, Cologne Center for eHumanities

Beginen sind geistliche Frauen, die ohne Klausur und Gelübde in mittelalterlichen Städten lebten und im Gegensatz zu den klösterlichen Nonnen im Stadtbild präsent waren. Rechtlich waren sie Laien und unterstanden der weltlichen Obrigkeit. Daher sind sie in Rechtsquellen der Stadt präsent, auch in den berühmten Schreinsbüchern der Stadt Köln.

Diese bieten eine für den Betrachtungszeitraum (13.–15. Jh.) außergewöhnlich dichte Überlieferungssituation (Miltzer 1999), da in ihnen bereits seit dem 12. Jahrhundert Rechtsgeschäfte erfasst wurden, zu denen insbesondere Immobilienübertragungen zählten. Durch eine seit den 1990ern andauernde Auswertung der Schreinsbücher konnten bisher rund 2100 Beginen identifiziert werden. Die Ergebnisse dieser Auswertung liegen in Form von Regesten in einer Textdatenbank vor, die zudem In-

formationen zu Verwandtschaftsbeziehungen, der sozialen Stellung der Frauen und ihren Wohnstätten in der Stadt verzeichnet.

Im März 2022 ist an der Universität zu Köln ein DFG-gefördertes Projekt (DFG 2022) unter der Leitung von Frau Dr. Letha Böhringer angelaufen (Böhringer 2022). Das Projekt verfolgt zwei interdependente Ziele: 1.) die erste umfassende Monographie zur Sozialgeschichte der Beginen in Köln vorzulegen, 2.) die Wissensbasis, d.h. die Quellenauswertung, digital so aufzubereiten, dass diese nach prosopographischen, sozialgeschichtlichen und sozialtopographischen Fragestellungen hin ausgewertet und in einer Webplattform öffentlich zugänglich gemacht werden kann. Diese Arbeit wird am Cologne Center for eHumanities durchgeführt, das erste Ergebnisse zur Datenmodellierung und Georeferenzierung mithilfe historischer Kartenmaterials der Stadt Köln vorstellen wird.

Datenmodellierung und Informationsextraktion

Die vorhandenen Regesten mit 4848 Einträgen sind in eine XML-Struktur überführt worden und müssen, da wesentlich als Fließtext festgehalten, durch regex-Anweisungen und XSL-Transformationen in relevante Informationseinheiten vorstrukturiert werden. Hierbei geht es primär um die Identifikation von Akteuren, topographischen Angaben und Relationen (Verwandtschaftsbeziehungen, Transaktionen). Methoden des NLP, wie Named Entity Recognition, werden hierbei ebenfalls ergänzend erwogen.

Bei der Datenmodellierung orientiert sich das Projekt an methodisch verwandten Vorhaben und etablierten Standards (Grünwald 2021). Auf der Ebene der semantischen Verknüpfung wird die eventbasierte Beschreibung durch CIDOC CRM Anwendung finden (s. Abb. 1). Die spezialisierte Bookkeeping Ontology for Historical Accounts (Pollin 2022), die auf CIDOC CRM basiert, wurde ebenfalls evaluiert.

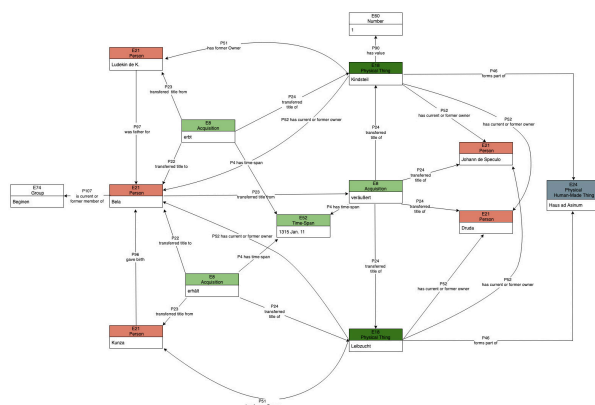


Abb 1.: Regest Schrb. 12 fol. 61r, nach CIDOC CRM V.7.0.1

Kartographie/Topographie

Immobilienbesitz und Wohnstätten der identifizierten Beginen lassen sich in großer Zahl und mit erstaunlicher Genauigkeit lokalisieren. Dies ist vor allem deshalb möglich, weil die Quellengrundlage überwiegend Immobilien-geschäfte zum Gegenstand hat und für die Stadt Köln ein Standardwerk zur städtischen Topographie vorliegt, das ebenfalls auf der Auswertung der Schreinsbücher beruht (Keussen 1910) und ein Referenzierungssystem für Häuser der mittelalterlichen Stadt bietet. Die Referenzen auf Keussens System wurden bei der Erstellung der Fließtext-Regesten bereits systematisch nachgehalten und können so digital genutzt werden.

Dank der Bereitstellung von historischem Kartenmaterial durch die Historische Gesellschaft Köln e.V., in dem die Straßen-Parzellen-Einteilung Keussens in ursprünglich für den Druck vorbereiteten proprietären Formaten des Programms Adobe Freehand abgebildet wurde (die nun durch das Beginen-Projekt verlustfrei in offene Formate überführt worden sind), steht dem Projekt eine vektorisierte Gesamtkarte Kölns zum hohen und späten Mittelalter Kölns nach der zweiten Stadterweiterung zur Verfügung (s. Abb. 2).

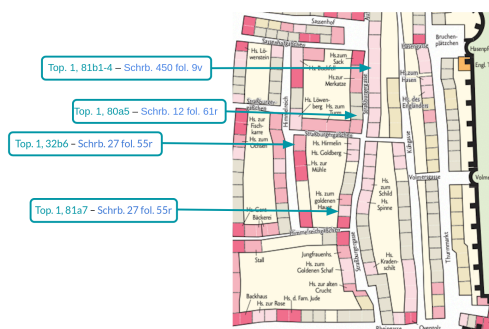


Abb 2.: J. Bigalke/Kliomedia/Historischen Gesellschaft Köln e.V./Grevs Verlag

Dies eröffnet eine Reihe von Möglichkeiten, die vom Minimalziel eines Exports als hochauflösende Rasterdatei, die über einen Map-Tile-Server in Webviewer eingebunden werden kann und das Markieren von Positionen und Clustern auf der Karte ermöglicht, bis hin zum Maximalziel eines Export der Vektoren für unterschiedliche topographische Typen (Mauer, Straßenzüge, Bezirksgrenzen, Kirchen, Klöster, "Parzellen") und Einbindung in GIS-Programme und Webpräsentation reicht.

Das Projekt wird die verschiedenen Optionen vorstellen und in seiner Posterpräsentation der Frage nachgehen, ob sich das Referenzsystem aus Keussens Topographie systematisch auf die eingezeichneten Haussegmente übertragen und so als dynamische Visualisierung auf der Grundlage von Datenbankabfragen realisieren lässt.

Zusammenfassung

Dank der Vorarbeiten in der Auswertung der Schreinsbücher kann das Beginen-Projekt auf Regesten zurück-

greifen, die eine Vielzahl an Informationen zu geistlichen Frauen im Köln des Mittelalters enthalten und es nach ihrer Strukturierung im weiteren Projektverlauf erlauben werden, prosopographische Netzwerkanalysen vorzunehmen. Daher wird das Poster als ersten Kernaspekt die Datenmodellierung eines exemplarischen Regesteneintrags visualisieren.

Zweiter Kernaspekt der DH-Komponente in dem Projekt und entsprechend der Posterpräsentation ist die Georeferenzierung und Verortung der geographisch gebundenen Daten mit den historischen Karten der mittelalterlichen Stadt Köln, sowie die Frage nach den damit verbundenen sozialtopographischen Auswertungsmöglichkeiten. Vorbehaltlich der lokalen Begebenheiten streben wir eine Live-Demonstration der Kartenvisualisierungen auf einem Laptop/Tablet an.

Jenseits dieser ersten vorläufigen Projektergebnisse möchten wir mit dem Poster zur Diskussion stellen, inwiefern das Projekt wichtige Vorarbeiten zum besseren Verständnis der Struktur und Semantik der Schreinsbucheinträge leisten kann, da die vollumfängliche Erschließung der Schreinsbücher ein lange bestehendes Desiderat der mediävistischen Forschung darstellt.

Bibliographie

Böhringer, Letha, und Barbara Weber. 2022. „Beginen - Frauengemeinschaften im Mittelalter Interv. Dr. Letha Böhringer geführt von Barbara Weber“. Deutschlandfunk, 21. April 2022. <https://www.deutschlandfunk.de/beginen-frauengemeinschaften-im-mittelalter-interv-dr-letha-boehringer-dlf-ca4a8b47-100.html>.

DFG. 2022. „DFG - GEPRIS - Beginen in Köln: Sozialgeschichte urbaner Frömmigkeit vom 13. bis zum 15. Jahrhundert“. Zugriffen 22. Juli 2022. <https://gepris.dfg.de/gepris/projekt/491803989>.

Grünwald, Korbinian. 2021. Die digitale Erfassung von mittelalterlichen Rechtsgeschäften - Beschreibung der semistrukturierten XML-Graph-Datenbank db_for_medieval_legal_transactions. https://dhd-blog.org/app/uploads/2021/10/DB_Presentation-2.pdf.

Keussen, Hermann. 1910. Topographie der Stadt Köln im Mittelalter. Band 1. Bonn. <https://www.ub.uni-koeln.de/cdm/ref/collection/rheinmono/id/54222>.

Militzer, Klaus. 1999. „Das topographische Gedächtnis: Schreinskarten und Schreinsbücher.“ In Quellen zur Geschichte der Stadt Köln 1, 165-68.

Pollin, Christopher, und Georg Vogeler. 2022. „Bookkeeping Ontology for Historical Accounts, Version 1.2“, <https://gams.uni-graz.at/o:depcha.bookkeeping>.

Buddhist Murals of Kucha on the Northern Silk Road. Ein Versuch der Semi-Automatisierung der Annotierung

Radisch, Erik

radisch@saw-leipzig.de

Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland

Die buddhistischen Höhlenkomplexe in der Region Kucha an der nördlichen Seidenstraße (Uigurisches Autonomes Gebiet Xinjiang, VR China) beherbergen beeindruckende Wandmalereien, die etwa aus dem 5. bis 10. Jhd. stammen. Die ersten Hinweise auf eine frühere buddhistische Kultur wurden zu Beginn des 20. Jahrhunderts entdeckt, woraufhin mehrere Länder Expeditionen in das Gebiet schickten, um die einst in der Region vorherrschende Religion zu erforschen. Es war eine Sensation, als verschiedene buddhistische Höhlenkomplexe entdeckt wurden. Damals wurden auch die ersten Fotografien vom Ist-Zustand der Höhlen angefertigt und Teile der Malereien aus den Höhlen entnommen und in die jeweiligen Nationalmuseen gebracht. Heute sind Fragmente der Wandmalereien über die ganze Welt verstreut, was eine Zuordnung zu den einzelnen Ursprungshöhlen sehr schwierig macht (Weitere Informationen: Yaldiz 1987; Popova 2008; Dreyer 2015).

Das hier vorgestellte Projekt hat es sich zur Aufgabe gemacht, die Wandmalereien in situ und die weltweit vorhandenen Einzelstücke zu dokumentieren, zu beschreiben und mit Hilfe von historischen Fotografien wieder in ihren ursprünglichen Kontext zu bringen.¹

Das Projekt bedient sich moderner Möglichkeiten der Digital Humanities, indem nicht nur eine umfangreiche textliche Beschreibung einzelner Szenen erfolgt, sondern auch die Bildinhalte der sich wiederholenden Darstellungen erfasst und mit digitalen Methoden angereichert werden. Zu diesem Zweck wird das digitale Bildannotationsstool Annotorious² (siehe Abbildung 1) genutzt, um die Inhalte direkt mit einer rund 1.000 Einträge umfassenden Taxonomie zu annotieren. Die erarbeiteten Forschungsdaten stehen online frei zur Verfügung.³

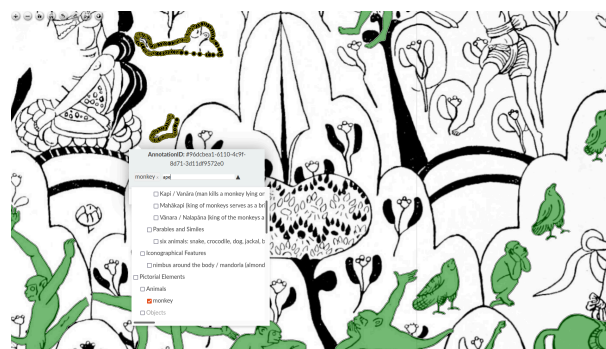


Abbildung 1: Annotieren mit Annotorious

Die Annotation von Objekten im Bild ermöglicht zwar eine wissenschaftliche Nachvollziehbarkeit der identifizierten Objekte, ist aber auch eine sehr umfangreiche und zeitaufwändige Aufgabe. Viele Bildinhalte wiederholen sich häufig in unterschiedlichen Zusammenhängen. Außerdem liegen manchmal mehrere Bilder eines Objekts aus verschiedenen Perspektiven vor oder es gibt Bilder aus der Zeit der Expeditionen, auf denen abgetrennte Teile noch in ihrem ursprünglichen Kontext zu sehen sind. Es besteht also die Notwendigkeit, manchmal sehr ähnliche oder gleiche Objekte mehrfach zu annotieren. Die Übertragung von Annotationen ist jedoch schwierig. Selbst wenn Fotos von denselben Objekten vorhanden sind, können unterschiedliche Blickwinkel und verschiedene Objektive dazu führen, dass die Bilder verzerrt sind. Es ist kaum möglich, diese Aufgabe mit herkömmlichen Computer-Vision-Methoden automatisch durchzuführen.

Aus diesem Grund wird im Rahmen des Projekts derzeit versucht, mit den bereits vorgenommenen Annotationen Region Based Convolutional Neural Networks (RCNNs)⁴ zu trainieren, um in Zukunft zumindest Teile der Annotation halbautomatisch (die Annotierenden werden die Möglichkeit haben, die gefundenen Regionen des RCNNs einzusehen und diese anzunehmen oder gegebenenfalls zu verbessern) durchführen zu können.

Bislang wurden RCNNs in den Digital Humanities vor allem zur Identifizierung, Lokalisierung und Ordnung von Objekten in Bildern eingesetzt (siehe z.B.: Howanitz et al. 2019; Arnold/Tilton 2019; Duhaime 2019; Duhaime 2019; Helm et al. 2021). Ihre Verwendung für eine halbautomatische Annotation ist zumindest zur Kenntnis des Autors dieses Posterproposals noch nicht umgesetzt worden. Da die Ränder automatisch erkannten Annotationen oft ausfransen oder nicht das gesamte Objekt erkannt wird, mag es auch gewagt sein, einen solchen Versuch zu starten.

Die Voraussetzungen des Kucha-Projekts sind sehr gut. Es existieren bereits über 9.000 Polygone, die in insgesamt etwa 12.000 Annotationen verwendet wurden (ein Polygon kann mit mehreren Elementen der Taxonomie verknüpft sein). Einige Objekte wurden mehrer hundert Mal annotiert. Es gibt jedoch auch einige Probleme, die zu berücksichtigen sind. So gibt es beispielsweise zwei grundlegend verschiedene Arten von Bildern im Korpus: zum einen Fotografien (historische und moderne), zum anderen Zeichnungen der Malereien. Die Erkennung auf Fotografien dürfte deutlich schwieriger sein, da hier die

Malereien oft in sehr schlechtem Zustand sind und selbst von einem geübten Auge nur schwer zu identifizieren sind.

Es wurden verschiedene Experimente durchgeführt, die die prinzipielle Nutzbarkeit von RCNNs für eine semi-automatisierte Annotation testen sollten. Diese werden im Poster näher präsentiert. In einem ersten Experiment wurden alle Trainingsdaten zusammen trainiert ($mAP_{IoU=0.75} = 5.85^5$). Ein zweites Experiment teilte die Daten nach Bildarten auf ($mAP_{IoU=0.75} = 3.40$ für Fotos und $mAP_{IoU=0.75} = 4.04$ bei Zeichnungen). Es war dabei auffällig, dass die Zeichnungen bessere Resultate erzielten, die aber niedriger als die Resultate des ersten Experimentes waren. Die folgenden beiden Experimenten wurden deswegen vorerst auf Zeichnungen beschränkt. Das dritte Experiment konzentrierte sich auf Klassen mit mehr als 50 Annotationen, was zwar den Großteil der Klassen außen vor ließ, aber vielversprechende Ergebnisse erzeugte ($mAP_{IoU=0.75} = 23.06$). Deswegen wurden in einem vierten Experiment alle Klassen von menschlichen Abbildungen zu einer Metaklasse zusammengelegt. Dieses letzte Experiment funktionierte besonders vielversprechend ($mAP_{IoU=0.75} = 59.99$; ein Beispiel kann in Abbildung 2 eingesehen werden). Zur Bewertung der Experimente kann der Output der Testbilder online eingesehen werden.⁶ Bis zur DHd im nächsten Jahr sollen einige weitere Experimente durchgeführt werden und ein erster Prototyp zur Anwendung kommen und im Rahmen des Posters präsentiert werden.

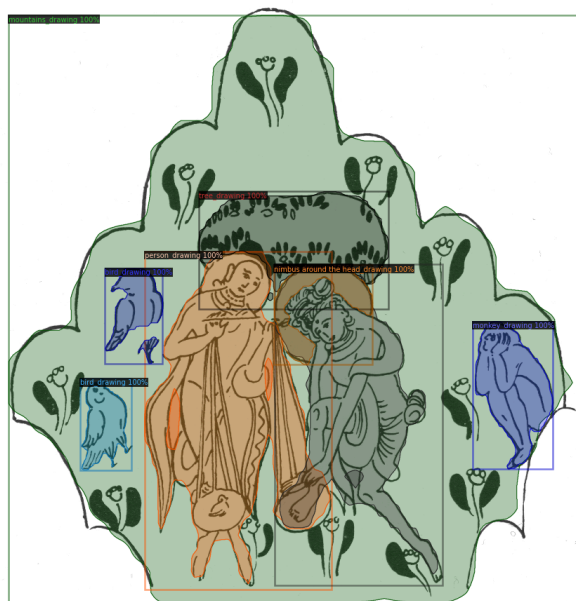


Abbildung 2: Beispiel-Output für Experiment 4

Fußnoten

1. Wissenschaftliche Bearbeitung der buddhistischen Höhlenmalereien in der Kucha-Region der nördlichen Seidenstrasse (SAW-Leipzig). Datenbank: <https://kucha-saw-leipzig.de>.

2. <https://recogito.github.io/annotorious/>. Das Projekt hat seine eigene Reihe «Leipzig Kucha Studies». Erster Band: Konczak-Nagel/Zin 2020

3. <https://kuchatest.saw-leipzig.de/>

4. Das Projekt nutzt hierfür detectron2 (Wu et al. 2019).

5. Die Maßeinheit gibt die Mean Average Precision bei einer Mindestübereinstimmung einer Region mit dem Goldstandard von 75% an.

6. <https://github.com/erikradisch/examplePics>

Bibliographie

Arnold, T. and Tilton, L. 2019. "Distant viewing: analyzing large visual corpora." In *Digital Scholarship in the Humanities* 36(1), DOI: <https://doi.org/10.1093/llc/fqz013>.

Dreyer, C. 2015. "Abenteuer Seidenstraße: Die Berliner Turfan-Expeditionen 1902–1914." Leipzig: Seemann.

Duhaime, D. 2019. "PixPlot." <https://github.com/YaleDHLab/pix-plot>.

Helm, W., Schmideler, S., Im, C., Mandl, T., Kollmann, S. and Müller, L. 2021. "Wie sich die Bilder ähneln. Vom Zufallsfund zur systematischen Forschung im Bereich der automatisierten Bildähnlichkeitssuche." In *Fabrikation von Erkenntnis: Experimente in den Digital Humanities*, ed. Burghardt, M., Dieckmann, L., Steyer, T., Trilcke, P., Walkowski, N.-O., Weis, J. and Wuttke, U. ZfdG (Sonderband 5). DOI: 10.26298/melusina.8f8w-y749-wsdb.

Howanitz, G., Bermeiter, B., Radisch, E., Gassner, S., Rehbein, M. and Handschuh, S. 2019. July 11. "Deep Watching - Towards New Methods of Analyzing Visual Media in Cultural Studies." *Digital Humanities 2019* (DH2019), Utrecht, Netherlands. <https://zenodo.org/record/3326470#.Y5uLSKfMIYw>.

Konczak-Nagel, I. and Zin, M. 2020. "Essays and Studies in the Art of Kucha" (Leipzig Kucha Studies 1). New Delhi: Dev Publishers.

Popova, I. F. (ed.). 2008. "Russian Expeditions to Central Asia at the Turn of the 20th Century: Collected Articles." St Petersburg: Slavia Publishers.

Wu, Y., Kirillov, A., Massa, F., Lo, W. and Girshick, R. 2019. "Detectron2." <https://github.com/facebookresearch/detectron2>.

Yaldiz, M. 1987. "Archäologie und Kunstgeschichte Chinesisch-Zentralasiens (Xinjiang)." *Handbuch der Orientalistik, Abteilung 7, Kunst und Archäologie, Band 3, Innerasien*. Leiden: Brill.

Building a virtual research environment to move from digital to distant Diplomatics (ERC project DiDip)

Vogeler, Georg

georg.vogeler@uni-graz.at
Universität Graz, Österreich

Luger, Daniel

daniel.luger@uni-graz.at
Universität Graz, Österreich

Nicolaou, Anguelos

angelos.nicolaou@uni-graz.at
Universität Graz, Österreich

Kovacs, Tamas

tamas.kovacs@uni-graz.at
Universität Graz, Österreich

Atzenhofer-Baumgartner, Florian

florian.atzenhofer-baumgartner@uni-graz.at
Universität Graz, Österreich

Lamminger, Florian

florian.lamminger@uni-graz.at
Universität Graz, Österreich

Aoun, Sandy

sandy.aoun@uni-graz.at
Universität Graz, Österreich

Decker, Franziska

franziska.decker@uni-graz.at
Universität Graz, Österreich

The ERC Project “From Digital to Distant Diplomatics” (DiDip, <https://didip.eu>) attempts to build an innovative and sustainable (virtual) research infrastructure and environment (VRE) to facilitate large-scale analyses of historical documents. It will extend the Monasterium.net infrastructure, which is provided in an aging software (Bürgermeister et al. 2018). However, Monasterium.net is still the largest repository of digital representations of medieval and early modern charters. For the future use of this corpus, it is crucial to make data, in particular gold standard annotations, and methods as open as possible. We plan to combine traditional approaches to analyzing such charters with state-of-the-art compu-

tational methods and artificial intelligence. The data produced and the methods used will be available under open licenses (code repository <https://github.com/Didip-eu>).

The project addresses an unsolved problem in the domain of diplomatics, i.e., the historical auxiliary science, studying medieval and early modern single sheet legal documents: With pure human intellectual capacity, the empirical part of this research had to focus on local, regional, or chancery level in the face of overwhelming quantities of charters (Hlavacek 2006). While earlier approaches to applying digital methods to the field focused on digital representation of individual descriptions (Ambrosio et al. 2014, Vogeler 2009, Bradley et al 2019), the large-scale “distant reading” approach has been scarce. This changed only recently: In the field of computer vision, Handwritten Text Recognition (HTR) has provided the first results in changing this (Hodel 2017). In addition, Computer Vision (CV) can provide quantified stylistic attributes of all graphical features of a charter. Leifert et al. 2020 and Christlein 2018 extracted graphical elements from charters (e.g., decorations, notarial signs). There are indications that CV can infer the date of a historical document (Cloppet 2017, Seuret 2021) and can classify the handwriting style from a paleographical perspective.

We conclude that a typical CV pipeline for analyzing a charter should consist of: layout analysis, HTR or word segmentation, and finally an analysis of non-text attributes such as style, material, seals etc. Most of these tasks utilize publicly available datasets that are focussed on manuscript books (Simistira 2016) that cannot encapsulate the diversity that is observed in large charter collections.

The most precious resource in such a pipeline is the diplomatist's time spent on annotating data. We drastically economize annotation effort by reformulating layout analysis as an object detection problem instead of the typical image segmentation approach. Indicatively, this allowed us to annotate the layout of 1175 charter images in a fraction of the time that would be needed normally. Figure 1 shows an example of this kind of annotation. Preliminary experiments demonstrate that this approach works well, e.g., it can detect seals with an accuracy above 95% when using a YOLOv5 (Jocher 2022) based model. With a 50% Intersection over Union (IoU) threshold this result is, of course, mainly usable for classification tasks, while segmentation will have to make use of approaches (Leipert et al. 2021). For tasks like HTR, writer identification, and layout analysis, we consider binarization as a useful step. We made experiments indicating that purely synthetic data could be used for the binarization step (Nicolaou 2022), yet comprehensive performance analysis on charters specifically would need manual annotation of the ground-truth pixel-by-pixel. Although the target CV pipeline will be under construction for a while, a few stand-alone methods required for such a pipeline have already been successfully developed and tested.

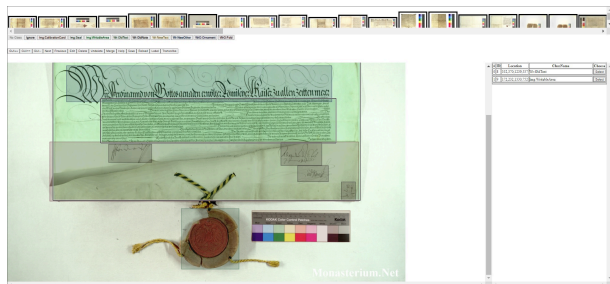


Figure 1: Image annotation example using FRAT (<https://github.com/anguelos/frat>)

A “distant reading” approach has been taken by Teli-huan et al. 2012 and 2014, Perreux 2021, Leclercq et al. 2021, who study statistical features of larger text corpora extracted from charters. We plan to generalize these approaches with the application of Natural Language Processing (NLP) as a custom, multi-level pipeline. It will resolve information retrieval from both HTR and human-produced data, i.e., address tasks like named entity recognition or relationship extraction, text reuse, and in particular formulaicity detection, but also reflect on the possibilities of named entity linking and text summarisation. We are currently working on laying the basis for this. The wide diversity of dialects and less-resourced languages (early vernacular) is one of the most significant analytic difficulties experienced in medieval and early modern charters. Additionally, the existing OCR of charter texts has insufficient quality for further NLP processing. We show the adoption of a multilingual generic system to tackle both problems by BERT (Devlin et al 2019) models. We create a domain-specific model currently under the pseudonym of “RatisBERT” for OCR post correction, that will be made available through the project’s GitHub repositories. It will be based on human controlled data from Monasterium.net itself, charter corpora like ALIM, CDLM, DEEDS, Glessgen 2016, CAO, and enriched by non-charter specific gold standard corpora like the reference corpora for medieval and early modern German (REM, REF, REN including Rhenish) or the PalaFro V2-2. The system takes into consideration a variety of errors originating from HTR and various historical periods and linguistic regions, and provides an effective and automated post-correction approach. We use XLM-RoBERTa for language and variant detection. Through the second layer, the pipeline identifies named entities in the formulaic language of charters, thus forming a solid subset for the abstract generating task, which creates a condensed version of a document in English and other modern languages while preserving its essential information in the standardised format of the charter abstract.

The project is planning to integrate these solutions built on the collection of Monasterium.net with generic Digital Humanities (DH) tools through RESTful application programming interfaces (API) and provides access to its own methods through their own, thriving to set up the domain-specific diplomatics VRE as part of the growing DH API infrastructure.

Bibliographie

ALIM - Archivio della Latinità Italiana del Medioevo <http://www.alim.dfill.univr.it/>

Ambrosio, Antonella, Sébastien Barret, and Georg Vogeler (Eds.). *Digital Diplomatics: The Computer as a Tool for the Diplomatist?* Archiv für Diplomatik. Beiheft 14. Köln, Wien: Böhlau Verlag, 2014. <https://www.degruyter.com/view/title/496882>.

Bradley, John, Dauvit Broun, Alice Rio, und Matthew Hammond. „Exploring a Model for the Semantics of Medieval Legal Charters“. *International Journal of Humanities and Arts Computing* 13, Nr. 1–2 (10. Juni 2019): 136–54. <https://doi.org/10.3366/ijhac.2017.0184>.

Bürgermeister, Martina, Schneider, Gerlinde, Makowski, Stephan, Jeller, Daniel, Bigalke, Jan, Theisen, Christian, und Vogeler, Georg. „Software Aging“ in den DH: Kritik des reinen Forschungswillens“. In *Kritik der digitalen Vernunft. DHd2018. Konferenzabstracts*. Köln: DHd, 2018: 308–11.

CAO - Corpus der altdeutschen Originalurkunden, elektronische Fassung ed. Kurt Gärter, Andrea Rapp. Trier, [2007]. <https://tcdh01.uni-trier.de/cgi-bin/iCorpus/CorpusIndex.tcl>.

Christlein, Vincent. „Automatic Detection of Illuminated Charters“. In *Illuminierte Urkunden. Beiträge Aus Diplomatik, Kunstgeschichte Und Digital Humanities*, herausgegeben von Gabriele Bartz und Markus Gneiß. Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde, Beihefte 15. Wien: Böhlau, 2018: 45–51.

———. „Technical Tools for the Analysis of High Medieval Papal Charters“. In *Papstgeschichte im Digitalen Zeitalter. Neue Zugangsweisen zu einer Kulturgeschichte Europas*, herausgegeben von Klaus Herbers. AfD Beiheft. Wien: Böhlau Verlag GmbH & Cie, 2018: 45–53.

CDLM - Codice Diplomatico della Lombardia Medievale. 2000–2022. <https://www.lombardiabeniculturali.it/cdlm/>

Cloppet, Florence, Véronique Eglin, Marlène Helias-Baron, Cuong Kieu, Nicole Vincent, und Dominique Stutzmann. „ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script“. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017: 1371–76. <https://doi.org/10.1109/ICDAR.2017.224>.

CorA-ReN - Reference Corpus of Middle High German (1050–1350) <https://www.linguistics.rub.de/rem/access/index.en.html>

DEEDS - Documents of Early England Data set. <https://deeds.library.utoronto.ca/>

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, und Kristina Toutanova. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. arXiv, 24. Mai 2019. <http://arxiv.org/abs/1810.04805>.

PALAFRAfro V2-2. In *palaFra*, 2016. <https://palaFra.github.io/fr/texts.html>.

Glessgen, Martin (ed.). *Documents linguistiques galloromans. Édition électronique*, 3eme ed. 2016 <https://www.rose.uzh.ch/docling/>

Gml-Mis - Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650). January 6, 2021. <https://www.fdr.uni-hamburg.de/record/9195>

Hlaváček, Ivan. „Das Problem der Masse: Das Spätmittelalter“. *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 52 (2006): 371–93.

Hodel, Tobias. „Sending 15th-Century Missives through Algorithms: Testing and Evaluating HTR with 2,200 Documents“. In *IMC Leeds 2017 Paper, 11th July*, 2017. <https://solascriptum.wordpress.com/2017/07/11/imc-leeds-paper-sending-15th-century-missives-through-algorithms-testing-and-evaluating-htr-with-2200-documents/>.

Jocher, Glenn, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, u. a. *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference* (Version v6.1). Zenodo, 2022. <https://doi.org/10.5281/ZENODO.3908559>.

Leipert, Martin, Georg Vogeler, Mathias Seuret, Andreas Maier, und Vincent Christlein. „The Notary in the Haystack – Countering Class Imbalance in Document Processing with CNNs“. In *Document Analysis Systems*, herausgegeben von Xiang Bai, Dimosthenis Karatzas, und Daniel Lopresti. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 246–61. https://doi.org/10.1007/978-3-030-57058-3_18.

Nicolaou, Angelos, Vincent Christlein, Edgar Riba, Jian Shi, Georg Vogeler, und Mathias Seuret. „TorMentor: Deterministic Dynamic-Path, Data Augmentations With Fractals“, 2022: 2707–11. https://openaccess.thecvf.com/content/CVPR2022W/ECV/html/Nicolaou_TorMentor_Deterministic_Dynamic-Path_Data_Augmentations_With_Fractals_CVPRW_2022_paper.html.

Perreaux, Nicolas. „Possibilities, Challenges and Limits of a European Charters Corpus (Cartae Europae Medii Aevi - CEMA)“. *arXiv:2105.00932 [cs]*, 21. April 2021. <http://arxiv.org/abs/2105.00932>.

ReF : „Referenzkorpus Frühneuhochdeutsch“, [2018]. <https://www.ruhr-uni-bochum.de/wegera/ref/>.

REM : Klein, Thomas; Wegera, Klaus-Peter; Dipper, Stefanie; Wich-Reif, Claudia (2016). Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0, <https://www.linguistics.ruhr-uni-bochum.de/rem/>. ISLRN 332-536-136-099-5.

REN : Schätzlein, Frank. „Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)“. <https://www.slm.uni-hamburg.de/ren.html>.

Seuret, Mathias, Angelos Nicolaou, Dalia Rodríguez-Salas, Nikolaus Weichselbaumer, Dominique Stutzmann, Martin Mayr, Andreas Maier, und Vincent Christlein. „ICDAR 2021 Competition on Historical Document Classification“. In *Document Analysis and Recognition – ICDAR 2021*, ed. by Josep Lladós, Daniel Lopresti, und Seiichi Uchida, Cham: Springer International Publishing, 2021: 618–34. https://doi.org/10.1007/978-3-030-86337-1_41.

Simistira, Foteini, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, und Rolf Ingold. „DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts“. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2016: 471–76. <https://doi.org/10.1109/ICFHR.2016.0093>.

Tilahun, Gelila, Andrey Feuerverger, und Michael Gervers. „Dating medieval English charters“. *The Annals of Applied Statistics* 6, Nr. 4 (Dezember 2012): 1615–40. <https://doi.org/10.1214/12-AOAS566>.

Tilahun, Gelila, Michael Gervers, und Andrey Feuerverger. „Statistical Methods for Applying Chronology to Undated English Medieval Documents“. In *Digital Diplomatics*, ed. by Antonella Ambrosio, Sébastien Barret, und Georg Vogeler. Köln: Böhlau Verlag, 2014: 211–24. <https://doi.org/10.7788/boehlau.9783412217020.211>.

XLNet-RoBERTa https://huggingface.co/docs/transformers/model_doc/xlnet-roberta.

Vogeler, Georg (ed.). *Digitale Diplomatik*. Köln: Böhlau Verlag, 2009.

Das Deutsche Kunstarchiv auf neuen Wegen

Fischeidl, Kathrin

k.fischeidl@gnm.de

Germanisches Nationalmuseum, Deutschland

Ausgangslage

Mit dem Sammlungsschwerpunkt auf schriftlichen Unterlagen aus dem Bereich der Bildenden Kunst unterscheidet sich das Deutsche Kunstarchiv im Germanischen Nationalmuseum in Nürnberg von Kommunal- oder Staatsarchiven. Die ca. 1450 Bestände, die seit den 1960er Jahren gesammelt werden, umfassen Vor- und Nachlässe aus dem deutschen Sprachraum der Gebiete Architektur, Bildhauerei, Bildwissenschaft, Design, Fotografie, Kunstgeschichte, Kunsthandel, Kunsthandwerk, Malerei und Restaurierung und decken den Zeitraum vom späten 19. Jahrhundert bis in die Gegenwart ab.

Durch den konkreten Schwerpunkt und die strukturelle Ähnlichkeit der einzelnen Bestände wuchs der Wunsch nach einem auf die Archivbestände zugeschnittenen Datenbanksystem.

2019 wurde der Entschluss gefasst, die seit 2009 genutzte hierarchische Datenbank durch eine semantische und graphbasierte Datenbank zu ersetzen. Ein Systemwechsel war unter anderem notwendig, weil die vorherige Archivsoftware nicht 64-bit fähig war. Anpassungen bei der Onlinepräsentation sind ebenso stark eingeschränkt und berücksichtigen nicht die adäquate Darstellbarkeit auf mobilen Geräten. Ein Upgrade auf eine neuere Version desselben Systems ist kostenpflichtig.

Während viele Archive zur Erschließung und Onlinestellung ihrer Bestände auf kommerzielle Softwarelösungen zurückgreifen, haben die Archive des Germanischen Nationalmuseums – das Historische Archiv und das Deutsche Kunstarchiv – einen alternativen Weg beschritten.

Anforderung

Aufgrund der negativen Erfahrung mit der vorherigen Archivsoftware, soll die neue Software Open Source sein. Darüber hinaus ist eine flexible Webpräsentation wichtig, die an die sich stetig ändernden Anforderungen der Nutzer:innen und der Geräte jeweils angepasst werden kann. Die im Archivinformationssystem zur Verfügung gestellten Daten müssen persistent und zitierfähig sein, um Nachhaltigkeit und Zukunftsfähigkeit sicherzustellen. Außerdem ist ein strukturell standardkonformes System zentral für die Implementierung von ISAD(G) (Internationale Grundsätze für die archivische Verzeichnung) und dem ISO-zertifizierten CIDOC (International Committee on Documentation) Conceptual Reference Model des International Council of Museums (ICOM).

Evaluierung

Die zugrundeliegende Ontologie für das semantische Datenmodell ist das CIDOC CRM. Ausschlaggebend für diese Entscheidung waren die Flexibilität und die Erweiterungsmöglichkeiten. So konnten die Anforderungen, die der Wechsel eines Erschließungssystems mit sich bringt, diskutiert und eine für die Bedürfnisse des Deutschen Kunstarchivs maßgeschneiderte aber zugleich den archivischen Standards entsprechende Erschließungsstruktur modelliert werden.

Bei der Konzeption der Erweiterung des CIDOC CRM wurden nicht nur die Vorgaben der inhaltlich beschreibenden Bestandserschließung berücksichtigt, sondern auch Anforderungen der Dokumentation, der Digitalisierung und der Vernetzung innerhalb der Bestände mitbedacht. Aus intensiven Diskussionen mit allen archiv- und datenbankerprobten Mitarbeiter:innen ging die Konzeption von Mechanismen zur Erfassung von archivinternen Workflows hervor. So wurden im Datenmodell Eingabefelder für restauratorische Maßnahmen, Akzessionierung oder das Depotmanagement ergänzt. Die weiteren beschreibenden Erfassungsfelder basieren auf den Erschließungsstandards ISAD(G) sowie RNAB (Ressourcenerschließung mit Normdaten in Archiven und Bibliotheken für Personen-, Familien-, Körperschaftsarchive und Sammlungen). Das entwickelte Archivinformationssystem bietet den Verzeichner:innen auf diese Weise die Möglichkeit auf allen Ebenen - vom Gesamtbestand bis hin zum Einzeldokument - inhaltliche Informationen mit Daten zur Bestandsverwaltung und -konservierung zu verknüpfen.

Die Anwendungsontologie des Deutschen Kunstarchivs bietet neben der Ontologie des International Council on Archives Records in Contexts (RiC-O) eine praxisnahe Diskussionsgrundlage für eine allgemeinere Erweiterung des CIDOC CRM für die archivische Erschließung. Bisher gibt es in diesem Kontext nur vereinzelte Versuche, wie den des Portuguese National Archive (vgl. Melo 2022) oder für Kommunalarchive allgemein (vgl. Vitzthum 2021). Das Portuguese National Archive weicht in der Modellierung des graphbasierten Datenmodells vom Deutschen Kunstarchiv ab. Das heißt, es werden unterschiedliche Klassen des CIDOC CRMs benutzt. Die Modellierung am Beispiel der Kommunalarchive

wendet die RiC-Ontologie auf die Archivbestände an. Das Archivinformationssystem des Deutschen Kunstarchivs zeigt exemplarisch, dass sich mithilfe des CIDOC CRM und WissKI Archivbestände adäquat strukturieren sowie erfassen lassen und sich darüber hinaus auch deren Binnenstruktur abbilden lässt.

Bei der Evaluierung existierender Systeme hinsichtlich der vorliegenden Anforderungen zeigte sich, dass für die Umsetzung des Graphnetzwerks mit einer spezifischen Domänenontologie lediglich WissKI alle Aspekte erfüllt.

Umsetzung, aktueller Stand und Ausblick

Innerhalb von zwei Jahren wurden die Daten ins neue Datenbanksystem migriert. Aktuell verzeichnen die Mitarbeiter:innen des Deutschen Kunstarchivs neue Bestände bereits im WissKI. Um in nationale und internationale Portale Daten liefern zu können, muss ein Export der auf CIDOC CRM basierenden Metadaten in die gewünschte EAD (Encoded Archival Description) xml Struktur entwickelt werden. Zusätzlich wird an einem Konzept zur Onlinestellung der Daten gearbeitet, das im folgenden Jahr umgesetzt werden soll.

Das Poster veranschaulicht zum einen die Entscheidungsgrundlagen und Vorstellungen der Mitarbeiter:innen des Deutschen Kunstarchivs bezüglich der Umsetzung eines Linked Open Data basierten Archivinformationssystems. Zum anderen wird die Modellierung der Domänenontologie auf Basis von CIDOC CRM abgebildet, die für die Umsetzung des Datenmodells notwendig ist.

Bibliographie

Bekiari, Chrysoula und George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, Athanasios Velios. 2022. „Definition of the CIDOC Conceptual Reference Model.“ https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.1.pdf (zugegriffen: 15.12.2022).

Brüning, Rainer und Werner Heegewaldt, Nils Brübach. 2002. „ISAD(G) - Internationale Grundsätze für die archivische Verzeichnung.“ 2. überarbeitete Ausgabe. https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_DE.pdf (zugegriffen: 15.12.2022).

Clavaud, Florence und International Council on Archives Expert Group on Archival Description. 2021. „Records in Contexts Ontology (ICA RiC-O) version 0.2.“ https://www.ica.org/standards/RiC/RiC-O_v0-2.html (zugegriffen: 15.12.2022).

Melo, Dora und Irene Pimenta Rodrigues, Davide Vagnolo. 2022. „A strategy for archives metadata representation on CIDOC-CRM and knowledge discovery.“ In *Semantic Web - Interoperability, Usability, Applicability*, Pre-Press 1-32.

Österreichische Nationalbibliothek, Schweizerische Nationalbibliothek und Staatsbibliothek zu Berlin - Preussischer Kulturbesitz. 2019. „Ressourcenerschließung mit

Normdaten in Archiven und Bibliotheken (RNAB) für Personen-, Familien-, Körperschaftsarchive und Sammlungen. Richtlinie und Regeln.“ Wien, Bern, Berlin. <https://dnb.info/1186104252/34> (zugegriffen: 15.12.2022).

Society of American Archivists, Library of Congress. 1999. „EAD Application Guidelines for Version 1.0“ <https://www.loc.gov/ead/> (zugegriffen: 15.12.2022).

Vitzthum, Maximilian. 2021. „kommis“ – Zur Umsetzung eines auf Records in Contexts basierenden Archivinformationssystems für Kommunalarchive auf Grundlage der virtuellen Forschungsumgebung „WissKI“. Unveröffentlichte Masterarbeit. Erlangen, Nürnberg: Friedrich-Alexander-Universität Erlangen-Nürnberg.

Das preußische Kriegsspiel als Forschungsobjekt

Henning, Pia

pia.henning@dbfz.de

Deutsches Biomasseforschungszentrum; Julius-Maximilians-Universität Würzburg

Historischer Hintergrund

Die Geschichte Preußens und seiner Armee wurde bereits intensiver erforscht. Das Kriegsspiel, das mit dieser verbunden ist, bisher jedoch kaum. Um ein Gesamtbild für diese Zeit und ihre Zusammenhänge erhalten zu können, ist eine kritische Auseinandersetzung mit dem Kriegsspiel ebenfalls nötig. Das 1824 vom preußischen Artillerieoffizier Georg Heinrich von Reisswitz ursprünglich der Armee vorgestellte *Preußische Kriegsspiel* war das erste offiziell eingeführte professionelle Kriegsspiel. Es handelt sich um eine kartenbasierte Simulation bei der Truppenfiguren maßstabsgetreu dargestellt sind. Der Würfel dient zur Ermittlung von Gefechtsausgängen und anderen nicht durch die Anleitungen festgelegten Entscheidungen. Die damaligen Offiziere versuchten militärische Auseinandersetzungen realitätsnah zu simulieren (Wintjes 2017, Henning 2021b). Bis zu Beginn des 20. Jahrhunderts wurde das Kriegsspiel von unterschiedlichen Autoren weiterentwickelt, so dass heute 20 Regelwerke vorliegen (Henning 2021a; 2021b).

Forschungsstand

Aus einem vorangegangenen Projekt liegen Faksimiles von 18 Originalen vor. Diese wurden mit *OCR4all* (Reul et al. 2019) erfasst. Da die Faksimiles in unterschiedlichen Frakturschrifttypen vorliegen, musste das OCR-Modell jeweils neu trainiert und für eine erste statistische Analyse einzeln vorverarbeitet werden. Aufgrund des Umfangs und der unterschiedlichen Bildqualität kann die Digitalisierung und Standardisierung des Korpus als zen-

trale Aufgabe betrachtet werden, bevor die eigentlichen Analysen stattfinden können. Die manuelle Nachkorrektur der Texte und die einheitliche Auszeichnung nach TEI ist ein laufendes Projekt. Weiterhin ist bekannt, dass diverse Übersetzungen der Regelwerke veröffentlicht wurden, allerdings sind diese noch gänzlich unerforscht und bisher nicht im Korpus integriert (Henning 2021a; 2021b, Wintjes 2022).

Teil der bisherigen Analyse war die Extraktion von distinktiven Merkmalen auf Textebene. Diese wurden im Rahmen meiner Masterthesis bereits erstellt und veröffentlicht (Henning, 2021a; Henning, 2021b). So konnte mit Hilfe einer Zeta-Analyse gezeigt werden, dass eine Zuordnung der einzelnen Texte zur Textsorte *Preußisches Kriegsspiel* basierend auf distinktiven Wörtern möglich sein kann. Eine auf BERTs (Devlin et al. 2019) Word Embedding (WE) folgende Support Vector Machine (SVM) zeigt, dass eine automatische Textklassifizierung in ‘Regelwerke’ und ‘militärtheoretische Literatur’ – als Vergleichskorpus – ebenfalls aussichtsreich ist. Auch die Untersuchung der *most frequent words* (MFW), *named entity recognition* (NER) und *part of speech* (POS) helfen Merkmale des Korpus, wie zum Beispiel die Verwendung der Truppengattungen und des Würfels, sowie den allgemeinen sprachlichen Aufbau (z.B. Wort-/Satzlänge, Verteilung von Substantiven/Adjektiven/Verben) der Textkategorie *Preußisches Kriegsspiel* nachzuweisen. Die nach zeitlichen Aspekten gebildeten Teilkorpora sowie das gebildete Vergleichskorpus können ebenfalls anhand der Merkmale unterschieden werden (Versuchsaufbauten: Henning 2021a; 2021b). Dadurch konnte die Hypothese gestärkt werden, dass sich die untersuchten Regelwerke und gebildeten Teilkorpora nicht nur gemeinsam gegen ein Vergleichskorpus abgrenzen lassen, sondern auch untereinander, wodurch eine historische Entwicklung deutlich wird. Durch distinktive Merkmale wie die Verwendung des Würfels als Zufallskomponente und die Veränderung der Gewichtung der Truppengattungen, sind drei Phasen innerhalb der Kriegsspielgeschichte definierbar (Henning 2021a; 2021b).

Sieben Regelwerke, entstanden zwischen 1862 und 1874, der Autoren Wilhelm von Tschischwitz (1862, 1866, 1870, 1874) und Thilo von Trotha (1867, 1870, 1874) wurden bereits einzeln

intensiver untersucht und jeweils zu einer Ausgabe zusammengeführt. Tschischwitz’ Werke bereits veröffentlicht (Wintjes, 2019), Trothas’ liegen in einer bisher unveröffentlichten Abschlussarbeit vor (Henning, 2018). Dadurch ist eine tiefere Analyse des gesamten Korpus angestoßen, sowie die Forschung in der praktischen Durchführung einer Simulation, mit Studenten oder an einer Führungsakademie, ermöglicht worden.

Geplantes Promotionsprojekt

Kernziel des Projekts ist die öffentliche Bereitstellung eines Regelwerkskorpus, inklusive dazugehöriger tabellarischer Beiwerke sowie Übersetzungen. Mit Hilfe von Methoden des *natural language processing* (NLP) sollen inhaltliche und sprachliche Unterschiede und Entwicklungen im Laufe der Entstehungsgeschichte der Kriegsspiele aufgezeigt werden. Ziel ist eine diachrone und ver-

gleichende Analyse der Gattung der Regelwerke sowie das Erstellen einer Ontologie für Regelwerke und nicht-literarischer Texte über das Kriegsspiel hinaus. Durch die Erstellung einer digitalen Sammlung und Ontologie, die auf vollständig annotierten und aufbereiteten Texten basiert, kann in einem Webtool eine direkte visuelle Vergleichbarkeit zwischen individuell gewählten Regelwerken geschaffen werden. Texte, Faksimiles, Analyseergebnisse und weitere für das preußische Kriegsspiel relevante Materialien sollen zentral gesammelt und für die Forschung sowie die interessierte Öffentlichkeit bereitgestellt werden. Beispielsweise verändert sich die Nutzung eines Würfels als zufallsgebende Instanz zur dritten Phase der Kriegsspiele. Auch die Veränderung bei der Verwendung der Truppengattungen, kann so verdeutlicht werden. Dies kann durch Textverweise und Hervorhebungen sowohl im Original, in einer aufbereiteten Version als auch als Graphik kenntlich gemacht werden. Visualisierungen über sprachliche und formale Veränderungen bieten zudem neue Zugänge zu dem Forschungsgegenstand.

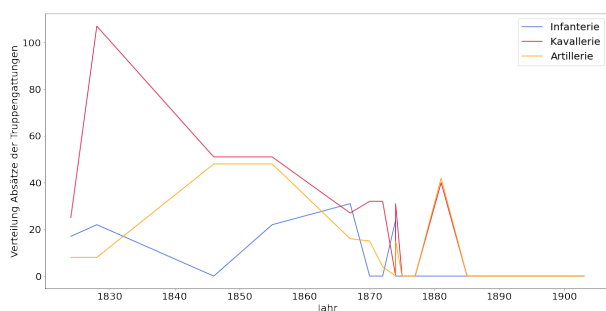


Abbildung 1: Veränderung der Truppengattungen in den Regelwerken

Bibliographie

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (1): 4171-4186. Minneapolis, Minnesota.

Henning, Pia. 2018. "Das Kriegsspiel des Thilo Wolf von Trotha". Bachelorthesis, Julius-Maximilians-Universität Würzburg.

Henning, Pia. 2021a. „Game on!": A Research Project on the Prussian Kriegsspiel". *British Journal for Military History* 7 (2): 174-83. doi:10.25602/GOLD.bjmh.v7i2.1561.

Henning, Pia. 2021b. „Die Entwicklung des preußischen Kriegsspiels". Masterthesis, Julius-Maximilians-Universität Würzburg.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner and Frank Puppe. 2019. „OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings" *Applied Sciences* 9(22): 4853. <https://doi.org/10.3390/app9224853>

Wintjes, Jorit. 2017. „When a Spiel is not a Game". *Vulcan*, 5(1): 5-28.

Wintjes, Jorit. 2019. „Das Kriegsspiel des Wilhelm von ". Hamburg.

Wintjes, Jorit. 2022. "A School for War – A Brief History of the Prussian Kriegsspiel", C. Turnitsa/C. Blais/A. Tolk (eds.), *Simulation and Wargaming*. Hoboken. 25-46

Das QhoD-Projekt: Digitale Edition von Quellen zur habsburgisch- osmanischen Diplomatie 1500–1918

Mayer, Manuela

manuela.mayer@oeaw.ac.at

Österreichische Akademie der Wissenschaften,
Österreich

Kurz, Stephan

stephan.kurz@oeaw.ac.at

Österreichische Akademie der Wissenschaften,
Österreich

Yilmaz, Yasir

yasir.yilmaz@oeaw.ac.at

Österreichische Akademie der Wissenschaften,
Österreich

Sonnberger, Jakob

jakob.sonnberger@uni-graz.at

Zentrum für Informationsmodellierung Universität Graz,
Österreich

Seit 2020 verbindet das Projekt *Digitale Edition von Quellen zur habsburgisch-osmanischen Diplomatie 1500–1918* (QhoD) schriftliche und dingliche Quellen zum diplomatischen Austausch zwischen dem Habsburgerreich und dem Osmanischen Reich. Der lange Zeitraum orientiert sich an der Aufnahme der diplomatischen Beziehungen der beiden Reiche zu Beginn des 16. Jahrhunderts bis zu ihrem politischen Ende 1918. Ediert werden Quellen zu fassbaren diplomatischen Missionen, die in diesen Zeitraum fallen. Bislang wurden bzw. werden vier solcher Missionen in Rahmen von Teilprojekten bearbeitet. Das QhoD-Projekt zeichnet sich durch eine breite Quellenbasis (sowohl quantitativ als auch medial und gattungsmäßig) und einen interdisziplinären Ansatz (Editionstechnik, Frühneuzeitforschung, Kunstgeschichte, Osmanistik, Turkologie) aus. Seit Juli 2022 sind die bislang bearbeiteten Quellen über < <https://qhod.net>

> frei nutzbar. Das den Nutzern zur Verfügung gestellte Material wird regelmäßig erweitert.

Das Poster beschreibt die Quellen und die editorischen wie technischen Überlegungen zu ihrer Edition. Die technische Implementierung erfolgt mithilfe der GAMS repository software zur Datenaufbereitung und -archivierung. Die XML-Codierung erfolgt nach TEI-Standards.

Details zum Projekt:

- Die im Rahmen des Projekts bearbeiteten Quellen bestehen aus handschriftlichen Quellen (u.a. Briefe, Instruktionen, Berichte, Protokollregister, Befehle), gedruckten Quellen (u.a. publizierte Reiseberichte, Flugschriften, Zeitungsartikel, Karten) und Objekten (u.a. Militaria, Kunstgegenstände).
- Bearbeitet werden sowohl habsburgische als auch osmanische Quellen (aus heute öffentlichen wie privaten österreichischen bzw. türkischen Archiven).
- Von allen Quellen werden Faksimiles und bibliographische Metadaten bereitgestellt.
- Objekte werden zusätzlich kunsthistorisch beschrieben.
- Schriftliche Quellen werden transkribiert und annotiert (Personen, Orte, Datierungen, textkritischer Apparat) – es wird aber auch eine Lesefassung ohne Annotationen angeboten. Osmanische Texte werden zusätzlich zur Transkription ins Englische übersetzt. Deutsche Texte bleiben unübersetzt, erhalten aber, wie auch die osmanischen Texte, ein englisches Abstract.

Technische/editorische Details:

- Gedruckte Quellen werden mittels Transkribus HTR erstbearbeitet. Handschriftliche Texte werden vorerst direkt in XML codiert. Derzeit laufen Versuche zur Erkennung osmanischer Texte in Transkribus und dem Training eines entsprechenden Modells.
- Für die Transkription der osmanischen Texte orientiert sich das Projekt an den Richtlinien des *#slam Ansiklopedisi Transkripsiyon Alfabeti*.
- Sammeln und Einspeisen von named entity data in das Austrian Prosopographical Information System (APIS), Anreichern mit GND-Identifiern.

Aktuell beinhaltet QhoD:

- 4 Teilprojekte zu einzelnen diplomatischen Missionen
- nach Mission: 15 Schreiben Selims II. an Maximilian II. (1566-1574); 31 schriftliche Quellen zur Internuntiaturs Johann Rudolf Schmid zu Schwarzenhorns (1649); 157 schriftliche und dingliche Quellen zu den gleichzeitig stattfindenden Großbotschaften Damian Hugo von Virmonts und Ibrahim Paschas (1719/20); Teilprojekt Nr. 4 zur Gesandtschaft Johann Jakob Kurtz von

Senftenaus (1623/24) wurde erst kürzlich begonnen und befindet sich aktuell in der Materialsichtung

- sprachlich: 60 deutsche, 42 osmanische Texte
- nach Genre: 60 Briefe, 20 Protokollregister, 5 in LIDO beschriebene Objekte, 4 Reiseberichte, 4 Berichte, 3 Instruktionen, 16 behördliche Schriftstücke

QhoD versteht sich als offene Plattform zu Austausch und Sammlung von nachnutzbaren Editionsdaten mit dem thematischen Fokus auf die diplomatischen Beziehungen zwischen der Hohen Pforte und dem Wiener Kaiserhof vom Erstkontakt bis zum Ende beider Imperien. Dieser Kontakt bestand auch in Zeiten, in denen sich beide Reiche im Krieg miteinander befanden. Die edierten Quellen liefern somit einen wertvollen Zusatz zur bekannten (Militär-)Geschichte. Laufend werden neue Subprojekte hinzugefügt. Eine Kooperation zur Nutzung der QhoD-Infrastruktur ist nicht gebunden an bestimmte Institutionen, Arten von Fördermittel oder akademische Grade der Beitragenden (einige Transkriptionen wurden im Rahmen von Qualifikationsarbeiten nach QhoD-Richtlinien erstellt).

Bibliographie

Strohmeyer, Arno. 2013a. "Die Theatralität interkulturellen Friedens: Damian Hugo von Virmont als Kaiserlicher Großbotschafter an der Hohen Pforte (1719/20)." In *Frieden und Friedenssicherung in der Frühen Neuzeit. Das Heilige Römische Reich und Europa. Festschrift für Maximilian Lanzinner zum 65. Geburtstag*, 413-438.

Strohmeyer, Arno. 2013b. "Kategorisierungseleistungen und Denkschemata in diplomatischer Kommunikation: Johann Rudolf Schmid zum Schwarzenhorn als kaiserlicher Resident an der Hohen Pforte (1629-1643)." In *Politische Kommunikation zwischen Imperien. Der diplomatische Aktionsraum Südost- und Osteuropa*, 21-29.

Strohmeyer, Arno. 2014. "Krieg und Frieden in den habsburgisch-osmanischen Beziehungen in der Frühen Neuzeit." In *Die Türkei, der deutsche Sprachraum und Europa. Multidisziplinäre Annäherungen und Zugänge*, hg. von Reiner Arntz, Michael Gehler, 31-50.

Yilmaz, Yasir. 2017. "Nebulous Ottomans vs. Good Old Habsburgs: a historiographical comparison." *Austrian History Yearbook* 48: 173-190.

Das Thüringische Flurnamenportal

Aehnlich, Barbara

barbara.aeahnlich@uni-jena.de
Friedrich-Schiller-Universität Jena, Deutschland

Kunze, Petra

petra.kunze@uni-jena.de
Thüringer Universitäts- und Landesbibliothek Jena

Flurnamen sind Benennungen von Örtlichkeiten der Siedlungsflur, die vor allem der Gliederung der Landschaft dienen. Es handelt sich um die Bezeichnungen für Wälder, Felder, Wiesen, Berge, Gewässer und alle anderen natürlichen oder durch den Menschen beeinflussten Geländegegebenheiten, an denen sich der Mensch in der Landschaft orientiert. Auffallend ist ihr eingeschränkter Kommunikationsradius. Sie werden meist nur von Einheimischen benutzt, gelegentlich auch nur von einzelnen Familien. Flurnamen reagieren stark auf gesellschaftliche Veränderungen sowie Gegebenheiten wie Besitzwechsel oder variierende Bodenbewirtschaftung und sind deshalb nicht so stabil wie andere Örtlichkeitsbezeichnungen. Ihre schriftliche Überlieferung ist mündlich geprägt. Flurnamen spiegeln das enge Verhältnis der Namensgeber zu ihrem Lebens- und Arbeitsumfeld wider. Da es zumeist die bäuerliche Landbevölkerung war, die Flurnamen vergab, wurden diese Bestandteil regionaler Identität.

Flurnamen gehören zum immateriellen Kulturerbe: Ihre Erforschung gibt Aufschluss über die Siedlungsgeschichte und ehemalige Raumstrukturen, sie liefert Erkenntnisse über die Entwicklung der deutschen Sprache und ihrer Dialekte. Viele andere Wissenschaftsbereiche wie die Volkskunde, die Wirtschafts- und Sozialgeschichte, die Rechtsgeschichte sowie die Botanik, Zoologie und Geologie profitieren ebenfalls von den Forschungsergebnissen.

In Thüringen wird seit über 110 Jahren Flurnamenforschung betrieben. Dabei wurden insgesamt etwa 200.000 Namen aus Thüringen und dem Süden Sachsen-Anhalts erhoben und in einem Zettelarchiv an der Universität Jena gesammelt. Zusätzlich wurden seit 1999 in einem von Barbara Aehnlich wissenschaftlich betreuten Ehrenamtsprojekt mehr als 500 Sammlungen mit insgesamt rund 40.000 Namen von über 360 aktiven Mitarbeitenden eingereicht. Insgesamt ist mit einem Bestand von 300.000 bis 350.000 Flurnamen für Thüringen zu rechnen.

Die Digitalisierung der umfangreichen Belegsammlung des Flurnamenarchivs wird seit 2019 von der Thüringer Staatskanzlei gefördert und zielt auf die öffentliche Sicht- und Nutzbarkeit des in Kooperation mit der Thüringer Universitäts- und Landesbibliothek (ThULB) entstehenden Portals.¹ In diesem Digitalisierungsprojekt werden die Belege transkribiert und in die Datenbank Collections@UrMEL² eingetragen. Abkürzungen und bibliographische Angaben werden dabei nach Möglichkeit aufgelöst. Etwa 72.000 Namenbelege sind bereits im Portal sichtbar.³ Parallel zur Abschrift werden die gescannten Belege mit den Gemarkungen verknüpft und in einem gemeinsamen Viewer zur Verfügung gestellt. Außerdem werden die Gemarkungen mit der zugehörigen Orts-ID der Gemeinsamen Normdatei (GND) und künftig die Flurstücke mit den offenen Geodaten des Thüringer Landesamtes für Bodenmanagement und Geoinformation verknüpft. Durch die in diesen Datensätzen enthaltenen Informationen wird eine Präsentation der Gemarkungen und Flurstücke in OpenStreetMap möglich. Das Portal macht also das Datenmaterial sicht- und nutzbar und ergänzt die Gemarkungen mit Kartenmaterial. Es stellt den bisher nur in Zettelform zur Verfügung stehen-

den Archivbestand dar und ist ein wichtiger Schritt, die Thüringer Flurnamenforschung in die Zukunft zu führen.

Aktuell wird die Zusammenarbeit von Bürger*innen und Wissenschaft intensiviert. In engem Austausch erheben Ehrenamtliche die Flurnamen und erfassen das lokale Wissen ihrer Heimatorte, indem sie die mündliche Lautung der Namen, Sagen und Legenden sowie Informationen zu landschaftlichen und historischen Gegebenheiten aufzeichnen und private Quellen erschließen. Die Bürger*innen werden durch verschiedene Maßnahmen geschult (Online- und Präsenzveranstaltungen, Flurnamensprechstunden, Workshops), so dass jede Altersgruppe ihren Fähigkeiten entsprechend mitwirken kann. Die Interaktion und der Austausch stehen dabei im Vordergrund. In der ThULB wird derzeit das Datenmodell des Flurnamenportals an diese Datenbestände angepasst und erweitert, damit die Belege passgenau eingegeben und dargestellt werden können. Aktuell wird die webbasierte Eingabemaske so angepasst, dass es in absehbarer Zeit möglich sein wird, dass die Bürger*innen ihre Ergebnisse selbst eintragen und diese nach einer wissenschaftlichen Überprüfung im Portal sichtbar werden. Die Arbeitsschritte im Projekt und deren Erfolge werden im Thüringischen Flurnamenportal, aber auch über Social-Media-Kanäle⁴ und Tagungsbeiträge der Öffentlichkeit vorgestellt.

Langfristig sollen die thüringischen Flurnamen flächendeckend erschlossen sowie sprach- und kulturwissenschaftlich ausgewertet werden. Alle Bestände sollen im Thüringischen Flurnamenportal zusammengeführt werden, um diese Informationen für die interessierte Öffentlichkeit und Forschende gleichermaßen zugänglich zu machen. Die digitale Auswertung und Darstellung des gesammelten sprachlichen Materials im Flurnamenportal verspricht im Sinne von Open Culture Auskünfte über die Geschichte der Orte, frühere Bodennutzungen und Landschaftsgestaltungen, Traditionen und Kultur, Siedlungsströme und Rechtsverhältnisse. Dies großräumig herauszuarbeiten und Verbreitungen auch kartografisch darzustellen, ist das große Ziel des thüringischen Flurnamenprojektes. Das Poster stellt die Arbeitsprozesse bei der Erstellung des Thüringischen Flurnamenportals sowie den bürgerwissenschaftlichen Ansatz vor und zeigt exemplarisch die Herausforderungen und unsere Lösungsansätze. Dabei werden auch der geplante Workflow für die Dateneingabe durch Ehrenamtliche, die Visualisierungsmöglichkeiten, die Social-Media-Aktivitäten und die Nutzung des Portals vorgestellt.

Fußnoten

1. <http://projekte.thulb.uni-jena.de/flurnamen> [letzter Abruf 25.11.2022]. Vgl. dazu ausführlich: Aehnlich (2019, 2021).
2. <https://archive.thulb.uni-jena.de/hisbest> [letzter Abruf 28.07.2021]. Die Universal Multimedia Electronic Library (UrMEL) ist die zentrale Zugangsplattform der Thüringer Universitäts- und Landesbibliothek (ThULB) für ihre multimedialen Angebote. Dabei handelt es sich um wissenschaftliche Informationsangebote und kulturelle Überlieferungen.
3. Stand: November 2022.

4. <https://www.facebook.com/thueringische.flurnamen> und <https://www.instagram.com/thueringische.flurnamen/>

Bibliographie

Barbara Aehnlich. 2021. „Das Thüringer Flurnamenportal – Ein Werkstattbericht“ *Namenkundliche Informationen (NI)* 113: 35–52.

Barbara Aehnlich. 2019. „Die thüringische Flurnamenforschung wird digital“ *Heimat Thüringen*, 26. Jahrgang, Heft 4: 21–24. (<https://www.heimatbund-thueringen.de/publikationen/zeitschrift-heimat-thueringen/heimat-thuerin-gen-heft-42019/>)

Der NFDI4Culture Helpdesk – ein Beratungsangebot für die Kulturwissenschaften

Mayer, Desiree

desiree.mayer@slub-dresden.de
Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (SLUB), Deutschland

Kailus, Angela

kailus@fotomarb.de
Philipps-Universität Marburg

NFDI4Culture¹ ist das Konsortium in der Nationalen Forschungsdateninfrastruktur (NFDI²), das eine bedarfsgerechte Infrastruktur für Forschungsdaten zu materiellen und immateriellen Kulturgütern schafft.

Die in den Blick genommene Forschungslandschaft von Architektur-, Kunst-, Musik- bis hin zur Theater-, Tanz-, Film- und Medienwissenschaften ist durch starke Diversität gekennzeichnet. Sie umfasst Universitätsinstitute, Kunst- und Musikhochschulen, Akademien, Galerien, Bibliotheken, Archive, Museen und einzelne Forscher:innen ebenso Anbieter von Services und Infrastrukturen für diese Bereiche. Das Konsortium will ein Netzwerk etablieren, das an den FAIR-Prinzipien orientierte Angebote für alle Phasen des Forschungsdatenlebenszyklus von der Datengenerierung und -anreicherung über Datenanalyse und -archivierung bis hin zur Daten-distribution und -nachnutzung sicherstellt. Eine Herausforderung sind dabei die sehr unterschiedlichen Anwendungsgebiete, Nutzungszwecke und Softwarelösungen, die in digital gestützten Prozessen berücksichtigt werden müssen, aber auch die häufig sehr unterschiedlichen Vorkenntnisse über einschlägige Verfahren des Forschungsdatenmanagements, Standards und nach-

nutzbare Angebote in einer bislang wenig vernetzt agierenden Fachgemeinschaft.

Profil des Helpdesks

Der NFDI4Culture-Helpdesk ist ein Beratungsangebot³, das eine individuelle, direkte Konsultation zu konkreten Fragestellungen ermöglicht und die Projektkontexte, institutionellen Gegebenheiten und Vorkenntnisse der Nutzenden berücksichtigt. Es ergänzt das Portfolio der Veranstaltungs- und Publikationsformate, mit denen NFDI4Culture Kenntnisse vermitteln und die Integration der Community voranbringen möchte. Der Helpdesk steht allen Akteuren der oben genannten Domänen zur Verfügung, unabhängig von Qualifikationsstufe oder institutioneller Zugehörigkeit. Die Beratung erfolgt gemäß den Schwerpunktthemen von NFDI4Culture:

- Organisatorische und technische Aspekte der Digitalisierung von Kulturgütern (2D, 3D, Audio, Video, AR/VR)
- Datenqualität, Standards, Datenkuratierung
- Umsetzung der FAIR-Prinzipien
- Entwicklung, Konsolidierung, Betrieb und Zertifizierung von nachhaltigen, interoperablen Forschungswerkzeugen und Datendiensten
- Publikationsprozesse, insbesondere für multimodale Publikationen und deren Archivierung in Repositorien
- datenrechtliche und ethische Fragen, etwa zum Urheberrecht, zu Eigentums- und Persönlichkeitsrechten, zum Umgang mit Open Science oder mit kulturell sensiblen Objekten
- Inhalte, Organisation und Gestaltung von Schulungen, Workshops, Lehrmodulen etc. zu Data und Code Literacy
- Informationen zu in Frage kommenden Förderprogrammen und Voraussetzungen für die Projektantragstellung
- Planung des Forschungsdatenmanagements für Projekte und institutionelle Workflows

Für die Beratung steht ein Team bereit, dessen Mitglieder der durchweg selbst über Expertise in einem oder mehreren der Schwerpunktthemen verfügen. Eine Rechtsberatung im juristischen Sinne kann allerdings nicht durchgeführt werden.

Ablauf der Beratung

Der Helpdesk ist über das Kontaktformular auf der NFDI4Culture-Website erreichbar. Die oder der Ratsuchende beschreibt dort die Fragestellung und kann sie bereits einem Schwerpunktthema zuordnen. Die eingegangene Anfrage wird einem Mitglied des Helpdesk-Teams mit einschlägigen Kenntnissen zugeordnet, die oder der als „Datenlotsin“ oder „Datenlotse“ bearbeitet und über den gesamten Beratungsprozess steuernd begleitet. Bei komplexeren Anfragen zieht sie oder er weitere Expertise aus anderen Bereichen des Konsortiums oder auch von anderen vernetzten Forschungsdatenma-

nagement-Initiativen, Serviceeinrichtungen oder Fachleuten hinzu. Gleichzeitig ist sie oder er Ansprechpartner:in für die oder den Ratsuchenden und ist immer informiert über die Aktivitäten rund um die Beratung.

Meist beginnt die Beratung mit einem Gesprächstermin, in dem das Problemfeld näher geschildert wird. Die Informationen werden auf Wunsch vertraulich behandelt. Die Bandbreite der Beratungen reicht von einmaliger Beantwortung einer konkreten Frage bis zur längerfristigen Begleitung eines Projekts von der Antragsphase an. Anfragen dürfen sich in alle Richtungen weiter entwickeln: Was beispielsweise als konkrete Frage zur Lizenzvergabe beginnt, kann in die Vermittlung einer Kooperation mit anderen Projekten münden.

Oft steht hinter einer Anfrage der Wunsch nach Vernetzung. Daher sieht der Helpdesk seine Aufgabe auch darin, Akteure, Projekte und Institutionen miteinander in Verbindung zu bringen, um so die Entwicklung der gesamten Domäne von NFDI4Culture zu unterstützen.

Bedarfsgerechte Ausrichtung

Die Beratungsergebnisse werden protokolliert, denn sie dienen auch dazu, die Community und die Herausforderungen besser kennenzulernen. So können weitere Angebote von NFDI4Culture bedarfsgerechter ausgerichtet werden. In regelmäßigen Treffen arbeitet das Team Helpdesk an der Optimierung seiner Services gemäß den Erfordernissen einer guten klientenzentrierten Beratung.

Akzeptanz des Angebots

Der Helpdesk-Service stößt auf gute Resonanz. Seit dem Projektstart von NFDI4Culture zu Jahresbeginn 2021 gingen 230 Anfragen ein (Stand 1.8.2022). Häufig beziehen sich Anfragen auf Aspekte der Projekt- und Workflowplanung bei der Digitalisierung, Erschließung und Sicherung von Beständen, auf rechtliche Themen, den Einsatz von Normdaten oder auf Softwaretools für bestimmte Anwendungskontexte der digital gestützten Forschung. Die überwiegende Anzahl der Anfragen stammte bis jetzt von Forscher:innen, die im institutionellen Kontext Projekte planen oder durchführen. Beachtlich ist aber auch das Interesse aus bestandshaltenden Institutionen und von Service- und Infrastruktureinrichtungen. Auch Anfragen aus Kontexten, die an den Schnittstellen zu anderen geisteswissenschaftlichen Disziplinen liegen, werden gern entgegen genommen.

Zunehmend entsteht auch eine Zusammenarbeit mit anderen NFDI-Konsortien, den Landesinitiativen zum Forschungsdatenmanagement und ähnlichen Angeboten zum Aufbau effektiver und aufeinander abgestimmter Beratungsservices, so dass der Helpdesk hier einen wertvollen Beitrag zur Vernetzung dieser Initiativen leistet.

Benötigen Sie Unterstützung zu Fragen rund um Forschungsdaten und -software? Kontaktieren Sie den NFDI4Culture-Helpdesk unter <https://nfdi4culture.de/de/kontakt.html>

Fußnoten

1. <https://nfdi4culture.de/>
2. <https://www.nfdi.de/>
3. <https://nfdi4culture.de/go/helpdesk>

Bibliographie

Altenhöner, Reinhard, Ina Blümel, Franziska Boehm, Jens Bove, Katrin Bicher, Christian Bracht, Ortrun Brand. 2020. NFDI4Culture - Consortium for research data on material and immaterial cultural heritage . Research Ideas and Outcomes. <https://doi.org/10.3897/ri-o.6.e57036>

Die digitale Schulbuch-Bibliothek GEI-Digital im neuen Gewand: Ein modernes Präsentationssystem öffnet digitalisierte Schulbücher für die Open Humanities

Klaes, Jan Sebastian

klaes@leibniz-gei.de

Leibniz-Institut für Bildungsmedien | Georg Eckert-Institut, Deutschland

Krüger, Katharina

katharina.krueger@leibniz-gei.de

Leibniz-Institut für Bildungsmedien | Georg Eckert-Institut, Deutschland

Leonhardt, Susann

leonhardt@leibniz-gei.de

Leibniz-Institut für Bildungsmedien | Georg Eckert-Institut, Deutschland

Nieländer, Maret

nielaender@leibniz-gei.de

Leibniz-Institut für Bildungsmedien | Georg Eckert-Institut, Deutschland

Sommer, Kai

sommer@leibniz-gei.de
Leibniz-Institut für Bildungsmedien | Georg Eckert-
Institut, Deutschland

Towara, Nadine

towara@leibniz-gei.de
Leibniz-Institut für Bildungsmedien | Georg Eckert-
Institut, Deutschland

Ausgangslage:

GEI-Digital startete als Digitalisierungsprojekt im Jahr 2009 und bietet Zugang zu digitalisierten Schulbüchern zahlreicher Fächer, Epochen und Regionen. Das Schulbuchkorpus umfasst die Themengebiete Geschichte, Geographie und Politik sowie Fibeln, Atlanten, Realien- und Lesebücher („GEI-Digital“ unter <https://gei-digital.gei.de/viewer/index/>).

Der Digitalisierungsprozess umfasst dabei die Herstellung hochauflösender digitaler Images, die Erzeugung von Meta- und Strukturdaten, die Volltexterfassung, die Präsentation der Digitalisate auf der GEI-Digital-Website und Maßnahmen zur Langzeitsicherung. Gegenwärtig liegt der Fokus von GEI-Digital auf deutschsprachigen Werken, die in der Zeit zwischen dem ersten Aufkommen des Schulbuches im 16. Jahrhundert bis zum Ende des Kaiserreiches im Jahre 1918 erschienen sind.

Die aktuell verfügbaren Sammlungen bzw. Teilkorpora auf GEI-Digital stellen eine Auswahl an historischen Schulbüchern dar, die als repräsentativ für das jeweilige Fach, die einzelnen Schulstufen und -formen und regional unterschiedlichen Ausgaben gelten kann. Im Hinblick auf die hohe Schulbuchproduktion im Deutschen Reich und die Bedeutung der Epoche für die Konstruktion einer nationalen Identität wird das digitale Korpus fachspezifisch und zeitlich in separaten Kollektionen auf GEI-Digital präsentiert. Internationale historische Schulbücher sind punktuell vorhanden und werden perspektivisch weiter ausgebaut (Hertling und Klaes 2018a, 21-44 und Hertling und Klaes 2018b 45-68).

Digitalisiert und unter einer Public Domain-Lizenz zur freien Nachnutzung zur Verfügung gestellt, werden nach dem geltenden Urheberrechtsgesetz (UrhG) bislang nur Werke, deren Autoren seit mehr als 70 Jahren verstorben sind. Derzeit beinhaltet GEI-Digital circa 7.400 Schulbücher mit über 1,8 Millionen Seiten. Davon konnten bereits 1,4 Millionen einer Volltexterkennung zugeführt werden. Zu den nicht im Volltext vorliegenden Beständen zählen aufgrund der überwiegend grafischen Darstellungen, Atlanten, sog. Schreiblesefibeln, und Bestände in Frakturschrift mit einem Erscheinungsjahr vor 1800. Die Implementierung einer Volltexterkennung insbesondere für ältere Frakturschriften bei mittelgroßen Sammlungen ist Gegenstand eines von der DFG geförderten Verbundprojekts „OCR-D OCR4all-libraries“ (Engl et.al. 2020 und „DFG – GEPRIS“ o.J.). Dabei wird in Zusammenarbeit mit dem Zentrum für Philologie und Digitalität (ZPD) der Universität Würzburg das Open-Source-Werkzeug OCR4all (Reul et. al. 2019) so erweitert und angepasst, dass es von

Bibliotheken und Archiven bei der Digitalisierung größerer Mengen eingesetzt werden kann.

In den letzten 10 Jahren haben sich die Anforderungen an das Design und Präsentation digitaler Inhalte massiv geändert. Vor dem Hintergrund hat sich das GEI entschieden den Web-Auftritt von GEI-Digital grundlegend zu erneuern und zu optimieren. Die Präsentation der Inhalte erfolgte bisher auf Basis des von der Firma Intranda und unter der Open Source-Lizenz stehenden Softwarelösung „Goobi Workflow“ und „Goobi Viewer“, die am GEI seit 2012 unverändert eingesetzt wird (Hankiewicz 2019, 77-88).

Vorgehen:

Bereits 2014 wurde eine umfangreiche Nutzer:innenbefragung durchgeführt, die wichtige Bedarfe sichtbar machte. In der Folge wurden gezielte Feedbackrunden mit Forschenden in den Digital Humanities etabliert, um die sichtbaren Bedarfe zu konkretisieren. Unter Berücksichtigung dieser Ausgangslage begann am GEI eine interdisziplinäre Arbeitsgruppe aus den forschenden Abteilungen und der Forschungsbibliothek einen Katalog mit Optimierungsanforderungen zusammen zu stellen. Dabei kristallisierten sich Volltextverfügbarkeit, Sammlungsverwaltung und Schnittstellen für die Datenbereitstellung als zukünftige Kernbedarfe dieser Zielgruppe heraus.

Allerdings erforderte die 2020 veraltete Softwareinfrastruktur des Goobi Viewers als Fundament für die Realisierung der Kernbedarfe ein umfassendes Upgrade. Um aber den weiterentwickelten Goobi Viewer auch für andere Anwenderinstitutionen möglichst einfach nachnutzen zu können, hat das GEI zusammen mit dem Entwicklerteam von Intranda entschieden, diese Software in Form einer unter Open Source laufenden Containervirtualisierung, auch Docker genannt, zur Verfügung zu stellen. Docker erlaubt eine vereinfachte Bereitstellung von Software-Anwendung, da alle relevanten Softwarepakete enthalten und somit schnell für eine Installation bereitstehen. Die gesamte Goobi Viewer-Software steht jetzt als Docker-Anwendung unter GitHub zur Nachnutzung bereit (Goobi-Viewer-Docker o.J.). Eng verbunden mit dem Implementierungsprozess war auch eine starke Fokussierung auf die Lokalisierung, Eingrenzung und Behebung von Problemen und Fehlern. Die interdisziplinäre Arbeitsgruppe konnte gerade im Sinne eines agilen Vorgehens schnell und effektiv Herausforderungen angehen und Lösungsansätze erarbeiten.

Zusammenfassung und Ausblick:

Das abgestufte Vorgehen mit einer Nutzer:innenbefragung, gezielten Feedbackrunden mit Forschenden und einer interdisziplinären Arbeitsgruppe erwies sich in dem hier vorliegenden Szenario als gewinnbringend. Mit Abschluss der Phase eines ersten technischen Upgrades und einer umfassenden Erneuerung der Software-Infrastruktur steht die digitale Schulbuchbibliothek GEI-

Digital in einem neuen Outfit für die Nutzer:innen in den Digital Humanities zur Verfügung. Gerade die interdisziplinäre Zusammenarbeit von Akteur:innen aus den Forschungsinfrastrukturen und den wissenschaftlichen Anwender:innen konnte einen erfolgreichen Neustart sicherstellen und die nachhaltige Grundlage für weitere Optimierungen legen.

Der kontinuierliche Betrieb einer digitalen Bibliothek und deren Optimierung sollte als eine Daueraufgabe verstanden werden. Vor dem Hintergrund ist vorgesehen, die Nutzungsmöglichkeiten von GEI-Digital für die vielfältigen Bedarfe weiter zu verbessern. Zu den nächsten Schritten zählen die Bereitstellung nachnutzbarer Datenformate nach aktuellen Standards und die Persistenz der Daten im Hinblick auf z.B. die verschiedenen Volltextversionen bei Anwendung verschiedener OCR-Engines im Kontext von Forschungsdaten.

Bibliographie

„DFG - GEPRIS - OCR4all-libraries - Volltexterkennung historischer Sammlungen“. o. J. Zugriffen 14. Dezember 2022. <https://gepris.dfg.de/gepris/projekt/460665940?context=projekt&task=showDetail&id=460665940&>.

Engl, Elisabeth, Konstantin Baierer, Matthias Boenig, Volker Hartmann, und Clemens Neudecker. 2020. „Volltexte – die Zukunft alter Drucke: Bericht zum Abschlussworkshop des OCR-D-Projekts“. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 7 (2): 1–4. <https://doi.org/10.5282/o-bib/5600>.

„GEI-Digital“. 2022. Zugriffen 10. Dezember 2022. <https://gei-digital.gei.de/viewer/index/>.

Goobi-Viewer-Docker. (2022), Abgerufen 3. August 2022, von <https://github.com/intranda/goobi-viewer-docker>

Hankiewicz, Steffen. 2018. „Goobi entwickeln – Eine Open-Source Software zur Verwaltung von Workflows in Digitalisierungsprojekten“. In *Digitalisierung in Bibliotheken*, herausgegeben von Gregor Neuböck, 77–88. De Gruyter. <https://doi.org/10.1515/9783110501094-006>.

Hertling, Anke, und Sebastian Klaes. 2018a. „Historische Schulbücher als digitales Korpus für die Forschung: Auswahl und Aufbau einer digitalen Schulbuchbibliothek“. In *Digital Humanities in der internationalen Schulbuchforschung*, Volume 9:21–44. Eckert. Expertise, Volume 9. V&R unipress. <https://doi.org/10.14220/9783737009539.21>.

Hertling, Anke, und Sebastian Klaes. 2018b. „»GEI-Digital« als Grundlage für Digital-Humanities-Projekte: Erschließung und Datenaufbereitung“. In *Digital Humanities in der internationalen Schulbuchforschung*, Volume 9:45–68. Eckert. Expertise, Volume 9. V&R unipress. <https://doi.org/10.14220/9783737009539.45>.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, und Frank Puppe. 2019. „OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings“. *Applied Sciences* 9 (22): 4853. <https://doi.org/10.3390/app9224853>.

Die Wahl der Mittel – Jupyter-Notebooks als Forschungsinfrastruktur

Jung, Kerstin

kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Hein, Pascal

pascal.hein@ilw.uni-stuttgart.de
Universität Stuttgart, Institut für Literaturwissenschaft

Blessing, André

andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Hess, Jan

jan.hess@dla-marbach.de
Deutsches Literaturarchiv Marbach

Kushnarenko, Volodymyr

volodymyr.kushnarenko@hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart

Mit Python als vielgenutzter Programmiersprache in den Digital Humanities¹ steigt auch der Bedarf an Möglichkeiten zur nachhaltigen Weitergabe und Wiederverwendbarkeit von Python-Quellcode. Softwareentwicklungsnahen Lösungen wie der Verfügbarmachung über Versionsverwaltungsrepositorien oder dem Einpflegen in eine Paketverwaltung wurde der ‚Notebook‘-Ansatz² zur Seite gestellt, der Dokumentation, Ausführung und Visualisierung verzahnt und eine Aufbereitung für verschiedene Zielgruppen ermöglicht.

Im Forschungskontext werden solche Notebooks daher verwendet, um auf einer (Web-)Seite Datensätze einzulesen, zu analysieren, visualisieren und die verwendete Methodik zu erläutern, ohne dies auf verschiedene Orte oder Zugänge verteilen zu müssen. In der (Nach-)Nutzung können z. B. Parameter in der Analyse oder Visualisierung direkt im Browser verändert werden und eine Anpassung ohne Programmierkenntnisse oder -erfahrung ermöglichen. Die Notebook-Dateien können wiederum über entsprechende Softwareentwicklungs-Repositorien zur Verfügung gestellt werden, was Anpassungen für weitere Datensätze oder Forschungsfragen erlaubt. Jupyter-Notebooks sind dabei als JSON-Dokumente strukturiert verarbeitbar.

Im Rahmen unseres Projekts geht es uns um die Möglichkeit, Jupyter-Notebooks so zur Verfügung zu stellen, dass sie für eine sehr heterogene Nutzengruppe (u. a. Autor*innen, Forschende, Schüler*innen) einen Mehr-

wert bedeuten.³ Wir möchten verschiedene Ebenen der Vorkenntnisse bedienen und gleichzeitig ermöglichen, eigene Forschungsfragen einzubringen. Im Beitrag soll aber auch der infrastrukturelle Aufwand verdeutlicht werden, der hinter der Möglichkeit nachhaltig ausführbarer Notebooks für die Forschung steht und auf einen Ausgleich zwischen Flexibilität der Nutzung und Sicherheit des Angebots hinausläuft. Letztendlich möchte der Beitrag die Diskussion befördern, inwieweit die Community von einer forschungsgetriebenen, unabhängigen und nachhaltigen Infrastruktur zum Umgang mit Jupyter-Notebooks profitieren würde, da ein entsprechendes Vorgehen für individuelle Projekte weniger umsetzbar ist.

Zu den Vorteilen der Bereitstellung von Zugängen zu Daten und Analysen durch Notebooks gehören (i) die Möglichkeit, ein Angebot an eine breite Nutzengruppe zu machen: Je nach Aufbereitung der Notebooks (interaktive Elemente wie Dropdown-Menüs oder Range-Sliders sind möglich) können sie fast ohne Vorkenntnisse mit Python betrieben werden und an das individuelle Forschungsinteresse angepasste Ergebnisse produziert werden, (ii) dass die technischen Voraussetzungen, z. B. benötigte Pakete, im Notebook selbst spezifiziert sind. Diese Vorteile kommen allerdings nur in einer konfigurierten Ausführungsumgebung zum Tragen. Werden nur die Jupyter-Notebook-Dateien bereitgestellt, setzt das bei den Nutzenden Kenntnisse in Python, Bash o. Ä. sowie im Umgang mit Jupyter voraus. Oft sind Pakete in aufeinander abgestimmten Versionen erforderlich oder in Abhängigkeit vom Betriebssystem verfügbar, so dass nur eine vorkonfigurierte Umgebung den Nutzenden tatsächlich die technischen Hürden abnimmt.

So stellt sich die Frage, in welchem Rahmen ausführbare Jupyter-Notebooks zur Verfügung gestellt werden können. Der Betrieb einer zugänglichen Ausführungsumgebung („Hub“) setzt Hardware, Administrations- und Wartungskapazitäten voraus. Eine Nutzungsverwaltung (Vergabe und Pflege von Accounts, Monitoring von Speicher- und Rechenkapazitäten) ist dabei ebenso unerlässlich wie Aktualisierungen mittels Updates auf Ebene von Maschine, Hub und Paketen und damit verbundene Wartungsarbeiten durch Abhängigkeiten in den Notebooks. Der Betrieb einer nachhaltigen Ausführungsumgebung setzt dies für einen längeren Zeitraum voraus, so dass die Idee der eigenen Ausführungsumgebung den Rahmen eines Forschungsprojekts oft übersteigt. Des Weiteren muss der Sicherheitsaspekt berücksichtigt werden, da es sich bei ausführbaren Jupyter-Notebooks um ausführbaren Quellcode handelt, der gewollt oder ungewollt Schaden am eigenen oder an externen Systemen verursachen kann.

Mit dem Service Colaboratory⁴ bietet Google an, Notebooks einzurichten, Pakete dafür dauerhaft zu installieren und diese Notebooks ggf. mit Zugangsbeschränkung zu veröffentlichen. Dies zeigt die technische Möglichkeit, einen Notebook-Service mit großen Gestaltungsmöglichkeiten für die Nutzenden zu hosten. Allerdings kann diese Lösung für die Forschung nicht als Standard vorgeschlagen werden – allein aufgrund der problematischen Speicherung aller Daten, aber auch weil hier aufgrund des unvorhersehbaren Umgangs mit den eigenen Diensten die Nachhaltigkeit nicht gesichert werden kann.

Eine Alternative hierzu kann der Betrieb einer stark restringierten Ausführungsumgebung sein, die zwar die vorhandenen Notebooks abspielen kann und Nutzende ggf. aus vorgegebenen Parametern wählen lässt, Forschenden aber kaum Flexibilität bezüglich einer eigenen Exploration oder Einbindung weiterer Pakete ermöglicht.

Sofern spezifische technische Expertise angenommen werden kann, ist eine weitere Möglichkeit, Docker-Container zum Download zur Verfügung zu stellen oder eine detaillierte Dokumentation zur Nutzung eines Notebooks innerhalb einer integrierten Entwicklungsumgebung zu liefern. Die Zielgruppe wird damit allerdings auf Nutzende der entsprechenden Infrastruktur eingeschränkt.

Notebooks, die über bestimmte Repositorien öffentlich zur Verfügung gestellt werden, können über Binder⁵ in eine Ausführungsumgebung gebracht werden. Dabei sind vor allem forschungsbezogene Repositorien wie Zenodo von Interesse, die gute Voraussetzungen für die langfristige Verfügbarkeit auf geschützten Servern bieten. Forschungsgetriebene Ansätze wie von GESIS⁶ mit Binder (Bleier und Erdogan 2020) sowie Forschungsumgebungen mit Nutzendenverwaltung wie das DHVLab⁷ und DH2go⁸ (Heckelen et al. 2022), die die Ausführung von Jupyter-Notebooks erlauben, ermöglichen ggf. auch den Umgang mit spezifischeren Daten, sind aber ggf. auf Nutzende aus bestimmten Fachbereichen oder Institutionen beschränkt.

Ein entsprechender Ansatz für die breite Forschungscommunity wäre ein großer Gewinn bezüglich der Verfügbarmachung, Nachnutzung und Dokumentation von Forschungsmethoden und -ergebnissen.

Fußnoten

1. Vgl. die Wellen der Sprachen Fortran, Prolog, Perl und Python in der Korpusstudie von Burghardt et al. (2022).
2. Jupyter-Notebooks, <https://jupyter.org/> (zugegriffen: 15. Dezember 2022).
3. Im Projekt SDC4Lit (Science Data Center for Literature, <https://www.sdc4lit.de/>, zugegriffen: 15. Dezember 2022) werden u.a. literarische Blogs zur Untersuchung zur Verfügung gestellt. Dabei kommen z.B. Tools zu Linkextraktion und Graphaufbau zum Einsatz, die über Jupyter-Notebooks bereitgestellt werden können.
4. <https://colab.research.google.com/> (zugegriffen: 15. Dezember 2022).
5. <https://mybinder.org/> (zugegriffen: 15. Dezember 2022).
6. <https://notebooks.gesis.org/binder/> (zugegriffen: 15. Dezember 2022).
7. <https://dhvlab.gwi.uni-muenchen.de/> (zugegriffen: 15. Dezember 2022).
8. <https://dh2go.ilw.uni-stuttgart.de/> (zugegriffen: 15. Dezember 2022).

Bibliographie

Bleier, Arnim und Kenan Erdogan. 2020. „A Persistent BinderHub: Democratizing Access to Computational Re-

sources in the Social Sciences." JupyterCon2020 Online Conference.

Burghardt, Manuel, Jan Luhmann und Andreas Niekler. 2022. "Tools as Epistemologies in DH? A Corpus-Based Exploration." In *Digital Humanities 2022. Conference Abstracts*, 144-146. Tokyo, Japan.

Heckelen, Malte, Claus-Michael Schlesinger und Fabienne Burkhard. 2022. "Dh2go - Lehr- Und Lernumgebung Für Die Digital Humanities." In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*. Zenodo. <https://doi.org/10.5281/zenodo.6328013>

DigEdTnT – Digital Edition Creation Pipelines: Tools and Transitions

Pollin, Christopher

christopher.pollin@uni-graz.at

Zentrum für Informationsmodellierung, Graz

Strutz, Sabrina

sabrina.strutz@uni-graz.at

Zentrum für Informationsmodellierung, Graz

Steiner, Christian

christian.steiner@dhcraft.org

Zentrum für Informationsmodellierung, Graz

Klug, Helmut

helmut.klug@uni-graz.at

Zentrum für Informationsmodellierung, Graz

Digitale Editionen sind ein Kernbereich der Digital Humanities; sie machen historische Quellen zugänglich. Dabei werden computergestützte Methoden zur Umsetzung, Verbreitung und Erforschung von wissenschaftlich fundierten Quellenveröffentlichungen herangezogen. Digitale Editionen umfassen dabei textuelle, visuelle und ggf. auch quantitative Daten und erfordern oft spezielle Benutzeroberflächen, um domänenspezifische Forschungsfragen zu bearbeiten. Obwohl jedes Editionsprojekt seine eigenen spezifischen Anforderungen hat, lassen sich einzelne Schritte identifizieren, die für Editionsprojekte generell notwendig sind. Das ist im weitesten Sinne die Digitalisierung der Quelle mit der Verwaltung von Bildern und Text, die Transkription, die Modellierung relevanter Textphänomene mittels adäquater Auszeichnungssprachen, die Annotation semantischer Informationen und Named Entities, die Erstellung von Indizes, sowie eine den FAIR-Kriterien entsprechende Publikation über das Web. In den letzten Jahren wurde eine Vielzahl an Tools entwickelt, die für all diese Schritte eingesetzt werden.

Editionen bauen in der Regel auf Bilddigitalisaten der Quelle auf, deren Erstellung und Zurverfügungstellung im Aufgabenbereich von Bibliotheken und Archiven liegt. Ein Zugriff darauf ist im Idealfall mittels iif möglich.

Die Transkription von Texten kann manuell, über Crowdsourcing (z. B. FromThePage) oder automatisiert (z. B. Transkribus) durchgeführt werden. Relevant dabei ist, dass unabhängig vom Werkzeug die Umwandlung des transkribierten Textes in XML/TEI möglich ist. Im bereits modellierten XML/TEI werden schließlich weitere Annotationen durchgeführt. Auch hier gibt es wieder eine breite Auswahl an Werkzeugen, deren Verwendung von projektspezifischen Anforderungen und Benutzergruppen abhängig ist. Einige sind für spezielle Forschungsbereiche konzipiert, wie z. B. LaKomp, andere bieten eine grafische Oberfläche für Editor*innen ohne tiefgreifende XML/TEI-Kenntnisse (CATMA). Wieder andere kombinieren eine Benutzeroberfläche mit bestimmten Funktionalitäten, wie z. B. einer Registerfunktion (ediarum). In einigen Anwendungsfällen bietet es sich darüber hinaus an, reines XML/TEI im Oxygen XML Editor zu schreiben.

Digitale Editionen produzieren Forschungsdaten und machen diese im Idealfall unter Einhaltung der FAIR-Prinzipien zugänglich. Dies erfordert die Einbindung von Normdaten oder kontrollierten Vokabularen. Werkzeuge hierfür können OpenRefine zur halbautomatischen Verknüpfung von Entitäten sein, oder Tools wie ba[sic]; stärker datenzentrierte Projekte verwenden Tools wie Fast Cat.

Die Veröffentlichung und Langzeitarchivierung kann schlussendlich über Repositories und Tools wie GAMS, ARCHE, teiPublisher oder ediarum.Web erfolgen. Für manchen Editionsprojekten ist es sinnvoll, domänenspezifische Werkzeuge oder APIs, wie z. B. correspSearch für Korrespondenzen, zu verwenden.

Ziel des Projekts *Digital Edition Creation Pipelines: Tools and Transitions (DigEdTnT)* ist es, Best-Practice-Pipelines und Tutorials für ausgewählte Tools und deren Übergänge (=Transitions) zu erstellen, die bei der Wahl der Tools und der Arbeit mit Tools zur Erstellung digitaler Editionen helfen sollen. Denn an den Übergängen ergeben sich mitunter besondere Herausforderungen, wenn beispielsweise Ergebnisse aus Transkribus nach ediarum zur weiteren Annotation überführt werden sollen. Das Projekt setzt dabei insbesondere auf eine Community-basierte Auseinandersetzung mit vorhandenen Tools. Daher sollen in zwei Workshops Nutzer*innen und Entwickler*innen zusammengeführt werden, um Anwendungsfälle und Feedback gemeinsam zu diskutieren. Wie die Übergänge zwischen einzelnen Tools abgewickelt werden können, wird letztlich in Tutorials und Guidelines sowie in Code Snippets beschrieben, die wiederum in einschlägigen Kontexten (KONDE Weißbuch, DARIAH Campus, etc.) zugänglich gemacht werden.

Das eingereichte Poster soll zum Projektstart von DigEdTnT die Diskussion zu diesem Thema eröffnen und interessierte Kolleg*innen sowie auch Toolentwickler*innen adressieren.

Bibliographie

Fafalios, Pavlos and Kostas Petrakis, Georgios Samaritakis, et. al. "FAST CAT: Collaborative Data Entry

and Curation for Semantic Interoperability in Digital Humanities". In *Journal on Computing and Cultural Heritage (JOCCH)*. 14, 4, Article 45 (2021), 1-20. <https://doi.org/10.1145/3461460>.

Fechner, Martin. "Eine nachhaltige Präsentationsschicht für digitale Editionen." DHd, (2018). [urn:nbn:de:kobv:b4-opus4-33277](https://nbn-resolving.org/urn:nbn:de:kobv:b4-opus4-33277).

Fritze, Christiane. "Wohin mit der digitalen Edition?". In *Bibliothek Forschung und Praxis* edited by Achim Bonte et. al., 43(3), (2019), 432-440.

Holstein, T., Störl, U. "Towards Supporting Tools for Editors of Digital Scholarly Editions for Correspondences". In *HCI International 2020 – Late Breaking Posters HCII 2020. Communications in Computer and Information Science* edited by C. Stephanidis, M. Antona, S. Ntoa., vol 1293. Springer, Cham. https://doi.org/10.1007/978-3-030-60700-5_25.

Horstmann, Jan. "Undogmatic Literary Annotation with CATMA" In *Annotations in Scholarly Editions and Research: Functions, Differentiation, Systematization* edited by Julia Nantke and Frederik Schlupkoth, 157-176. Berlin, Boston: De Gruyter, 2020. <https://doi.org/10.1515/9783110689112-008>.

Klug, Helmut W. and Selina Galka and Elisabeth Steiner. "KONDE Weißbuch im HRSM Projekt 'Kompetenznetzwerk Digitale Edition'". <https://www.digitale-edition.at>.

RIDE – A review journal for digital editions and resources, <https://ride.i-d-e.de>, 20.6.2022.

Sahle, Patrick. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. Schriften des Instituts für Dokumentologie*, 2013. <https://kups.ub.uni-koeln.de/5352>.

Söring, Sibylle. Technische und infrastrukturelle Lösungen für digitale Editionen: DARIAH-DE und TextGrid. In *Bibliothek Forschung und Praxis* edited by Achim Bonte et. al. 40(2), (2016), 207-212. <https://doi.org/10.1515/bfp-2016-0040>.

Digitale Editionen von historischen Reiseberichten öffnen: Open Text und Open Data mit einheitlicher Textauszeichnung, semantischer Annotation und ontologiebasierter Datenmodellierung

Balck, Sandra

balck@ios-regensburg.de
IOS Regensburg, Deutschland

Frank, Ingo

frank@ios-regensburg.de
IOS Regensburg, Deutschland

Einleitung

Innerhalb der Literatur- und Kulturwissenschaften ist die historische Reiseforschung ein beliebtes Forschungsfeld, dennoch gibt es innerhalb der Digital Humanities wenige nennenswerte Fortschritte bei der Erschließung, Verarbeitung und Visualisierung von Reiseberichten. Konventionellen digitalen Editionen fehlt bisher die notwendige Ausdruckskraft und Flexibilität, um die diversen Anwendungsfälle und Forschungsfragen der historischen Reiseforschung zu erschließen. Anstatt einer starren Auszeichnung mit TEI zu folgen, müssen Informationen in digitalen Editionen erkannt, identifiziert, mit zusätzlichen Daten angereichert und Narrationen der Ereignisse explizit modelliert werden.¹

Dieser Beitrag beschäftigt sich mit der Frage der Öffnung digitaler Reiseberichte für die wissenschaftliche Analyse und Visualisierung (von Zeit, Raum, Ereignissen u.a.) durch Textauszeichnung, semantische Annotation und ontologiebasierte Modellierung. Hierbei verfolgen wir einen disziplinübergreifenden, iterativen Ansatz, welcher sowohl geistes- als auch informationswissenschaftliche Perspektiven einbezieht.² Erprobt wird dieser Ansatz an der digitalen Edition des Reiseberichts Franz Xaver Bronners (1758–1850), der 1810 als Professor für theoretische Physik von Aarau in der Schweiz an die russische Universität Kasan an der Wolga ging und 1817 in die Schweiz zurückkehrte.

Problemstellung

Datenmodellierung kann laut Flanders und Jannidis (vgl. 2015) in zwei Gruppen unterschieden werden: Curation-Driven und Research-Driven. Curation-driven beschreibt die Praktiken von Bibliotheken und Archiven, Objekte mit Hilfe von Standards einheitlich zu erfassen, um so die Auffindbarkeit und Transparenz zu gewährleisten. Die dafür notwendige Reduktion führt jedoch zu Ungenauigkeiten und Lücken im Datenmaterial. Research-Driven hingegen zielt auf die Beantwortung spezifischer Forschungsfragen ab und die Datenerfassung/Modellierung folgt einem konkreten Forschungsinteresse. Dabei werden nur selten Standards berücksichtigt. Dieser Gegensatz zwischen Forschungs- und Kuratierungspraxis erschwert die Kompatibilität und damit die Vergleichbarkeit der Daten. Um digitale Reiseberichte für wissenschaftliche Analysen verwertbar zu machen, müssen sie beiden Ansprüchen gerecht werden.

Im Bereich der digitalen Editionen haben sich die TEI-Guidelines zum De-facto-Standard entwickelt. "[TEI] is the most systematic effort so far to create standards for scholarly memory in an evolving digital culture." (teic.org 2019) Die Guidelines beinhalten aktuell 585 Elemente und sind flexibel gestaltet, um für ein breites Spektrum von Forschungsfragen anwendbar zu sein. Ein Problem ist jedoch, dass dieselbe Information unterschiedlich kodiert und interpretiert werden kann (z. B. <rs>, <persName>) und der „Standard“ damit nicht zwingend interoperabel ist (siehe Unsworth 2011; Burrows et al. 2021; Giovannetti und Tomasi 2022). Die TEI versucht die Brücke zwischen Standardisierung und Granularität zu schlagen, verliert damit aber ihre Eindeutigkeit (vgl. Kudella und Jefferies 2019).

Textauszeichnung und Ontologie-Entwurfsmuster

Einen Lösungsansatz für das Interoperabilitätsproblem auf Textseite bietet das DTA-Basisformat (Haaf et al. 2015), welches sich zum Ziel gesetzt hat: "[...] eine umfassende Textaufbereitung [zu] ermöglichen und dabei gleichzeitig Variationsspielräume bei der Annotation so ein[schränken], dass die Kohärenz [...] untereinander gewährleistet wird." (DTABf 2011-2020) Das DTABf richtet sich dabei nach den P5-Richtlinien der TEI. Um darüber hinaus auch spezifische Forschungsinteressen der historischen Reiseforschung zu adressieren, entwickeln wir auf Datenseite einen ontologiebasierten Textanreicherungs- und Bearbeitungsworkflow: Ontologie-Entwurfsmuster werden iterativ aufgebaut und für die Klassifizierung von Reise(teil)ereignissen (Abreise, Ankunft usw.) und Reisebeobachtungen (z. B. besuchte öffentliche Orte, Gewohnheiten von Personen) angewandt.

Während TEI für die Textauszeichnung verwendet wird, dient CRM zur Anreicherung des Textes mit explizitem Wissen, welches in einer Datenbank gespeichert wird. Wir modellieren mit den Ontologie-Entwurfsmustern nicht die Erzählung als solche³, sondern den rekonstruierten und stellenweise interpretierten Reiseverlauf als Repräsentation der Realität. Aus narratologischer Sicht

machen wir mit dem ereigniszentrierten Modellierungsansatz von CRM also nur die Fabula (chronologische Reihenfolge der Ereignisse) eines Reiseberichts explizit. Das Sujet (Erzählreihenfolge) kann allerdings bei Bedarf anhand der annotierten Textstellen abgefragt und rekonstruiert werden.⁴

Zur Erstellung des Annotationsschemas und der damit verbundenen Ontologie-Entwurfsmuster wenden wir die Frame-Semantik als theoretischen Rahmen an. Frames können als n-äre Relationen⁵ repräsentiert und daher zur Entwicklung von Ontologie-Entwurfsmustern für Reiseereignisse und -beobachtungen verwendet und darüber hinaus als "knowledge patterns" zur Validierung der Entwurfsmuster herangezogen werden (vgl. Presutti et al. 2012).⁶ Die Ontology Design Patterns dienen uns als „Schablonen“ zum Anlegen an den Text – wobei deren Orientierung an den Frames sehr hilfreich ist – um Reisedaten, Beobachtungen und Tätigkeiten unterwegs und während Zwischenstopps zu erfassen. Im ständigen Austausch mit der Bearbeitung von Forschungsfragen am Text werden die Frames und Design Patterns laufend überprüft und angepasst. Diese Form der forschungsgeleiteten Standardisierung (mittels expliziter und einheitlicher Modellierung) macht digitale historische Reiseberichte interoperabel und öffnet sie damit für vergleichende Analysen. Die möglichen Ansätze zur Verknüpfung von TEI-kodiertem Text und RDF-Daten mittels semantischer Annotation evaluieren wir (vgl. Eide 2015 und Borriello et al. 2016) und stellen im Poster Lösungswege mit EARMARK (Barabucci et al. 2013) und NIF (Hellmann et al. 2013) vor.

Verwandte Arbeiten und Schlussfolgerung

Es gibt einige Beispiele wie das Hespont-Projekt (Mambrini 2016) oder die Semantic Blumenbach-Edition (Wettlaufer 2015), die TEI-kodierten Text und CRM-modellierte Daten miteinander verknüpfen. Unser Ansatz geht jedoch über die bestehenden Projekte hinaus, da wir Ontologie-Entwurfsmuster für eine explizitere Datenmodellierung entwickeln und anwenden.⁷ Kurz gesagt, wir lösen die Interoperabilitäts- und Ausdrucksprobleme von TEI und CRM mit Hilfe von DTABf und Frame-basierten Ontologie-Entwurfsmustern, was wiederum die Kategorien von Reiseereignissen und Reisebeobachtungen in historischen Reiseberichten für die weitere Analyse und Visualisierung explizit macht.

Fußnoten

1. Bei der Erstellung der digitalen Edition folgen wir dem Historical Information Life Cycle (Meroño-Peñuela et al. 2014), wobei wir Ontologien nicht nur in der Anreicherungs-, Bearbeitungs- und Retrieval-Phase des Lebenszyklus, sondern auch in der Analyse- und Visualisierungsphase einsetzen.
2. Unser Ontologie-Entwurfsansatz orientiert sich an der eXtreme Design-Vorgehensweise (Presutti et al. 2009), bei der sog. Competency Questions aus anfäng-

lichen User Stories abgeleitet werden, um die Anforderungen an die Datenmodellierung und das Information Retrieval zu definieren.

3. siehe hierzu den Modellierungsansatz von Bartalesi et al. 2017

4. Das kann interessant für Analyse und Visualisierung sein, weil, wie Maurer (2015, S. 391 f.) anmerkt, Reiseberichte von wissenschaftlichen Forschern oft einer thematischen Organisation folgen, um „Reiseergebnisse“ zu präsentieren, anstatt einfach einer chronologischen Reihenfolge zu folgen.

5. Annotationswerkzeuge wie brat (Stenetorp et al. 2012) oder INCEpTION (Klie et al. 2018) bieten Annotations- und Visualisierungskomponenten für n-äre Relationen (Frames oder Ereignisse inkl. der Rollen von Akteuren).

6. Siehe dazu etwa das allgemeine Frame *Travel* mit Frame-spezifischen Eigenschaften zur Beschreibung von reisender Person, Reiseziel, Verkehrsmittel usw. in FrameNet: <http://framenet.lexicalsemantics.org/frameIndex.xml?frame=Travel>

7. Die Datenmodelle GeoJSON-T, Linked Places und Linked Traces (siehe Grossner et al. 2017) sind bereits gut etabliert, aber es handelt sich dabei nur um Formate für den Datenaustausch und kommen daher in unserem Projekt nicht für den Aufbau der Datenbank in Frage.

Bibliographie

Barabucci, Gioele, Angelo Di Iorio, Silvio Peroni, Francesco Poggi, and Fabio Vitali. 2013. “Annotations with Earmark in Practice: A Fairy Tale.” In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*. DH-Case '13. Association for Computing Machinery, 10.1145/2517978.2517990

Bartalesi, Valentina, Carlo Meghini, and Daniele Metilli. 2017. “A Conceptualisation of Narratives and Its Expression in the Crm.” In *International Journal of Metadata, Semantics and Ontologies* 12 (1): 35–46.

Burrows, Toby, Matthew Holford, David Lewis, Andrew Morrison, Kevin Page, and Athanasios Velios. 2021. “Transforming Tei Manuscript Descriptions into Rdf Graphs.” In *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, 143–54. Nordstedt: BoD, <http://www.digitalhumanities.org/dhq/vol/16/2/000605/000605.html>

Eide, Øyvind. 2015. “Ontologies, Data Modeling, and TEI.” In *Journal of the Text Encoding Initiative*, no. 8 (December), <https://doi.org/10.4000/jtei.1191>

Füssel, Marian, Tim Neu. 2021. “Akteur-Netzwerk-Theorie und Geschichtswissenschaft”. Paderborn: Brill; Ferdinand Schöningh.

Giovannetti, Francesca; Tomasi, Francesca. 2022. “Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation” In *Digital Humanities Quarterly* 16 (2). <http://www.digitalhumanities.org/dhq/vol/16/2/000605/000605.html>.

Grossner, Karl, Merrick Lex Berman, and Rainer Simon. 2017. “Linked Places: A Modeling Pattern and Software for Representing Historical Movement.” In *Digital*

Humanities 2017: Conference Abstracts, 463–65, <https://dh2017.adho.org/abstracts/204/204.pdf>

Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2015. “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*, no. 8, doi:10.4000/jtei.1114

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. “Integrating NLP using Linked Data”. In: 12th International Semantic Web Conference, 21–25 October 2013, Sydney, Australia.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation”. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA. <https://aclanthology.org/C18-2002>

Kudella, Christoph, and Neil Jefferies. 2019. “How Do We Model the Republic of Letters?” In *Reassembling the Republic of Letters in the Digital Age*, edited by Howard Hotson and Thomas Wallnig, 41–53. Göttingen: Göttingen University Press, 10.17875/gup2019-1146

Mambrini, Francesco. 2016. “Treebanking in the World of Thucydides. Linguistic Annotation for the Hespertion Project.” In *Digital Humanities Quarterly* 10 (2), <http://www.digitalhumanities.org/dhq/vol/10/2/000251/000251.html>

Maurer, Michael. 2015. “Reiseberichte als Wissensspeicher.” In *Wissenspeicher Der Frühen Neuzeit: Formen und Funktionen*, edited by Frank Grunert and Anette Syndikus, 391–412. Berlin, Boston: De Gruyter, 10.1515/9783050086637-015

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leene Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank van Harmelen. 2015. “Semantic Technologies for Historical Research: A Survey.” *Semantic Web* 6 (6). IOS Press: 539–64, 10.3233/SW-140158

Presutti, Valentina, Enrico Daga, Aldo Gangemi, and Eva Blomqvist. 2009. “eXtreme Design with Content Ontology Design Patterns.” WOP.

Presutti, Valentina, Eva Blomqvist, Enrico Daga, and Aldo Gangemi. 2012. “Pattern-Based Ontology Design.” In *Ontology Engineering in a Networked World*, edited by Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, 35–64. Berlin: Springer.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. “Brat: a Web-based Tool for NLP-Assisted Text Annotation”. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107.

tei-c.org. 2019. “A very gentle introduction to the TEI markup language.” <https://tei-c.org/Vault/Tutorials/mueller-index.htm>.

Unsworth, John. 2011. “Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the Tei.” In *Journal of the Text Encoding Initiative*, no. 1 (June), 10.4000/jtei.215

Wettlaufer, Jörg, Christopher Johnson, Martin Scholz, Mark Fichtner, and Sree Ganesh Thotempudi. 2015. “Semantic Blumenbach: Exploration of Text-Object Relati-

onships with Semantic Web Technology in the History of Science.” In *Digital Scholarship in the Humanities* 30 (suppl_1): i187–i198, 10.1093/llc/fqv047

Digitale Interaktion auf Augenhöhe – drei Wege zu partizipativer Forschung und FAIRer Lehre an der UB Kiel

Christ, Andreas

christ@ub.uni-kiel.de

Christian-Albrechts-Universität zu Kiel, Deutschland

Diebel, Richard

diebel@ub.uni-kiel.de

Christian-Albrechts-Universität zu Kiel, Deutschland

Henzel, Katrin

henzel@ub.uni-kiel.de

Christian-Albrechts-Universität zu Kiel, Deutschland

Petersen, Britta

b.petersen@rz.uni-kiel.de

Christian-Albrechts-Universität zu Kiel, Deutschland

Vetter, Angila

vetter@ub.uni-kiel.de

Christian-Albrechts-Universität zu Kiel, Deutschland

Ausgangslage

Das Selbstverständnis wissenschaftlicher Bibliotheken hat sich grundlegend gewandelt (z.B. Auberer et al. 2022): Neben ‚traditionellen‘ Aufgaben im Bewahren, Digitalisieren, Erschließen und dauerhaften Bereitstellen offener Kulturdaten übernehmen insbesondere Universitätsbibliotheken zunehmend Aufgaben im Forschungsdatenmanagement, stellen Arbeits- und Infrastrukturen inklusive Schulungs- und Beratungsdienste für DH-Projekte bereit. Ihnen kommt damit „eine wichtige Rolle bei der Sicherung guter wissenschaftlicher Praxis durch professionellen Umgang mit Forschungsdaten“ zu (Rösch 2021, 136). Hierin wird die *vermittelnde* Rolle von Bibliotheken deutlich, gerade auch im Umgang mit Standards für Open Data. Der von den Bibliotheken selbst wie auch an sie herangetragene Anspruch an Offenheit (Berg-Weiß et al. 2022) ist jedoch nicht garantiert, vielmehr zeigt sich aufgrund der ausgesprochen heterogenen Ausgangs- und Rahmenbedingungen für Open

Science (vgl. UNESCO 2021, 6) erst in konkreten Anwendungsfällen, ob der Zugang zu Kulturdaten, ihre Erstellung und (Nach-)Nutzung gemäß rechtlicher Vorgaben tatsächlich offen sind.

Ein solches Anwendungsszenarium ist mit der FAIRen Produktion und Nachnutzung von OER im Kontext der Lehre Gegenstand dieses Posters. Als Beispiel dient ein an der UB Kiel angesiedeltes Projekt, welches wiederum in ein partizipatives Forschungsdatenmanagement eingebettet ist, das mit Inklusion und Citizen Science zwei weitere Schwerpunkte bildet. Alle drei Bereiche teilen dabei die Grundidee der digitalen Interaktion auf Augenhöhe, die essentiell für das Gelingen partizipativer Forschung und FAIRer Lehre ist.

FAIRe Open Educational Resources

Freier Zugang zu und niedrigschwellige Nachnutzung von Lern- und Lehrmaterialien gehören zu einer offenen Wissenschaftspraxis. Open Educational Resources (OER) „kommt eine wichtige Funktion bei einem chancengerechten Wandel in der Bildung [...] zu“ (BMBF 2022, 2). Informations- und Kommunikationstechnologien bieten dabei, wie die UNESCO in ihrer Empfehlung zu OER betont, „ein hohes Potenzial für effektiven, chancengerechten und inklusiven Zugang zu OER und deren Weiterverwendung, Bearbeitung und Weiterverbreitung“ (UNESCO 2019, I.3). Dieses Potenzial umzusetzen ist jedoch anspruchsvoll, da neben der Verwendung offener Lizenzen und Dateiformate auch automatische Auffindbarkeit und didaktische Kontextualisierung gewährleistet werden müssen (Twillo o.J.). Herkömmliche Lernmanagement-Systeme an Hochschulen eignen sich für die Produktion FAIRer OER nur sehr bedingt (Dietrich, Zug 2020). Im zentralen Forschungsdatenmanagement der CAU kommt daher LiaScript – ein erweiterbarer, freier Markdown-Dialekt (Dietrich 2022) – zum Einsatz. Er wurde für die Erstellung digitaler Lern- und Lehrressourcen entwickelt und gewährleistet einfache Editier- und Versionierbarkeit der Materialien mittels Open-Source-Software ohne den Einsatz proprietärer Autor:innensysteme. Durch die Verwendung von Markdown bleiben die Materialien formatunabhängig und können für verschiedene Anwendungsfälle in passende Formate (u.a. HTML, PDF, SCORM für Lernmanagement-Systeme) exportiert werden. In zwei Kieler Pilotprojekten werden modulare LiaScript-Bausteine zu Digital-Literacy-Inhalten gemeinsam mit den Fachwissenschaften entwickelt, welche die didaktische Strukturierung und fachliche Einbettung sichern. Ein von der UB Kiel mit dem Germanistischen und Historischen Seminar der CAU Kiel durchgeführtes Projekt hat die Integration dieser Bausteine in bestehende curriculare Lehrveranstaltungen zum Ziel. Hierbei spielen organisatorisch-infrastrukturelle, fachliche, didaktische und technische Aspekte eine Rolle, die zusammen mit ersten Evaluationsergebnissen zum Einsatz des Tools in Lehrveranstaltungen diskutiert werden.

Einbindung in ein Gesamtkonzept eines partizipativen und inklusiven Forschungsdatenmanagements

Eng verbunden mit der Forderung nach offener und freier Bildung sind die Themen Partizipation der Zivilgesellschaft mittels Citizen Science und eine inklusiv gedachte und praktizierte Wissenschaft als Bestandteile einer offenen Wissenschaftskultur. Universitätsbibliotheken bieten aufgrund ihrer bestehenden Erfahrung und Infrastruktur die Möglichkeit, partizipative, trans- und interdisziplinäre Forschungsprojekte anzustoßen und zu begleiten, zur Vernetzung zwischen Bürger:innen und der Scientific Community beizutragen, Citizen-Science-Projekte sichtbar zu machen und zu beraten (Vohland et al. 2021, 114; siehe auch Wiederkehr 2021). Die Transkription von Texten, Georeferenzierung von Objekten und Kategorisierung von kulturellen Artefakten bietet, etwa über eine App, einen niederschweligen Zugang, um Bürger:innen an Forschungsprozessen zu beteiligen (vgl. Studie zu 'Public Participation in Scientific Research', Bonney et al. 2009) und sie, nicht zuletzt mittels OER, beim Erwerb von Datenkompetenzen zu unterstützen.

Offenheit und Teilhabe sind wesentliche Forderungen der FAIR-Prinzipien. Wie aber gestalten sich FAIRe Daten für und barrierearmes Arbeiten in inklusiven Forschungsgruppen? Versteht man Behinderung als komplexes Phänomen und als Interaktion zwischen Individuum und Umwelt (UNESCO 2022, Abs. 1f.), liegt die Vermutung nahe, es könne keine allgemeingültigen Standards für einen inklusiven Umgang mit Forschungsdaten geben. Doch zeigen z.B. die Regelungen im Webdesign (nach WCAG 2.1 und BITV 2.0), dass sich die Rahmenbedingungen sehr wohl ändern und Barrieren abbauen lassen. Für das Forschungsdatenmanagement soll Ähnliches erreicht werden, hier gilt es den Fokus über die Präsentation von Daten hinaus um Datenerhebung, -analyse und -archivierung zu erweitern.

Bibliographie

Auberer, Benjamin, Alexander Berg-Weiß, Vanessa Gabriel und Martin Spenger. 2022. „Potentiale nutzen und Verbindungen herstellen. Neue fachliche Aufgabenbereiche für Bibliotheken am Beispiel Forschungsdatenmanagement.“ *O-Bib* 9 (2), 1–16. 10.5282/o-bib/5783.

Berg-Weiß, Alexander, Sibylle Hermann, Miriam Kötter, Caroline Leiß, Christoph Müller und Annette Strauch-Davey. 2022. „Openness in Bibliotheken. Positionspapier der Kommission für Forschungsnahe Dienste des VDB.“ *O-Bib* 9 (2):1–4. 10.5282/o-bib/5826.

Bundesministerium für Bildung und Forschung (BMBF). 2022. *OER-Strategie. Freie Bildungsmaterialien für die Entwicklung digitaler Bildung*. Berlin: BMBF. https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/3/691288_OER-Strategie.html (zugegriffen: 14. Dezember 2022).

Dietrich, André und Sebastian Zug. 2020. *From Hero to Zero with Learning Management Systems. Why and How LMSs fail in distributing knowledge*. <https://aizac.herokuapp.com/from-hero-to-zero-with-learning-management-systems> (zugegriffen: 28. Juli 2022).

Dietrich, André. 2022. *Share your knowledge and build online courses with simple Markdown!* <https://liascript.github.io> (zugegriffen: 01. August 2022).

Göbel, Claudia, Justus Henke und Sylvi Mauermeister. 2020. *Kultur und Gesellschaft gemeinsam erschaffen. Überblick und Handlungsoptionen zu Citizen Science in den Geistes- und Sozialwissenschaften*, unter Mitarbeit von Susann Hippler, Nicola Gabriel und Steffen Zierold, Institut für Hochschulforschung (HoF) an der Martin - Luther - Universität, Halle - Wittenberg. <https://www.hof.uni-halle.de/web/dateien/pdf/HoF-Handreichungen14.pdf> (zugegriffen: 14. Dezember 2022).

Heinisch, Barbara, Kristin Oswald, Maik Weißpflug et al. 2021. „Citizen Humanities.“ In *The Science of Citizen Science*, hg. von Kathrin Vohland, Anne Land-Zandstra, Luigi Ceccaroni et al., Cham: Springer, 97–118. 10.1007/978-3-030-58278-4_6.

Bonney, Rick et al. 2009. *Public participation in scientific research: Defining the field and assessing its potential for informal science education: A CAISE inquiry group report*. <https://files.eric.ed.gov/fulltext/ED519688.pdf> (zugegriffen: 14. Dezember 2022).

Rösch, Hermann. 2021. „Forschungsethik und Forschungsdaten.“ In *Praxishandbuch Forschungsdatenmanagement*, hg. von Markus Putnings, Heike Neuroth und Janna Neumann, Berlin, Boston: De Gruyter Saur, 115–140. 10.1515/9783110657807.

Twillo. o.J. *Digitaler Leitfaden Teilen von Bildungsmaterialien*. <https://twillo-lehre-teilen.github.io/leitfaden-oer-workshop/#/> (zugegriffen: 01. August 2022).

Wiederkehr, Stefan. 2021. „Citizen Science: Eine Chance für wissenschaftliche Bibliotheken.“ *O-Bib* 8 (4), 1–13. 10.5282/o-bib/5727.

UNESCO. 2019. *UNESCO Empfehlung zu Open Educational Resources*. Paris: UNESCO. https://www.unesco.de/sites/default/files/2020-05/2019_Empfehlung%20Open%20Educational%20Resources.pdf (zugegriffen: 14. Dezember 2022).

UNESCO. 2021. *UNESCO Recommendation on Open Science*. Paris: UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en> (zugegriffen: 14. Dezember 2022).

UNESCO Institute for Statistics. 2022. „Disability“, in *UIS Glossary*. <https://uis.unesco.org/en/glossary-term/disability> (zugegriffen: 14. Dezember 2022).

Digitale Methoden kritisch reflektieren – Die Erweiterung des Werkzeugkastens der Historiker:innen

Althage, Melanie

melanie.althage@hu-berlin.de

Humboldt-Universität zu Berlin, Deutschland

Historiker:innen setzen sich zur Erforschung der Geschichte mit einer Vielzahl ganz unterschiedlicher Quellenarten auseinander. Um das historische Quellengut entsprechend der individuellen Fragestellungen angemessen und kritisch-reflektiert zu bearbeiten, wurden mit dem historischen Werkzeugkasten eine Reihe von Methoden etabliert, die dieser Vielfalt Rechnung tragen (Überblick: von Brandt 18. Aufl. 2012). Die konkreten Verfahren müssen dabei nicht spezifisch geschichtswissenschaftlich sein, sondern können zum Teil auch aus anderen Wissenschaften entstammen, etwa den Philologien oder Wirtschaftswissenschaften. Entscheidend für die Auswahl des jeweiligen methodischen Vorgehens ist die konkrete Fragestellung in Kombination mit dem Quellenkorpus sowie die Befolgung methodischer Grundsätze, die auf die Plausibilität der Darstellung historischer Wirklich- und Wahrscheinlichkeiten zielen. Insofern wir als Historiker:innen also abhängig von unseren Fragestellungen das Untersuchungsinstrumentarium immer neu bestimmen, ist die Kritik der jeweiligen Methodik unverzichtbar (Sellin 2008, 84-96). Sie zielt darauf, die der Methode impliziten Einschränkungen zu verdeutlichen und bewusst zu machen. Erst auf dieser Basis ist dann eine angemessene Interpretation der gewonnenen Ergebnisse möglich, die wiederum Grundlage historiographischer Erzählungen ist.

Mit dem „digitalen Zeitalter“ und den Digital Humanities kamen vor allem seit den 2000er-Jahren durch die digitalen Methoden innovative Möglichkeiten zur Quellenlektüre und -auswertung für die Geschichtswissenschaften hinzu. Damit stehen nunmehr Werkzeuge zur Verfügung, die einerseits vormalige analoge Tätigkeiten digital abbilden und unterstützen, wie Data Mining für die Historische Statistik, andererseits bieten quantitative Verfahren des Maschinellen Lernens wie etwa Topic Modeling mit ihrem explorativen Modellierungsansatz neuartige Ansätze, indem Texte als Daten verstanden und entsprechend flexibel skalierbar in Masse ausgewertet werden können. Für die Geschichtswissenschaften ist es entscheidend, sich mit den epistemologischen und methodologischen Konsequenzen dieser Digitalität in Bezug auf ihre Quellen und Methoden auseinanderzusetzen. Denn auch die digitalen Methoden wurden in anderen Disziplinen mit je eigenen theoretischen und methodologischen Annahmen respektive Erkenntnisinteressen entwickelt. Sie sind daher nicht ohne Weiteres auf historische Anwendungsfälle übertragbar. Um sie

dennoch produktiv in den "Werkzeugkasten" der Historiker:innen zu integrieren, ist daher zunächst die "Kluft" zwischen historischer Fachdisziplin und fachfremder Methode zu identifizieren und durch geeignete Strategien zu überwinden.

Diesem Anliegen widmet sich das diesem Poster zugrunde liegende Dissertationsprojekt "Mining the Historian's Web – Methodenkritische Reflexion quantitativer Verfahren zur Analyse genuin digitaler Quellen am Beispiel der historischen Fachkommunikation". Zwar wird in jüngerer Zeit zunehmend untersucht, welche Implikationen mit digitalen Methoden für die Arbeit mit historischen Quellen sowie für den Erkenntnisbildungsprozess einhergehen (u.a. Hiltmann et al. 2021; Fickers 2020; König 2017; Braake et al. 2016; Wettlaufer 2016), allerdings fehlt es weitestgehend noch an einer systematischen Werkzeug- und Methodenkritik, die den verantwortungsvollen Umgang mit digitalen Methoden in den Geschichtswissenschaften begleiten muss. In Anlehnung an Diskussionen rund um *Tool-* und *Algorithmic Criticism* (Es/Schäfer/Wieringa 2021; Dobson 2019; Ramsay 2011) ist es Ziel der Dissertation, diese Lücke zu schließen. Dazu werden vergleichend etablierte Methodenkomplexe zunächst theoretisch-konzeptionell unter Berücksichtigung ihrer Entwicklungsgeschichte erarbeitet und anhand exemplarischer historischer Fragestellungen praktisch angewendet. Anschließend werden vor dem Hintergrund der besonderen Charakteristika historischer Fragestellungen und Daten die Erkenntnispotenziale und -grenzen kritisch geprüft. Auf dieser Basis wird dann abstrahierend ein Kriterien- und Fragenkatalog für die methodenkritische Evaluation und Auswahl digitaler Methoden entwickelt sowie konkrete Anwendungsempfehlungen gegeben.

Das Poster wird erste anwendungsbezogene Erkenntnisse am Beispiel von Topic Modeling vorstellen. In den digitalen Geistes- und Geschichtswissenschaften ist die 2003 vorgestellte *Latent Dirichlet Allocation* (LDA, Blei/Ng/Jordan 2003) hierfür am populärsten. Sie wird eingesetzt, um über die Identifikation statistisch signifikanter Sprachgebrauchsmuster in umfangreichen Textsammlungen beispielsweise die Entwicklung von Publikationstrends zu untersuchen (exemplarisch: Wehrheim 2019). Der frequente Einsatz von LDA scheint vor allem auf der hohen Verfügbarkeit zu basieren¹ und weniger auf einer Evaluation der Eignung im Vergleich zu anderen Ansätzen. Für historische Fragestellungen etwa, die insbesondere die Temporalität und Kontextgebundenheit der Quellen fokussieren, weist LDA einige Limitierungen auf: Die Topic-Modellierung, deren vorrangiger Zweck die maschinelle Klassifikation umfangreicher und unstrukturierter Daten ist, berücksichtigt weder Relationen zwischen den Topics noch die Historizität der Daten. Diese im Algorithmus inhärenten Einschränkungen waren vielfach Anlass für technische Weiterentwicklungen (überblickshaft: Chauhan/Shah 2021; Vayansky/Kumar 2020); der Stand der Methodenentwicklung wird in den digitalen Geistes- und Geschichtswissenschaften allerdings bislang kaum rezipiert. Hier werden vor allem die grundsätzlichen Herausforderungen und Konsequenzen diskutiert, die mit LDA einhergehen und entsprechende Workflows zur nachhaltigen Integration in den Forschungsprozess vorgeschlagen (u.a. Hodel/Möbus/Serif 2022; Ugianova/Gius 2020; Maier et al. 2018;

Fechner/Weiß 2017; Andorfer 2017). Da die Wahl des konkreten Modellierungsverfahrens aber darüber entscheidet, welche Aussagen sich über die sprachliche Struktur einer Dokumentensammlung treffen lassen, soll das Poster durch eine methodenkritische Bestandsaufnahme auch andere Topic-Modeling-Algorithmen für die historische Forschung präsentieren.

Fußnoten

1. LDA ist in zahlreichen etablierten Programmbibliotheken und gebrauchsfertigen Werkzeugen implementiert, siehe etwa MALLET (<http://mallet.cs.umass.edu/topics.php>), Gensim (<https://radimrehurek.com/gensim/>) oder DARIAH-DE TopicsExplorer (<https://dariah-de.github.io/TopicsExplorer/>).

Bibliographie

- Andorfer, Peter. 2017. "Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich." *Zeitschrift für digitale Geisteswissenschaften* 2 10.17175/2017_002.
- Braake, Serge ter, Antje Fokkens, Niels Ockeloen und Chantal van Son. 2016. "Digital History: Towards New Methodologies." In *Computational History and Data-Driven Humanities*, hg. von Bojan Bozic, Gavin Mendel-Gleason, Christophe Debruyne und Declan O'Sullivan, 23–32. Cham: Springer.
- Brandt, Ahasver von. 2012. *Werkzeug des Historikers. Eine Einführung in die historischen Hilfswissenschaften*. Stuttgart: Kohlhammer 18. Aufl.
- Blei, David M., Andrew Y. Ng und Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.
- Chauhan, Uttam und Apurva Shah. 2021. "Topic Modeling Using Latent Dirichlet allocation: A Survey." *ACM Computing Surveys* 54, 7 10.1145/3462478.
- Dobson, James E. 2019. *Critical Digital Humanities. The Search for a Methodology*. Urbana (Illinois): University of Illinois Press.
- Es, Karin van, Mirko T. Schäfer und Maranke Wieringa. 2021. "Tool Criticism and the Computational Turn. A "Methodological Moment" in Media and Communication Studies." *Medien & Kommunikationswissenschaft* 69, 1: 46–64.
- Fechner, Martin und Andreas Weiß. 2017. "Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts." *Zeitschrift für digitale Geisteswissenschaften* 2 10.17175/2017_005.
- Fickers, Andreas. 2020. "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?" *Zeithistorische Forschungen/Studies in Contemporary History* 17, 1: 157–168.
- Hiltmann, Torsten, Jan Keupp, Melanie Althage und Philipp Schneider. 2021. "Digital Methods in Practice. The Epistemological Implications of Applying Text Re-Use Analysis to the Bloody Accounts of the Conquest of Jerusalem (1099)." *Geschichte und Gesellschaft* 46, 1: 122–156 10.13109/gege.2021.47.1.122.
- Hodel, Tobias, Dennis Möbus und Ina Serif. 2022. "Von Inferenzen und Differenzen. Ein Vergleich von Topic-Modeling-Engines auf Grundlage historischer Korpora." In *Von Menschen und Maschinen: Mensch-Maschine-Interaktionen in digitalen Kulturen*, hg. von Selin Gerlek, Sarah Kissler, Thorben Mämecke und Dennis Möbus, 181–205. Hagen: Hagen University Press. 10.57813/20220620-160005-0.
- König, Mareike. 2017. "Digitale Methoden in der Geschichtswissenschaft: Definitionen, Anwendungen, Herausforderungen." *BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen* 30, 1–2: 7–21. <https://www.budrich-journals.de/index.php/bios/article/download/33241/28560> (zugegriffen: 2. August 2022).
- Maier, Daniel, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri und S. Adam. 2018. "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology." *Communication Methods and Measures* 12, 2–3: 93–118 10.1080/19312458.2018.1430754.
- Ramsay, Stephen. 2011. *Reading Machines. Toward an Algorithmic Criticism*. Urbana (Illinois): University of Illinois Press.
- Sellin, Volker. 2008. *Einführung in die Geschichtswissenschaft*. Göttingen: Vandenhoeck & Ruprecht 2. Aufl.
- Uglanova, Inna und Evelyn Gius. 2020. "The Order of Things. A Study on Topic Modeling of Literary Texts." *CEUR Workshop Proceedings* 2723: 57–76. <http://ceur-ws.org/Vol-2723/long7.pdf> (zugegriffen: 2. August 2022).
- Vayansky, Ike und Sathish A. P. Kumar. 2020. "A review of topic modeling methods." *Information Systems* 94 10.1016/j.is.2020.101582.
- Wehrheim, Lino. 2019. "Economic history goes digital: topic modeling the Journal of Economic History." *Cliometrica* 13: 83–125.
- Wettlaufer, Jörg. 2016. "Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern." *Zeitschrift für digitale Geisteswissenschaften* 1 10.17175/2016_011.

duerer.online
– Virtuelles
Forschungsnetzwerk
Albrecht Dürer

Große, Peggy

grosse@ub.uni-heidelberg.de
Universität Heidelberg, Deutschland

Das seit 2020 von der DFG geförderte Projekt "duerer.online - Virtuelles Forschungsnetzwerk Albrecht Dürer"¹ baut eine Forschungsumgebung mit vollständigem Werkverzeichnis der Druckgraphik, Gemälde und Zeichnungen des bedeutenden Renaissance-Künstlers sowie

dessen Nachlebens auf. In der bis 2023 laufenden ersten Projektphase steht die Erfassung von Werken im Fokus, die im schriftlichen Nachlass explizit genannt sind. Die Bestände der kooperierenden Sammlungen (Kunstsammlungen der Stadt Nürnberg, Albrecht-Dürer-Haus-Stiftung e.V., Nürnberg, Germanischen Nationalmuseum und Albertina, Wien) dienen als Erschließungsgrundlage. Grundsätzlich können und werden bereits Informationen zu Beständen weiterer Sammlungen erfasst. Ziel ist es nach Projektende der Fachcommunity zu ermöglichen, neue Forschungsergebnisse zu den in der Datenbank aufgenommenen Werken zu ergänzen und Werke bzw. Exemplare hinzuzufügen. Nach Projektende soll es der Fachcommunity ermöglicht werden, neue Forschungsergebnisse zu den in der Datenbank aufgenommenen Werken zu ergänzen und Werke bzw. Exemplare hinzuzufügen. Die Fortführung, Pflege und Sichtbarkeit für die Fachcommunity über die Laufzeit des Projektes hinaus wird durch die Integration in das Angebot des Fachinformationsdienstes Kunst, Fotografie, Design - arthistoricum.net an der UB Heidelberg² gesichert.

Im Unterschied zu den bereits im Druck vorliegenden Werkverzeichnissen des Künstlers (wie z.B. Schoch/Mende/Scherbaum 2001-2004 für das graphische Werk, Anzelewsky 1991 für Gemälde und Winkler 1936-1939 für Zeichnungen) werden in der Forschungsumgebung Dürers Werke gattungsübergreifend nach einheitlichen Kriterien erschlossen, ebenso Werke der bis in die heutige Zeit andauernden Rezeption. Dabei versteht sich das Projekt nicht als ein weiteres autoritatives Werkverzeichnis, d.h. es werden keine Werke zu- oder abgeschrieben, sondern offen und transparent historische und aktuelle Diskussionen bezüglich der Autorschaft abgebildet und somit eine Grundlage für weitere Forschungen rund um das Werk Dürers und seiner Nachfolge geschaffen.

Neben der Erschließung der Werke werden ausgewählte Quellen des schriftlichen Nachlasses transkribiert und ediert. Durch die Auszeichnung mittels TEI³ und der Anreicherung der ausgezeichneten Werke, Personen und Orte mit Normdaten der Gemeinsamen Normdatei (GND) ist eine Durchsuch- und Recherchierbarkeit dieses Materials möglich, die für Forschende eine Neuheit darstellt. Über eine programmierte Pipeline werden zudem Registerdateien erzeugt, die zusätzlich mit Informationen zu Kunstwerken und Personen aus der Datenbank angereichert werden. Ebenso wird im Datensatz zum jeweiligen Werk/Person automatisch die Verlinkung auf den entsprechenden Registereintrag abgelegt, sodass sich in „duerer.online“ Nennungen der Entitäten in Quellen nachvollziehen lassen.

Für das Portal wird die „Wissenschaftliche Kommunikationsinfrastruktur (WissKI)“⁴ eingesetzt, eine virtuelle Forschungsumgebung, die den Aufbau von Anwendungen im Bereich der Digital Humanities unter Nutzung von semantic web-Standards ermöglicht. Die Einordnung und Speicherung der erhobenen Daten erfolgt ontologiebasiert mittels einer auf CIDOC CRM⁵ (ISO-Standard 21127) basierenden Anwendungsontologie und unter Nutzung der Gemeinsamen Normdatei (GND), des Getty AAT⁶ und Iconclass⁷. Das Projekt nutzt die „Heidelberger Anwendungsontologie für Werkverzeichnisse“ (Sobriell 2022) nach, die um Klassen und Eigenschaften erweitert wurde, die zur Dokumentation eines

Kunstwerkes und dessen Werkbiographie benötigt werden. So wurden vor allem Properties aufgenommen, die Beziehungen zwischen Werken beschreiben. Die Relationen (z.B. has copy/ is copy after) wurden nach dem Lightweight Information Describing Objects (LIDO) Schema modelliert (Sobriell 2022, 43).

Das mit der Anwendungsontologie umgesetzte Datenmodell berücksichtigt die Erfassung von Unikaten und Werken in mehreren Ausführungen. Das Werkkonzept beschreibt die inhaltliche Entstehung und nimmt alle Informationen auf, die jede Ausführung betreffen und konzeptioneller Natur sind. In der Ebene Ausführung/Exemplar werden spezifische Angaben zu einer Ausführung bzw. einem Exemplar dokumentiert. Das Datenmodell kann somit semantisch korrekt die Rezeption von Originalen darstellen, da diese immer vom Inhalt/Konzept ausgeht. Außerdem müssen keine Informationen mehrfach erfasst werden und zukünftig können graphische Sammlungen Inhalte zu ihrer speziellen Ausführung dem entsprechenden Werk-Datensatz hinzufügen. Das beschriebene Datenmodell und die darauf beruhende konsequente Erschließung von Werk- und Ausführungs-/Exemplarebene ist in ihrer Durchsuchbarkeit in sammlungsübergreifenden Graphik-Portalen derzeit singulär.

Neben Beziehungen von Werken untereinander werden u.a. Informationen zu Verkaufsereignissen und Ausstellungen erfasst. Durch Verknüpfung mit den jeweils verkauften oder ausgestellten Werken bzw. Ausführungen können historische nicht mehr existierende Sammlungen bzw. die Ausstellungshistorie eines Exemplars/Ausführung sichtbar gemacht werden. Zudem wird, wenn möglich, seitengenau auf die zur Verfügung stehenden Digitalisate von Auktions- und Ausstellungskatalogen verlinkt, sodass der Forschende schnell an die jeweiligen Nachweise gelangt.

Die Projekte zum Werk von Lucas Cranach d. Ä.⁸ und Rembrandt van Rijn⁹ sind dem beschriebenen Portal vergleichbar in ihrer Ausrichtung, doch für das Werk Albrecht Dürers besteht bisher kein Angebot, das gattungs- und sammlungsübergreifend Informationen zur Verfügung stellt und dabei auf anschlussfähige offene Formate gemäß der FAIR-Prinzipien unter Verwendung von Semantic-Web-Standards setzt. Damit steht der Dürer-Forschung ein Instrument zur Verfügung, das nicht nur zukünftige Forschungsergebnisse aufnehmen und sichtbar machen kann, sondern auch alle Informationen über Schnittstellen (SPARQL Endpoint) verfügbar macht, sodass es Forschenden möglich ist, mit eigenen Anwendungen Forschungsfragen weiter zu verfolgen.

Fußnoten

1. duerer.online [<https://sempub.ub.uni-heidelberg.de/duerer.online/>] ist seit März 2022 als First View veröffentlicht, <https://blog.arthistoricum.net/beitrag/2022/03/22/first-view-erste-werke-auf-duerer-online>. Dr. Franziska Ehrl ist für die wissenschaftliche Erschließung und Dokumentation der Werke verantwortlich.
2. <https://www.arthistoricum.net/themen/wvz/albrecht-duerer>.

3. <https://tei-c.org/>.
4. <https://wiss-ki.eu/>.
5. CIDOC CRM Conceptual Reference Model wird von der Special Interest Group des International Council for Museums entwickelt, <https://www.cidoc-crm.org/>.
6. Getty Art and Architecture Thesaurus des Getty Research Institute [<https://www.getty.edu/research/tools/vocabularies/aat/>].
7. Iconclass der Henri van de Waal Foundation [<https://iconclass.org/>].
8. Das seit 2009 von der Mellon Foundation "Cranach Digital Archive" konzentriert sich überwiegend auf die Gemälde des Künstlers [<https://lucascranach.org/>].
9. The Rembrandt Database ist eine vom RKD - Netherlands Institute for Art History bereitgestellte Forschungsplattform zu den Gemälden Rembrandts [<https://rembrandtdatabase.org/>].

Bibliographie

Anzelewsky, Fedja. 1991. Albrecht Dürer. Das malerische Werk. 2. neu bearbeitete Auflage Berlin: Deutscher Verlag für Kunstwissenschaft.

Schoch, Rainer, Matthias Mende, Anna Scherbaum, Anna. 2001-2004. Albrecht Dürer. Das druckgraphische Werk. 3 Bde., München u.a.: Prestel.

Sobriol, Nicole. 2022. Semantische Datenmodellierung für Werkverzeichnisse (Universitätsbibliothek Heidelberg). <https://doi.org/10.11588/data/64KP3N>, heiDATA, V1 (zugegriffen: 3. August 2022).

Winkler, Friedrich. 1936-1939. Die Zeichnungen Albrecht Dürers. 4 Bde., Berlin: Deutscher Verein für Kunstwissenschaft.

Dunkelgrün, blassgrün, fenchelgrün oder: Über die Konkretisierung des Vokabulars im deutschsprachigen Roman (1760–1920)

Hilger, Agnes

agnes.helene.hilger@gmail.com

Julius-Maximilians-Universität Würzburg, Deutschland

Hintergrund

Im Jahr 2012 entdecken Ryan Heuser und Long Le-Khac, dass eine Reihe von semantisch verwandten Wörtern in englischsprachigen Romanen über das 19. Jahrhundert hinweg immer häufiger verwendet wird (vgl.

Heuser/Le-Khac 2012). Diese Wörter sind tendenziell konkret.¹ Sie bezeichnen Körperteile wie *finger* oder *hair* oder Farben wie *red* oder *scarlet*. Ted Underwood zeigt in einer späteren Untersuchung, dass dieser Trend bereits 1760 einsetzt. Er weist angesichts des mehrere Epochen umfassenden Anstiegs auf eine Lücke im bisherigen literaturgeschichtlichen Wissen hin.

Eine Masterarbeit, die dem Poster zugrunde liegt, verfolgte das Ziel, die Beobachtungen von Heuser, Le-Khac und Underwood zunächst versuchsweise für die deutschsprachige Literatur nachzuvollziehen und so dann eine erste Eingrenzung derjenigen Bereiche zu leisten, die von der Entwicklung betroffen sind. Die Ergebnisse sollen hier vorgestellt werden.

Korpus und Methode

Das Korpus basiert auf den bei TextGrid und Projekt Gutenberg digital zur Verfügung gestellten Texten (vgl. Neuroth u.a. 2015; Reuters o.J.). Es enthält 1147 zwischen 1760 und 1920 erschienene Romane. Diese sind jedoch nicht gleichmäßig über den Zeitraum verteilt (s. Figure 1). Die Unausgewogenheit soll in anschließenden Arbeiten angegangen werden.

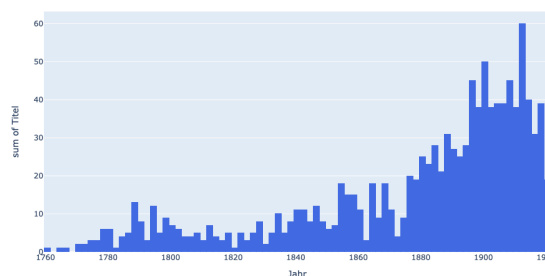


Figure 1: Übersicht über die im Korpus enthaltenen Romane pro Jahr.

Um die Wortfrequenzen zu ermitteln, wurden zunächst Wortlisten erstellt. Als Heuristik diente eine sehr grobe Einteilung in drei Gruppen: 1) Informationen zu Figuren, 2) Informationen zu Räumen und 3) Informationen zu Beschaffenheit und Material. Auf Basis dieser Unterscheidung wurde eine Liste von 31 Wortfeldern erstellt.² Anschließend wurden auf Basis von GermaNet Listen von zu diesen Wortfeldern gehörenden Wörtern erstellt (vgl. Henrich/Hinrichs 2010). Ein Beispiel für ein solches Wortfeld sind Farbwörter. Ausgehend vom Knoten *farbspezifisch* wurden dessen Hyponyme – zum Beispiel *grün* – und wiederum dessen Hyponyme – beispielsweise *dunkelgrün* und *blassgrün* – extrahiert.

Um der historischen Sprachstufe und der Domäne Roman gerecht zu werden, wurden die Wortlisten anschließend mit einem Word-Embedding-Modell erweitert. Dafür wurde ein auf CommonCrawl trainiertes Fasttext-Modell auf dem Roman-Korpus weitertrainiert (vgl. Bojanowski 2016). Aufgrund guter Performance in ähnlichen Tasks schien ein solches Fasttext-Modell ausreichend (vgl. Ehrmanntraut u.a. 2021). Um die Wortlisten zu erweitern, wurden zu den extrahierten Wörtern abhängig

von der Länge der Liste die zwei bis zwanzig nächsten Nachbarn ermittelt und, sofern nicht schon vorhanden, der Liste angehängt. Neben Wörtern wie *grün* oder *dunkelgrün* enthielt die Liste nun auch sehr spezifische wie *fenchelgrün*. Anschließend wurden die Listen von Hand bereinigt.

Die Romane wurden mit dem Python-Paket *spaCy* lemmatisiert und für jedes Wortfeld die zugehörigen Wortfrequenzen berechnet (vgl. Honnibal/Montani 2017).

Ergebnisse

Nimmt man die Wortfrequenzen aller 31 konkreten Wortfelder zusammen, ergibt sich ein signifikanter Anstieg (Mann-Kendall-Test, $\# = 0,01$, $p = 2,22e-16$). In Figure 2 repräsentiert jeder Punkt einen Roman, die y-Achse gibt jeweils die Wortfrequenz an. Im Korpus gibt es also einen ähnlichen Trend wie in den englischsprachigen Texten.

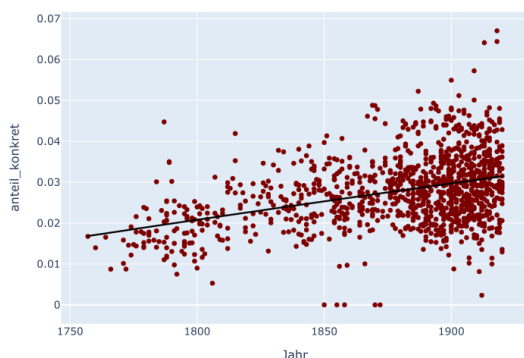


Figure 2: Frequenzen der konkreten Wörter pro Roman und Jahr.

Insbesondere bei den Wortfeldern, die Figuren, Gebäude und Innenräume beschreiben, gibt es signifikante Anstiege. Bei den Wortfeldern, die der Darstellung von Naturräumen (z.B. die Wortfelder *Gartenanlage*, *Gewächs*, *Bewaldung*) dienen dürften, gibt es dagegen keine signifikanten Trends. Figure 3 gibt einen Überblick über die Ergebnisse zu den untersuchten Wortfeldern ($\# = 0,01$).

Figure 3: Übersicht über die Trends für alle untersuchten Wortfelder.

	Anstieg	Rückgang	Kein Trend
1) Räume	Bau/Gebäude Gebäudeteil Zimmer Einrichtungsgegenstand/ Möbel Heimtextilie Haushaltsgegenstand/ Haushaltsprodukt Dekorationsgegenstand/ Ziergegenstand Gras/Grünfläche	Pflanzenteil	Gartenanlage/ Grünanlage Weg Gewächs/Pflanze Bewaldung/Wald Kunstobjekt Bild
2) Figuren	Körperteil Bekleidung/Kleidung Bekleidungsstück/ Kleidungsstück Aufmachung/Outfit Aussehensspezifisch Tasche		Ge- schmeide/Schmuck
3) Beschaffenheit	Gewebe/Stoff/Textil Farbspezifisch Helligkeitsspezifisch Oberflächenspezifisch Muster/Musterung Formspezifisch Geruch		Ornament/Verzierung Holz

Fazit

Die Ergebnisse der Arbeit deuten auf eine grundlegende Veränderung im untersuchten Korpus hin, die von der germanistischen Literaturgeschichte bislang nicht wahrgenommen wurde: Die Art und Weise, in der Romane ihre fiktive Welt physisch gestalten, wandelt sich über einen mehrere Epochen umfassenden Zeitraum hinweg erheblich. Zudem ermöglichen die Ergebnisse der Arbeit eine erste Differenzierung. Besonders betroffen scheinen Informationen über das physische Erscheinungsbild von Figuren und Orten, an denen Figuren leben. Eine anschließende Arbeit soll diese Ergebnisse konkretisieren.

Fußnoten

1. Die Zuschreibung ‚konkret‘ ist als Heuristik zu verstehen. Ausschlaggebend ist für die Zuordnung anders als in der sprachwissenschaftlichen Abgrenzung von Konkreta und Abstrakta nicht die Gegenständlichkeit, sondern die Möglichkeit der sinnlichen Wahrnehmbarkeit. Dabei orientiere ich mich an der Definition, die Sabine Schulte im Walde und Maximilian Köper für ihren Datensatz benutzen (vgl. Köper/Schulte im Walde 2016).
2. Der Begriff ‚Wortfeld‘ ist hier weit gefasst und bezeichnet anders als in der Sprachwissenschaft keine Gruppe synonymen Wörter, sondern eine Gruppe von semantisch verwandten Wörtern, die sich unter einen Oberbegriff subsumieren lassen. Für die Wortfelder, die Gegenstände umfassen, fällt der Begriff daher mit dem des Sachfeldes zusammen (vgl. Fries 2016, S. 774).

Bibliographie

Bojanowski, Piotr, Edouard Grave, Armand Joulin und Tomas Mikolov. 2016. „Enriching word vectors with sub-word information“. *arXiv preprint arXiv:1607.04606*.

Ehrmanntraut, Anton, Thora Hagen, Leonard Konle, und Fotis Jannidis. 2021. „Type- and Token-Based Word Embeddings in the Digital Humanities“. *Proceedings of the Conference on Computational Humanities Research*, 16–38.

Fries, Norbert. 2016. „Wortfeld“. In *Metzler Lexikon Sprache*, hg. von Helmut Glück und Michael Rödel, Stuttgart: Metzler.

Henrich, Verena und Erhard Hinrichs. 2010. „GernE-diT - The GermaNet Editing Tool“. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, 2228–2235.

Heuser, Ryan, und Long Le-Khac. 2012. „A Quantitative Literary History of 2,958 Nineteenth-Century British Novels. The Semantic Cohort Method“. *Stanford Literary Lab Pamphlets*.

Honnibal, Matthew, und Ines Montani. 2017. „spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing“.

Köper, Maximilian, und Sabine Schulte im Walde. 2016. „Automatically Generated Norms of Abstractness, Arousal, Imageability and Valence for 350,000 German Lemmas“. In *Proceedings of the 10th Conference on Language Resources and Evaluation (LREC)*.

Neuroth, Heike, Andrea Rapp, und Sibylle Söring (Hgg). 2015. *TextGrid. Von der Community - für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt.

Reuters, Hella (Hg.). o.J. Projekt Gutenberg-DE, online unter: www.projekt-gutenberg.org [zuletzt aufgerufen am 14.12.2022].

Underwood, Ted. 2019. *Distant horizons. Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.

D-WISE – Digitale Wissenssoziologische Diskursanalyse

Fischer, Tim

tim.fischer@uni-hamburg.de
Universität Hamburg, Deutschland

Eiser, Isabel

isabel.eiser@uni-hamburg.de
Universität Hamburg, Deutschland

Schneider, Florian

florian.schneider-1@uni-hamburg.de
Universität Hamburg, Deutschland

Petersen-Frey, Fynn

fynn.petersen-frey@uni-hamburg.de
Universität Hamburg, Deutschland

Biemann, Chris

christian.biemann@uni-hamburg.de
Universität Hamburg, Deutschland

Koch, Gertraud

gertraud.koch@uni-hamburg.de
Universität Hamburg, Deutschland

Das D-WISE Projekt

Das Verbundprojekt D-WISE (www.dwise.uni-hamburg.de) ist ein BMBF gefördertes interdisziplinäres Kooperationsprojekt an der Universität Hamburg zwischen den Geisteswissenschaften, vertreten durch das Institut für Empirische Kulturwissenschaft, und dem Fachbereich Informatik, vertreten durch die Arbeitsgruppe Language Technology (Koch et al. 2022). D-WISE entwickelt neue informatische Analyseverfahren unter Einsatz von kontextorientierten Embedding-Repräsentationen und eine Arbeitsumgebung (die D-WISE Tool Suite) als digitale Unterstützung von wissenssoziologischen Diskursanalysen (Keller 2011). Dabei orientiert sich D-WISE an zwei grundlegende Fragestellungen: (1) Zu welchen Zwecken, wann und wie können Automatisierung und DH-Methoden sinnvoll in qualitative diskursanalytische Ansätze und Wissensproduktion integriert werden; welche bestehenden Methoden und Tools können übernommen werden und welche müssen neu entwickelt werden? (2) Wie können hermeneutische Methoden durch die Nutzung (halb-)automatisierter Forschungsprozesse weiterentwickelt werden?

Das D-WISE Projekt zielt darauf ab, den Mangel an digitalen Lösungen für multimodale Diskursanalysen zu beheben. Die Lösungen sollen in der Lage sein, mit der Multimodalität von Materialien sowie der Pluralität von Bedeutungen umzugehen, welche vielfältige Herausforderungen und hohe Komplexitätsanforderungen für digitale Lösungen darstellen. Gleichzeitig müssen die ständig wachsende Zahl digitaler Materialien bewältigt werden, indem die Methodik durch digitale Methoden erweitert wird.

Die Überbrückung der Lücke zwischen strukturellen Mustern, die mit digitalen Methoden aufgedeckt werden, und interpretativen Prozessen menschlicher Bedeutungsproduktion steht im Mittelpunkt des kollaborativen Ansatzes von Kulturanthropologie und Sprachtechnologie im D-WISE-Projekt. Durch die Kombination von manuellen und digitalen Ansätzen zur Diskursanalyse wird die D-WISE Tool Suite für die digitale qualitative Diskursanalyse in enger Co-Creation Arbeitsweise (Eiser et al. 2023) zwischen Wissenschaftler:innen aus den Geisteswissenschaften und der Informatik geschaffen.

Die D-WISE Tool Suite für KI-gestützte Wissenssoziologische Diskursanalysen

Die D-WISE Tool Suite¹ (DWTS) wird als offen zugängliche Webapplikation entwickelt und ausgehend von der Evaluation und Reflektion von etablierten DH-Methoden sowie digitalen Tools (u.a. MAXQDA, CATMA, WebAnno) konzipiert. DWTS erweitert das bestehende Angebot und verbessert den Forschungsprozess, insbesondere Diskursanalysen, mittels künstlicher Intelligenz (KI) gestützten Funktionalitäten und teilautomatisierten Prozessen. Hierbei spielen die sich ergänzenden Konzepte von Human-in-the-Loop (Holzinger 2016) und AI-in-the-Loop (Koch et al. 2022) eine zentrale Rolle und werfen für D-WISE relevante Fragestellungen hinsichtlich der Interaktion von Mensch und Algorithmus zur Verbesserung von KI-Forschungssystemen und deren Auswirkungen auf den menschlichen Forschungsprozess auf. Zur Umsetzung werden, extern wie intern, Open-Source-Software, -Bibliotheken und -Lösungen verwendet.

Das Projekt befindet sich in seinem zweiten Projektjahr mit einer entwickelten Tool Suite, in der zunächst Standard-Funktionalitäten implementiert wurden, wie das Suchen und Filtern, Kodieren und Annotieren und Dokumentieren. Insbesondere wurde auch ein besonderer Fokus auf projekt-zentriertes und kollaboratives Arbeiten, sowie die Unterstützung aller diskursanalytischen Schritte gelegt – vom Crawling und Filtern von Daten, über die Spezifikation des Forschungsphänomens bis hin zu Analyse, Auswertung und Interpretation, aber auch Dokumentation und Reflexion. Die DWTS unterstützt dabei die drei Kodierungsphasen des offenen, axialen und selektiven Kodierens der Grounded Theory, wodurch Muster und Konzepte identifiziert, verknüpft und analysiert werden können. Auch implementiert wurde die kollaborative Arbeit und Memofunktion. Die D-WISE Tool Suite bietet ferner die automatische Erkennung und Codierung von Entitäten², wie z.B. Akteur:innen, Organisationen oder Orten sowie Bild-Annotationen. Das integrierte Logbuch ermöglicht einerseits eine automatisierte Erfassung aller getätigten Änderungen zur Dokumentation des Forschungsprozesses sowie die Möglichkeit eines manuell zu führenden Feldtagebuchs.

In den letzten Jahren gab es bedeutende Durchbrüche bei verschiedenen Aufgaben der Computer Vision (ViT, Dosovitskiy et al. 2021), der Verarbeitung von Natural Language Processing Tasks (BERT, Devlin et al. 2019) sowie bei multimodalen Aufgaben (CLIP, Radford et al. 2021). Entsprechende Erkenntnisse und Entwicklungen erlauben dem D-WISE Team multimodale Daten, die aus Text, Bild, (zukünftig auch Video oder Audio) oder einer Mischung aus allen Modalitäten bestehen, innerhalb der D-WISE Tool Suite zu verarbeiten. Insbesondere wird dadurch eine multi-modale Ähnlichkeitssuche³, also beispielsweise das Auffinden ähnlicher Bilder zu einem Suchtext, ermöglicht. Diese semantische und multi-modale Suchfunktionalität ist neben einer lexikalischen Suche in DWTS integriert.

Im Gegensatz zu meist lizenzbasierten und kostenpflichtigen, umfangreichen ‚All-in-one-Lösungen‘ für

qualitative Datenanalyse wie MAXQDA, Atlas.ti oder NVivo, wird die D-WISE Tool Suite als frei verfügbare Open-Source-Software mit Fokus auf die Wissenssoziologische Diskursanalyse konzipiert. Daher sind hermeneutisch-zirkuläre Methoden, Filterung und skalierbares Lesen vorherrschende Konzepte. Die Tool Suite unterstützt alle diskursanalytischen Schritte und erlaubt einen zyklischen Forschungsprozess, der eng an den hermeneutischen Zirkel angelehnt ist: DWTS soll beim Hinterfragen und der ständigen Erweiterung des Wissenstandes unterstützen, um zu neuen Fragestellungen, Erkenntnissen und somit zu einem tieferen Verständnis des Phänomens zu gelangen.

Im nächsten Projektjahr wird KI-gestützte Sprach- und Videoverarbeitung sowie die Entwicklung von Datenanalysefunktionen vorangetrieben.

Fußnoten

1. <https://github.com/uhh-lt/dwts>
2. Mittels multi-lingualen Modellen von spaCy (Honnibal 2020).
3. Anhand von Word Embeddings (Vasiwani 2017; Devlin 2019) durch Sentence Transformer (Reimers 2019) und CLIP (Radford 2021) zur Berechnung semantischer Repräsentationen der Texte und Bilder.

Bibliographie

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "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, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, and Xiaohua Zhai. Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *9th International Conference on Learning Representations, Volume*, 451, pp. 3–7. Online.

Eiser, Isabel, Tim Fischer, Florian Schneider, Gertraud Koch, Chris Biemann, Fynn Petersen-Frey. 2023. "Open Science Prinzipien und interdisziplinäre Kollaboration in D-WISE: Zwischen Hermeneutik und Digitaler Methode in der Diskursanalyse". *DHd2023 Luxemburg/Trier. Open Humanities – Open Culture. Konferenzabstracts*.

Holzinger, Andreas. 2016. "Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop?" *Brain Informatics* 3, no. 2: 119–131.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, Adriane Boyd. 2020. "spaCy: Industrial-strength Natural Language Processing in Python".

Keller, Reiner. 2011. "Wissenssoziologische Diskursanalyse: Grundlegung eines Forschungsprogramms." 3. Aufl.

Interdisziplinäre Diskursforschung. Wiesbaden: VS Verlag für Sozialwissenschaften.

Koch, Gertraud, Chris Biemann, Isabel Eiser, Tim Fischer, Florian Schneider, Teresa Stumpf, and Alejandra Tijerina Garza. 2022. "D-WISE Tool Suite for the Sociology of Knowledge Approach to Discourse." In *Culture and Computing*, edited by Matthias Rauterberg, 68–83. Cham: Springer International Publishing.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." In *Proceedings of the 38th International Conference on Machine Learning*, edited by Marina Meila and Tong Zhang, 139:8748–63. Proceedings of Machine Learning Research. PMLR. <https://proceedings.mlr.press/v139/radford21a.html>.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–92. Hong Kong, China: Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30*, herausgegeben von I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, und R. Garnett, 5998–6008. Curran Associates, Inc.

Fabrikation von Erkenntnis: Experimente in den Digital Humanities

Dieckmann, Lisa

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

Steyer, Timo

t.steyer@tu-braunschweig.de
Universitätsbibliothek Braunschweig, Deutschland

Walkowski, Niels-Oliver

niels-oliver.walkowski@uni.lu
Universität Luxemburg, Luxemburg

Weis, Jolle

weis@uni-trier.de
Universität Trier, Deutschland

Wuttke, Ulrike

ulrike.wuttke@fh-potsdam.de
Fachhochschule Potsdam, Deutschland

Die vDHd2021-Tagung, die als Ersatz für die coronabedingte Verschiebung der DHd-Jahrestagung 2021 stattfand, stand unter dem Motto "Experimente" und wurde von einem Publikationsexperiment begleitet, das von der Community initiiert wurde. Das Herausgeber*innengremium bildete sich in der Folge im Rahmen der vDHd-Planungen. Unter dem Titel „Fabrikation von Erkenntnis: Experimente in den Digital Humanities“ wurde schließlich eine digitale Publikation herausgegeben, die das experimentelle Potenzial der Digital Humanities in unterschiedlichen Beitragstypen ergründet (Pawlicka-Deger 2020, Lane 2016, Earhart 2015, Knorr-Cetina 1991).

Zwar ist der Band losgelöst von den Beiträgen der vDHd2021, dennoch greift er den Anspruch der DHd-Jahrestagungen auf, zur Sichtbarkeit aktueller DH-Aktivitäten im deutschsprachigen Raum beizutragen, wobei er selbst einen experimentellen Ansatz als "living publication" verfolgt. Das Poster thematisiert die experimentellen Aspekte der Publikation und die damit verbundenen Erfahrungen.

Am Beginn stand ein Call for Publications (Burghardt et al. 2021), der die Erkundung des experimentellen Potenzials der DH in den Vordergrund stellte. Willkommen waren thematische und formale Experimente, die in einem rollenden Verfahren als Ko-Publikation und Sonderband der Zeitschrift für digitale Geisteswissenschaften und Melusina Press erscheinen sollten.

Drei Typen von Einreichungen waren möglich:

- Fachartikel zu experimentellen Methoden und Verfahren der DH
- Daten-Experimente / Publikation von Datensätzen
- Code-Experimente / Publikation von ausführbaren Notebooks

Die Kategorie der Fachartikel stellt die etablierteste Publikationsart innerhalb des Sonderbands dar. Das Experiment bei dieser Beitragsart bestand nicht im Bereich der eigentlichen Publikation, sondern im Reviewverfahren. Die Autor*innen konnten zwischen einem traditionellen Double-Blind und einem Open-Review-Verfahren wählen. So bot sich der Band an, Erfahrungen mit einem offenen Begutachtungsprozess zu sammeln und wertvolle Impulse für die Diskussionen innerhalb des DHd-Verbands bezüglich des zukünftigen Reviewverfahrens bei den Jahrestagungen zu liefern. Die überwiegende Mehrzahl der Autor*innen entschlossen sich für das offene Verfahren, das ein öffentliches Kommentieren eines Preprints vorsah. Auch der Community stand die Möglichkeit offen, die Artikel zu kommentieren und sich damit in den Veröffentlichungsprozess einzubringen. Zusätzlich verfassten die Gutachter*innen eine abschließende Bewertung mit Hinweisen zu Stärken und Verbesserungsmöglichkeiten. Wenn sich alle Seiten einverstanden erklärten, wurden auch diese Gutachtenformulare mit dem Artikel publiziert. Nach der Überarbeitung der eingereichten Fassung wurde der fertige Artikel publiziert, wobei Preprint und Anmerkungen auch weiterhin publiziert bleiben.

Die Kategorie der Datenexperimente stellt innerhalb des Sonderbandes eine neuere Publikationsart dar. Sie sind an das Beispiel sogenannter Data Papers (Schöpfel et al. 2019, 3) angelehnt und bestehen aus einem Artikel, der einen oder mehrere Datensätze beschreibt, vorrangig in Text und Bild, und der Publikation des zugehörigen Datensatzes. Im CfP wurde explizit nach eher unkonventionellen Datensätzen (*corpora obscura*) gefragt, wichtig war aber insbesondere, dass die Datensätze nach den FAIR-Prinzipien frei verfügbar sind beziehungsweise gemacht wurden.

Der Schwerpunkt der Data Papers liegt auf der Beschreibung der Datensätze bzw. der Potenziale für die Forschung, nicht so sehr auf den damit erzielten Forschungsergebnissen (Schöpfel et al. 2019, 3). Data Papers zählen zu den neueren, in den Geisteswissenschaften noch nicht weit verbreiteten Publikationsarten (Schöpfel et al. 2019, 9) und gehören ursprünglich nicht zur Bandbreite der ZfdG. Daher mussten die Herausgeber*innen eigene Richtlinien und Empfehlungen für die Autor*innen zur Umsetzung des Formats entwickeln und an diese auch die Reviewempfehlungen anpassen. Für die Entwicklung dieses Kriterienkatalogs orientierte sich das Herausgeber*innenteam an bereits bestehenden sogenannten Data Journals (wie z. B. das JOHD oder RDJ). Eine Herausforderung war das Finden geeigneter Reviewer*innen (bei den Data Papers single-blind Verfahren), da sowohl fachliches als auch technisches Verständnis für den jeweiligen Datensatz notwendig war.

Ein weiteres Experiment dieses Sammelbands ist die Veröffentlichung von Code Experimenten in Form von Executable Publications (ein ähnliches Format bietet etwa das JDH). Es handelt sich dabei um interaktive Jupyter Notebooks (Python und R-basiert), die zusammen mit den verwendeten Daten und einer technischen Dokumentation als Git-Repository eingereicht wurden. Alle Notebooks haben neben den Code-Abschnitten eine klar strukturierte und verständliche textuelle Komponente, durch die die häufig anzutreffende Rollenverteilung von Text als Mittel der Interpretation und Daten/Code als Ort empirischer Stringenz aufgebrochen wird. Dank der Code Experimente werden Methodik und Material zu evaluierbaren Gegenständen und die Publikation damit zum Medium sich ergänzender Mittel des Argumentierens. Bei dieser Publikationsform war, neben dem Review Prozess, vor allem der Aufbau einer entsprechenden Infrastruktur bei Melusina Press eine Herausforderung.

Das Poster wird nicht nur kurz auf alle Einreichungstypen eingehen und die Begutachtungsverfahren darstellen, sondern den Band auch statistisch evaluieren und besonders auf die gewonnenen Erfahrungen innerhalb des Herausgeber*innenteams eingehen (vgl. dazu auch Walkowski 2022). Dabei stehen, wie bereits angedeutet, besonders Lessons Learned in Bezug auf die Koordination von zwei Publikationsorten, technische Herausforderungen, den allgemeinen Kommunikationsaufwand und die Annahme des offenen Begutachtungsverfahrens im Mittelpunkt.

Bibliographie

Burghardt, Manuel, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis und Ulrike

Wuttke. "Call for papers Fabrikation von Erkenntnis: Experimente in den Digital Humanities." vDHD 2021 Experimente. <https://vdhd2021.hypotheses.org/uber/call-for-papers> (zugegriffen: 01. August 2022).

Burghardt, Manuel, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis und Ulrike Wuttke. 2021/2022. *Fabrikation von Erkenntnis. Experimente in den Digital Humanities*. Wolfenbüttel: ZfdG; Esch-sur-Alzette: Melusina Press 10.17175/sb005

Earhart, Amy E. 2015. *The digital humanities as a laboratory*. MIT Press.

Journal of Digital History (JDH).

Journal of Open Humanities Data (JOHD).

Knorr-Cetina, Karin. 1991. *Die Fabrikation von Erkenntnis: Zur Anthropologie der Naturwissenschaft*. Suhrkamp.

Lane, Richard J. 2016. *The Big Humanities: Digital Humanities/Digital Laboratories*. Routledge.

Pawlicka-Deger, Urszula. 2020. "The Laboratory Turn: Exploring Discourses, Landscapes, and Models of Humanities Labs." *Digital Humanities Quarterly*, 14 (3). <http://www.digitalhumanities.org/dhq/vol/14/3/000466/000466.html> (zugegriffen: 01. August 2022).

Research Data Journal for the Humanities and Social Sciences (RDJ)-

Schöpfel, Joachim, Dominic J. Farace, Hélène Prost und Antonella Zane. 2019. "Data Papers as a New Form of Knowledge Organization in the Field of Research Data." *HAL halshs-02284548*

Walkowski, Niels-Oliver. 2022. "Fabrikation von Erkenntnis: Experimente in den Digital Humanities." Media Centre Uni.lu. Präsentation der Publikation, 9:08. https://videos.uni.lu/media/Fabrikation+von+ErkenntnisA+Experimente+in+den+Digital+Humanities/1_bmhj3jpx (zugegriffen: 01. August 2022).

Fanfiction Semantics – Eine quantitative Analyse sensibler Themen in deutscher Fanfiction

Häußler, Julian

julian.haeussler@stud.tu-darmstadt.de

Technische Universität Darmstadt, Deutschland

Hintergrund

Als Fanfiction können all jene Texte bezeichnet werden, die zur Veröffentlichung in eigens dafür eingerichteten Internetforen bestimmt sind und in „appropriativ-derivativer bzw. [...] transformativer“ (Stemberger 2021, 10) Weise Bezug auf Mainstreammedien (meist Romane,

Serien oder Filme) nehmen. Für die Autor*innen von Fanfiction, die meist unter Pseudonym schreiben, ist es somit möglich, in einem geschützten Raum (der Fancommunity) neben ersten Schreibversuchen, zu ihren Lieblingsfiguren auch das Verfassen von die Bezugsmedien kontrastierenden Erzählungen zu wagen. Catherine Tosenberger beschreibt letztere Kategorie als "a freedom especially felt with regard to non-normative and taboo forms and representations of sexuality" (Tosenberger 2014, 17). Fanfiction, die Literatur als Bezugsquelle gewählt haben, bauen dabei meistens auf großen Mainstream-Jugendbuchreihen aus dem angelsächsischen Raum auf. So sammeln sich in der deutschsprachigen Community auf fanfiktion.de unter der Kategorie ‚Bücher‘ die meisten Autor*innen in den Fandoms zu Harry Potter (über 55.000 Texte gesamt), Bis(s) (knapp 14.000) und Herr der Ringe (als allg. Gruppe Mittel Erde definiert mit über 8.000 Texten). Praktisch befassen sich die Autor*innen oft mit Rekombinationen des Figurenarsenals der Originaltexte. Eine wichtige Rolle spielt dabei das Stereotyp der Slashes (männliche Figuren, die in eine romantisch-sexuelle Beziehung gesetzt werden; vgl. Brottrager et al. 2022). Die Moralvorstellungen der Originale können dabei auch übergangen werden, wenn zum Beispiel der Schüler Harry mit seinem Lehrer Severus Snape eine (romantisch-sexuelle) Beziehung eingeht.

Ziele des Projekts

Das Masterprojekt „Fanfiction Semantics – A Quantitative Analysis of Sensitive Topics in German Fanfiction“ hat zum Ziel, diese Kontrastierungen in verschiedenen Fandoms zu betrachten. Gefragt wird dabei einerseits, wie virulent sensible Themen (sensitive topics) in Fanfiction sind und andererseits, wie diese Themen Wortbedeutungen beeinflussen. Als sensibel werden dabei jene Themen definiert, die in Originaltexten mehrheitlich ausgespart und von Fanfiction-Autor*innen demonstrativ eingeführt werden (v. a. extreme Gewalt und sexuelle Inhalte). Das Korpus entstammt einem seit 2020 andauernden Webscrapings am LitLab der Technischen Universität Darmstadt (vgl. Weitn 2022). Hier wurden alle in diesem Zeitraum bearbeiteten Texte sowie die von der Plattform definierten Metadaten (neben Selbstangaben der Autor*innen auch Angaben zu Altersbeschränkung der Texte und der eigenen Definition von Genre) heruntergeladen. Das Projekt bearbeitet darauf aufbauend in zwei Schritten zunächst die Metadaten zu den größten Fandoms im Bereich ‚Bücher‘ und wertet anschließend die dazugehörigen Texte quantitativ aus. Der untersuchte Ausschnitt des Fanfiction-Korpus umfasst über 8.000 Einzeltexte mit knapp 130 Mio. Tokens.

Methodik

1) Im ersten Schritt, bei der Analyse der Metadaten, kann beispielsweise eine Verteilung davon, welche der Kategorien zur Altersbeschränkung für wie viele Texte verwendet wurde, einen Hinweis auf die Unterschiede zum Bezugsmedium liefern. Die auf fanfiktion.de gesetzte Kategorisierung reicht dabei von Texten, die ab

6 Jahren geeignet sein sollen, bis hin zu solchen, die als „entwicklungsbeeinträchtigend“ (fanfiktion.de) gelten. Entsprechende Verteilungen werden auch zwischen Fandoms verglichen.

2) Die Analyse der Texte erfolgt im folgenden Schritt mithilfe von Word Embedding Modellen (hier word2vec; vgl. Mikolov et al. 2013). In einem Word Embedding Modell können Ähnlichkeiten zwischen verschiedenen Schlüsselwörtern effizient verglichen werden. Die algorithmische Grundlage bietet dabei die Verwendung der Wörter in einem festgesetzten Kontextfenster (vgl. Distributionshypothese nach Firth; vgl. Firth 1957). Durch den Abgleich von Schlüsselwörtern, die auf verschiedene Themen verweisen, können Schnittmengen und Unterschiede untersucht werden. Die Schlüsselwörter wurden hierzu aus den Wortfeldern Gewalt und sexuelle Inhalte gewählt, welche Wörter unter den (hier) 30 ähnlichsten zu einem Zielwort zu finden sind, beschreibt in gewisser Weise wie dieses Zielwort verwendet wird (vgl. Abb. 1). Für das Subkorpus der *Harry Potter*-Reihe werden zudem dieselben Operationen auch für die Originaltexte durchgeführt. Abschließend wird ein Verfahren der Word Embedding basierten Sentiment Analyse nach SentiArt (vgl. Jacobs 2019; Brottrager et al. 2022) durchgeführt, welche anhand der Variablen Valence (emotionale Wertigkeit) und Arousal (emotionales Erregungspotential) Schlüsselwörter beschreibt.

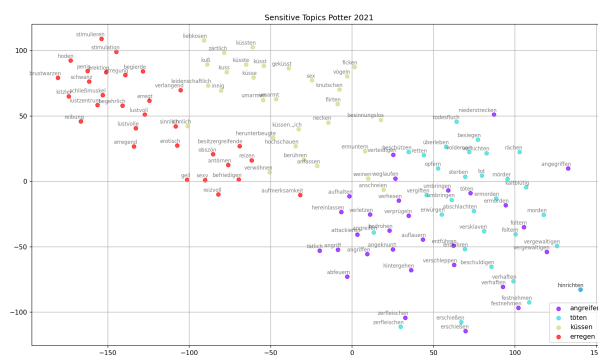


Abb. 1: t-SNE-Visualisierung der ähnlichsten Wörter zu vier Schlüsselwörtern in einem Word2Vec-Modell zu *Harry Potter*-Fanfiction.

Ausblick

Das Poster soll die Ergebnisse der oben beschriebenen Analysen grafisch darstellen und einen Eindruck über die thematische Zusammensetzung der Fanfiction-Korpora bieten. Das zugrundeliegende Korpus kann aus urheberrechtlichen Gründen nicht in der verwendeten Form veröffentlicht werden. Dennoch ist es beabsichtigt, nach Abschluss des Projekts alle verwendeten Codes auf GitHub bereitzustellen, um Transparenz zu schaffen und die Methodik für andere Projekte nachnutzbar zu machen (orientiert an der in den FAIR-Prinzipien definierten Wiederverwendbarkeit (vgl. GO FAIR 2022)). Anonymisierte Metadatatabellen sowie die Word Embedding-Modelle sollen zudem auch zugänglich gemacht werden.

Bibliographie

Brottrager, Judith, Joël Doat, Julian Häußler, und Thomas Weitin. 2022. „Character Shifts in Harry Potter Fanfiction“. Herausgegeben von Thomas Weitin. *LitLab Pamphlet*, Nr. #10.

fanfiktion.de. 2022. „Richtlinien für die Alterskennzeichnung“. <https://www.fanfiktion.de/p/ageadvice/0> (zugegriffen: 03. August 2022)

Firth, John. 1957. „A Synopsis of Linguistic Theory, 1930–1955“. In *Selected Papers of J.R. Firth 1952–1959*, herausgegeben von Frank Palmer, 168–205. London: Longman.

GO FAIR. 2022. „FAIR Principles“. <https://www.go-fair.org/fair-principles/> (zugegriffen: 03. August 2022).

Jacobs, Arthur M. 2019. „Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics“. *Frontiers in Robotics and AI* 6. <https://doi.org/10.3389/frobt.2019.00053>.

Mikolov, Tomas, Kai Chen, Greg Corrado, und Jeffrey Dean. 2013. „Efficient Estimation of Word Representations in Vector Space“. *arXiv*. <https://arxiv.org/abs/1301.3781>.

Stemberger, Martina. 2021. *Homer meets Harry Potter: Fanfiction zwischen Klassik und Populärkultur*. Dialoge. Tübingen: Narr Francke Attempto.

Tosenberger, Catherine. 2014. „Mature Poets Steal: Children's Literature and the Unpublishability of Fanfiction“. *Children's Literature Association Quarterly* 39 (1): 4–27. <https://doi.org/10.1353/chq.2014.0010>.

Weitin, Thomas. 2022. „Litlab“. <https://www.litlab.tu-darmstadt.de/institutlinglit/mitarbeitende/wei-tin/litlab/index.de.jsp> (zugegriffen: 03. August 2022).

Kaleidoskopische Muster des Protests. Visuelle und textuelle (Selbst-) Repräsentationen osteuropäischer Protestkulturen aus qualitativer und quantitativer Perspektive

Howanitz, Gernot

gernot.howanitz@uibk.ac.at

Universität Innsbruck, Österreich

Kaltseis, Magdalena

Magdalena.Kaltseis@aau.at

Universität Innsbruck, Österreich

Motivation

Die Posterpräsentation stellt das Projekt „Kaleidoskopische Muster des Protests“ vor, das visuelle und textuelle (Selbst-)Repräsentationen osteuropäischer Protestkulturen sowohl aus qualitativer als auch aus quantitativer Perspektive untersucht. Dieses Projekt wird von der Österreichischen Akademie der Wissenschaften gefördert und startet im Jänner 2023.

Politische Protestbewegungen haben in Osteuropa, insbesondere in Russland, der Ukraine und Belarus in den letzten zehn Jahren großen Aufschwung erlebt. In diesen drei Ländern besetzten Protestierende öffentliche Plätze und verwendeten visuelle Symbole und Slogans, um andere Menschen dazu bewegen, sich den Protesten anzuschließen. Gleichzeitig wurden Bilder der Protestierenden auch von den autoritären Regierungen dieser Länder genutzt, um die Protestbewegungen zu delegitimieren.

Eine wichtige Rolle bei der Organisation der Proteste sowie deren Wahrnehmung spielen soziale Netzwerke (Smyth & Oates 2015; Onuch 2015) und die Medien, vor allem das staatlich kontrollierte Fernsehen, das nach wie vor als das wichtigste Informationsmedium in diesen drei Ländern gilt (vgl. Szostek 2018). Aus diesem Grund sind medial vermittelte (Selbst-)Repräsentationen, u.a. YouTube-Videos, Blogbeiträge, Kommunikation über soziale Netzwerke, Fernsehnachrichten sowie Dokumentarfilme, ein integraler Bestandteil der Proteste selbst. Symbole und Slogans werden dazu genutzt, um die Ideen und Forderungen der Protestierenden zu verbreiten, weshalb Proteste *per se* als „kommunikativer Akt“ charakterisiert werden können (Kuße 2021). So signalisierten beispielsweise bei den Protesten in Russland (2011/12) weiße Bänder eine regierungskritische Haltung, während eine regierungsfreundliche Einstellung durch schwarz-orangene Georgsbänder ausgedrückt wurde. Auf dem Euromaidan in der Ukraine (2013/14) waren neben der Europaflagge auch rechtsnationale Symbole präsent. In Belarus (2021) diente die weiß-rot-weiße Flagge als Hauptsymbol für die Proteste gegen die Regierung (vgl. Gaufman 2021).

Slogans verstehen wir im Sinne von Friedman (2019) als die „Spitze diskursiver Eisberge“ sowie als „wichtige symbolische Schlüssel“ zu sozialen Bewegungen, wobei wir uns bewusst sind, dass deren ursprüngliche Bedeutung mit der Zeit transformiert und neuinterpretiert werden kann. Im Unterschied zu Symbolen definieren wir Slogans als rein textuelle Erscheinungen.

Symbole lesen wir hingegen in Anlehnung an Cassirer (1996) als „soziale Phänomene“, die an die Stelle der rein sprachlichen Kommunikation treten. Sie sind mit einer bestimmten Bedeutung oder Bedeutungen aufgeladen, die sowohl individuell als auch im Kontext (im sozialen Raum) interpretiert werden können.

Im Laufe unserer quantitativen und qualitativen Analyse werden wir laufend überprüfen, ob diese Definitio-

nen noch Gültigkeit haben oder überarbeitet bzw. angepasst/nachgeschärft werden müssen. Die Veränderung der Symbolik möchten wir jedenfalls miteinbeziehen und aus unserer Sicht ist das aufgrund der unterschiedlichen Perioden und Länder, die wir im Projekt abdecken, möglich.

Methoden

Um die (Selbst-)Repräsentationen von Protest in den oben genannten osteuropäischen Ländern zu erfassen, betrachten wir diese aus drei verschiedenen Perspektiven: (1) die Selbstrepräsentation der Protestkulturen auf YouTube sowie in den sozialen Netzwerken, (2) deren offizielle Darstellung in regierungstreuen sowie in unabhängigen TV-Nachrichtensendungen und (3) ihre cineastische Darstellung in drei Dokumentarfilmen. Aufgrund dieser verschiedenen Blickwinkel sprechen wir in unserem Projekttitel auch metaphorisch von „kaleidoskopischen Mustern des Protests“, die unsere Forschung sichtbar machen soll. Zu diesem Zweck kombinieren wir eine automatische Symbolerkennung mittels künstlicher neuronaler Netze (R-CNN) mit der Multimodalen Diskursanalyse (MDA) aus der Linguistik (Kress 2011). Mit diesem Ansatz folgen wir N. K. Hayles' (2010) Kombination von „close reading“, „machine reading“, und „hyper reading“; übertragen auf die visuellen Medien wird dabei aus dem „machine reading“ Arnold/Tiltons (2019) „distant viewing“.

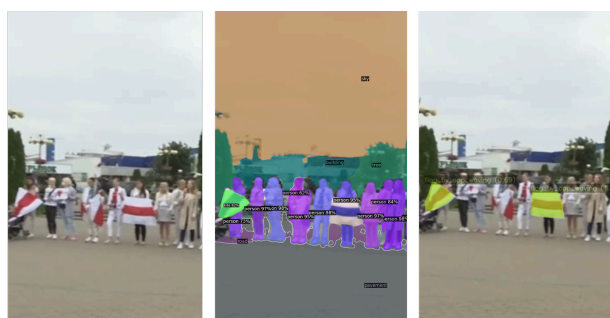


Abb. 1: Beispielframe aus einem YouTube-Video (links), Annotationen von einem vortrainierten Netz für panoptische Segmentierung (Mitte), Annotationen von unserem selbst trainierten Netz (rechts)

Der Rückgriff auf „distant viewing“ ist nicht zuletzt aufgrund der großen Datenmenge notwendig: Seit Februar 2022 haben wir über 74.000 YouTube-Videos zu Belarus, Russland und der Ukraine inklusive Metadaten gesichert – mit einer Gesamtlaufrzeit von knapp eineinhalb Jahren. Dieses Videokorpus wird von einem selbst trainierten neuronalen Netz nach nationalistischen Symbolen aus Osteuropa durchsucht. Konkret verwenden wir das Detectron2 Framework, weil es erlaubt, schnell eigene Netze zu trainieren, deshalb wird es verwendet; ergänzend werden wir auch vortrainierte Netze verwenden, etwa eine zero-shot-detection mit CLIP, um Szenenbeschreibungen zu bekommen.

Das eigene Training ist insofern notwendig, als frei verfügbare vortrainierte Netze das notwendige domänen-spezifische Wissen nicht mitbringen. Das Beispiel in Abb.

1 zeigt, wie das vortrainierte Netz zwar die grundlegende Konfiguration des Videoframes beschreiben kann; die für uns wichtigen Flaggen werden jedoch nicht oder falsch erkannt. Die kontextspezifische Information liefert somit unser eigenes Netzwerk, dessen Betaversion wir im Sinne der Open Humanities bereits veröffentlicht haben (Howanitz/Radisch 2022). Dieses Netz erreicht zur Zeit eine AP50 von 75.794 bei einem Trainingskorpus von 4.045 Bildern mit 8.156 Annotationen und einem Testkorpus mit 1.032 Bildern und 2.109 Annotationen. Wir haben 44 Symbolklassen definiert, und es sind knapp 100 Bilder pro Klasse im Korpus. Beim Training haben wir 20.000 Iterationen mit Batches von je 512 Bildern verwendet.

Die Resultate der Symbolerkennung werden zunächst evaluiert, visualisiert und anschließend für die MDA in MAXQDA aufbereitet. Die automatisch erstellten Annotationen dienen dabei nicht unmittelbar einem "close viewing". Vielmehr soll der quantitative Teil des Projekts helfen, gezielt Videos für die qualitative Analyse auszuwählen. Darüber hinaus interessiert uns eine statistische Auswertung des Symbolrepertoires im Korpus (z.B. Symbolverteilung über einen gewissen Zeitraum). Schließlich analysieren wir mithilfe der MDA ausgewählte Videos, um den Kontext, die Akteur:innen und ihre Rolle sowie die Interaktion verbaler und visueller Informationen zu untersuchen. Dabei konzentrieren wir uns auf die Fragen, welche allgemeinen Muster visueller und textueller (Selbst-)Repräsentationen von Protest erkennbar sind, welche Gemeinsamkeiten und Unterschiede es zwischen den einzelnen Ländern sowie den unterschiedlichen Medien gibt sowie auf die Frage, wie Protestsymbole und -slogans in den jeweiligen Medien (re-)kontextualisiert werden.

Ziele

Ziel unseres Projekts ist einerseits, ein Best-Practice-Beispiel für die Analyse eines großen visuellen Korpus zu liefern. Darüber hinaus erforschen wir visuelle und textuelle (Selbst-)Repräsentationen von Protestkulturen in Osteuropa und untersuchen, wie Bilder und Texte in verschiedenen Medien und Kontexten (wieder)verwendet werden. Schließlich stellt unser Projekt auch eine Momentaufnahme verschiedener osteuropäischer Protestkulturen dar und beantwortet die Frage, was von den Protesten nach Ablauf einer gewissen Zeit übrig bleibt: Insbesondere in autoritären Staaten wie Belarus und Russland ist nicht davon auszugehen, dass staatliche Medien die Erinnerung an regierungskritische Proteste bewahren.

Bibliographie

- Arnold, Taylor und Lauren Tilton. 2019. "Distant Viewing: Analyzing Large Visual Corpora." *DSH*, fqz013, <https://doi.org/10.1093/digitalsh/fqz013> (letzter Zugriff 26. 7. 2022).
- Cassirer, Ernst. 1996. *Versuch über den Menschen. Einführung in eine Philosophie der Kultur*. Hamburg: Meiner.
- Friedman, Jonathan. 2019. "Preface." In Makovsky, Nicolette; Trémon, Anne-Christine & Zandonai, Sheyla S.

(eds.): *Slogans: Subjection, Subversion, and the Politics of Neoliberalism*. London/New York: Routledge, xiii-xvii.

Gaufman, Elizaveta. 2021. "The Gendered Iconography of Belarus Protest." *New Perspectives* 29.1, 80–89.

Hayles, Nancy K. 2010. "How We Read: Close, Hyper, Machine." *ADE Bulletin* 150, 62–79.

Howanitz, Gernot und Erik Radisch. 2022. "Nationalist(ic) Symbols from Eastern Europe." *Github.com*, 21. 2. 2022 (v1.0.0). <https://zenodo.org/record/6206733#.YuEpDLuxWEI> (letzter Zugriff 26. 7. 2022).

Kress, Gunther. 2011: "Multimodal Discourse Analysis." In *The Routledge Handbook of Discourse Analysis*, hg. von James P. Gee und Michael Handford, 35–50. London u.a.: Routledge.

Kuße, Holger (Hg.). 2021. *Kommunikacija v #pochu protestov*. Berlin: Peter Lang.

Onuch, Olga. 2015. "EuroMaidan Protests in Ukraine: Social Media versus Social Networks." *Problems of Post-Communism* 62.4, 217–235

Smyth, Regina und Sarah Oates. 2015. "Mind the Gaps: Media Use and Mass Action in Russia." *Europe-Asia Studies* 67.2, 285–305.

Szostek, Joanna. 2018. "The Mass Media and Russia's 'Sphere of Interests': Mechanisms of Regional Hegemony in Belarus and Ukraine." *Geopolitics* 23.2, 307–329.

Klassifikation von Figurenauf- und -abtritten in XML-kodierten Dramen

Ehlers, Lena

st161358@stud.uni-stuttgart.de
Universität Stuttgart, Deutschland

Andresen, Melanie

melanie.andresen@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

In diesem Beitrag wird eine regelbasierte Methode vorgestellt, um Figurenauf- und -abtritte in den Regieanweisungen dramatischer Texte zu klassifizieren. In der Forschung wurde Regieanweisungen meist nur wenig Beachtung geschenkt, etwa weil sie während einer Theateraufführung nicht textuell in Erscheinung treten (Schößler, 2017, S.3). Als eine der wenigen quantitativen Untersuchungen stellen Trilcke et al. (2020) fest, dass sich die vermutete Episierung des Dramas im Laufe der Jahrhunderte am Korpus GerDraCor bestätigt. Eine Differenzierung nach Funktionen der Regieanweisungen erfolgt nicht.

Dabei ist die (Ko-)Präsenz von Figuren auf der Bühne eine häufig genutzte Grundlage für quantitative Dramenanalysen, insbesondere in der Netzwerkanalyse (z.B. Marcus 1973, Krautter et al. 2018, Fischer et al. 2018). Trilcke et al. (2017) konnten bereits am Beispiel von Lessings Emilia Galotti zeigen, dass in statischen Netzwer-

ken, die das ganze Drama auf einmal darstellen, wichtige Informationen zur Dynamik der Beziehungen zwischen den Figuren verloren gehen können. Unserer Ansicht nach ist außerdem zu berücksichtigen, dass Figuren auch innerhalb von Szenen auf- und abtreten und die Anwesenheit von zwei Figuren in einer Szene nicht zwangsläufig bedeutet, dass diese Figuren auch gleichzeitig auf der Bühne stehen.

Korpus und manuelle Annotation

Das Deutsche Dramenkorpus GerDraCor enthält über 550 deutschsprachige, TEI-kodierte Dramentexte aus dem Zeitraum von 1650 bis 1947 (Fischer et al. 2019). Die Regieanweisungen sind als *stage*-Elemente kodiert, die bisher keine weiteren Attribute besitzen, die Auskunft über den Inhalt der Regieanweisungen, wie z.B. das Auf- oder Abtreten von Figuren, geben würden.

Insgesamt wurden 16 Dramentexte manuell annotiert, wovon vier zur Implementierung¹ und zwölf zur Evaluation genutzt wurden. Die Guidelines für die manuelle Annotation stehen unter <https://doi.org/10.5281/zenodo.6951465> zur Verfügung. Die Annotation erfolgte direkt im XML-Format. Abbildung 1 zeigt, wie ein *stage*-Element mithilfe der Attribute *type* und *who* um die extrahierten Informationen erweitert wird.

```
<sp who="#wagner">
  <speaker>WAGNER.</speaker>
  <lg>
    <l>Ich hätte gern nur immer fortgewacht,</l>
    <l>Um so gelehrt mit Euch mich zu besprechen.</l>
    <l>Doch morgen, als am ersten Ostertage,</l>
    <l>Erlaubt mir ein' und andre Frage.</l>
    <l>Mit Eifer hab' ich mich der Studien beflissen;</l>
    <l>Zwar weiß ich viel, doch möcht' ich alles wissen.</l>
  </lg>
  <stage>Ab.</stage>
</sp>
<stage type="exit" who="#wagner">Ab.</stage>
```

Abbildung 1: Exemplarische Ergänzung eines *stage*-Elements

Regelbasierte Annotation

Das entwickelte Verfahren klassifiziert den unstrukturierten Text innerhalb der *stage*-Elemente mithilfe manueller erstellter Regeln. Diese basieren auf Schlüsselwörtern und -phrasen („treten herein“, „gehen ab“, ...) und nutzen reguläre Ausdrücke. Zusätzlich wird die Position einbezogen, etwa bei Regieanweisungen am Anfang einer Szene, die oftmals ausschließlich die Namen der anwesenden Figuren enthalten.

Wurde im ersten Schritt ein Auf- oder Abtritt erkannt, folgt als zweiter Schritt die Zuordnung der betroffenen Figuren. Hierfür wird die im XML enthaltene Liste der Namen aller sprechenden Figuren genutzt, die auch die Abbildung auf die Figuren-IDs ermöglicht. Der Auf- oder

Abtritt wird entweder einer in der Regieanweisung genannten Figur oder derjenigen Figur, deren Rede die Regieanweisung zugeordnet ist (vgl. Fall in Abb. 1), zugeschrieben.

Evaluation

Die Evaluation erfolgt anhand von zwölf manuell annotierten Texten, die nicht zur Aufstellung der Regeln herangezogen wurden. Evaluieren werden 1) die Klassifikation der Regieanweisungen in Figurenauf- und -abtritte und 2) die Zuordnung der betroffenen Figuren. Im zweiten Schritt werden nur die Elemente in die Evaluation einbezogen, die im ersten Schritt korrekt klassifiziert wurden. Tabelle 1 zeigt, dass die durchschnittlichen Werte für Precision, Recall und F1-Score für die Auf-/Abtritterkennung bei 0,85 liegen. Auch die Figurenerkennung liefert gute Ergebnisse (F1 = 0,87). Zwischen den Texten zeigt sich allerdings eine erhebliche Variation in der Qualität.

Tabelle 1: Evaluationsergebnisse

Auto- r*in	Titel	Jahr	Auf-/ Abtritt- erkennung	Figure- erkennung						
			P	R	F1	n	P	R	F1	n
Gottsched	Das Testament	1745	0,94	1	0,97	543	0,95	0,65	0,77	63
Schlegel	Canut	1746	1	1	1	23	1	1	1	23
Gellert	Die zärtlichen Schwes- tern	1747	0,95	0,98	0,97	129	0,99	0,99	0,99	63
Pfeil	Lucie Wood- vil	1756	0,95	0,97	0,96	141	1	0,9	0,95	78
Lenz	Der Hof- meister	1774	0,86	0,83	0,85	315	0,89	0,8	0,84	70
Schiller	Die Räuber	1781	0,8	0,79	0,79	544	0,86	0,74	0,79	75
Goethe	Die natü- rliche Toch- ter	1803	0,93	0,97	0,95	101	0,97	0,98	0,98	28
Kleist	Die Fa- milie Schrof- fenstein	1803	0,85	0,7	0,76	320	0,91	0,73	0,81	71
Gün- derode	Magie und Schick- sal	1805	0,93	0,95	0,94	84	0,98	0,88	0,93	37
Gün- derode	Udohla	1805	0,94	0,62	0,74	41	1	0,96	0,98	16
Wei- ßen- thurn	Das Manu- script	1817	0,7	0,89	0,79	675	0,85	0,72	0,78	64
Hof- manns- thal	Der Ro- senkavalier	1911	0,39	0,47	0,43	715	0,76	0,41	0,53	38
Mittel- werte			0,85	0,85	0,81	0,93	0,87			

Aufgrund des regelbasierten Verfahrens schneiden Texte, die stark von den zur Erstellung der Regeln verwendeten Texten abweichen, in der Evaluation schlechter ab. Besonders Texte mit langen Regieanweisungen sorgen dafür, dass viele Schlüsselwörter auch in anderen Kontexten vorkommen. Das zeigt sich insbesondere beim *Rosenkavalier*, dessen Regieanweisungen mit 12 Tokens im Mittel doppelt so lang sind wie der Durchschnitt aller Dramen. Ein weiteres Problem stellen von der Figu-

renliste abweichende Namen dar, etwa Varianten des Eigennamens oder Appellativa. Ersteres könnte in Zukunft durch die Nutzung von Ähnlichkeitsmaßen adressiert werden, die aber zu mehr Falsch-Positiven führen können.

Analyse

Abbildung 2 zeigt, dass die meisten Auf- und Abtritte tatsächlich innerhalb von Szenen stattfinden (ca. 46%). Etwas weniger erfolgen am Beginn einer Szene (41%) und etwa 13% am Ende. Erwartungsgemäß handelt es sich am Anfang der Szene fast ausschließlich um Auftritte, am Ende um Abtritte. Innerhalb der Szenen komme Auf- und Abtritte zu jeweils gleichen Anteilen vor.

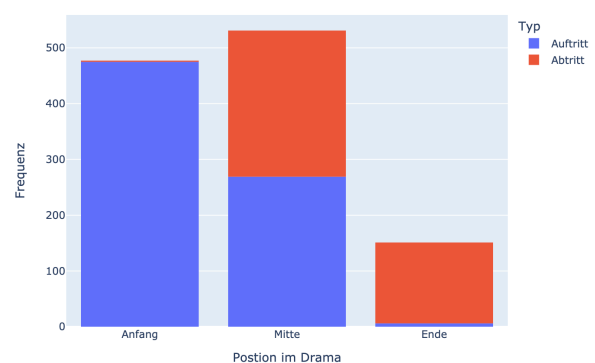


Abbildung 2: Verteilung der annotierten stage-Elemente auf den Beginn, die Mitte und das Ende einer Szene.

Fazit

In diesem Paper haben wir ein mit Figurenauf- und -abtritten annotiertes Teilkorpus zu GerDraCor präsentiert und einen regelbasierten Algorithmus vorgestellt, der diese Annotationen mit einem mittleren F1-Wert von über 0,85 reproduzieren kann. Ein Großteil der annotierten Auf- und Abtritte erfolgt innerhalb von Szenen. Diese Veränderungen in den Figurenkonstellationen werden bei einer szenenweisen Betrachtung der Figurenpräsenz nicht berücksichtigt, haben aber potenziell Auswirkungen auf beispielsweise netzwerkanalytische Arbeiten. Alle Daten und Skripte zu diesem Beitrag sind unter <https://github.com/quadrama/enter-exit> verfügbar.

Fußnoten

1. Es handelt sich um *Emilia Galotti* (Lessing, 1772), *Götz von Berlichingen* (Goethe, 1773), *Iphigenie auf Tauris* (Goethe, 1787) sowie *Maria Stuart* (Schiller, 1800).

Bibliographie

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hecht, Christopher Kittel, Carsten Milling, und Peer Trilcke. 2019. „Programmable Corpora – Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor“. In *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 194–97. <https://doi.org/10.5281/zenodo.2596095>.

Fischer, Frank, Peer Trilcke, Christopher Kittel, Carsten Milling, und Daniil Skorinkin. 2018. „To Catch a Protagonist: Quantitative Dominance Relations in German-Language Drama (1730–1930)“. In *DH 2018. Book of Abstracts*, 193–201. Mexiko City.

Krautter, Benjamin, Janis Pagel, Nils Reiter, und Marcus Willand. 2018. „Titelhelden und Protagonisten – interpretierbare Figurenklassifikation in deutschsprachigen Dramen“. *LitLab Pamphlets* 7 (November). https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf.

Marcus, Solomon. 1973. *Mathematische Poetik. Linguistische Forschungen* 13. Frankfurt a.M.: Athenäum.

Schöbeler, Franziska. 2017. *Einführung in die Dramenanalyse*. 2. Aufl. J. B. Metzler. <https://link.springer.com/book/10.1007/978-3-476-05285-8>.

Trilcke, Peer, Frank Fischer, Mathias Göbel, Dario Kampkaspar, und Christopher Kittel. 2017. „Netzwerkdynamik, Plotanalyse – Zur Visualisierung und Berechnung der ›progressiven Strukturierung‹ literarischer Texte“. In *4. Jahrestagung des DHd*, 175–80. Bern: Zenodo. <https://doi.org/10.5281/ZENODO.4622799>.

Trilcke, Peer, Christopher Kittel, Nils Reiter, Daria Maximova, und Frank Fischer. 2020. „Opening the Stage – A Quantitative Look at Stage Directions in German Drama“. In *Digital Humanities 2020 – Book of Abstracts*, 422–25. Ottawa, Kanada.

KoMuX – Der Kompositamuster-Explorer

Brunner, Annelen

brunner@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Deutschland

Katrin, Hein

hein@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Deutschland

KoMuX, der Kompositamuster-Explorer, (www.owi-d.de/plus/komux) ist eine Webanwendung, die es ermöglicht, mehr als 50.000 nominale Komposita des Deutschen gezielt nach abstrakten oder lexikalisch-teilspezifizierten Mustern zu durchsuchen. Unterschiedliche Visualisierungen helfen dabei, Strukturen und Zusammenhänge innerhalb der Ergebnismenge zu erfassen.

Mit KoMuX machen wir einen Teil der Datengrundlage frei verfügbar, auf der unsere empirischen Forschungen zur Wortbildung basieren und integrieren Analysen und Visualisierungen aus unseren Arbeiten. Der Explorer ist damit auch ein Beitrag zu OpenScience, indem er es ermöglicht, unsere Forschungsergebnisse in Teilen nachzuvollziehen und zu reproduzieren.

Forschungshintergrund

In der Wortbildungsforschung stellt die Einbeziehung authentischen Sprachmaterials nach wie vor ein Desiderat dar (vgl. z.B. Hein ersch. 2023; Elsen und Michel 2007). KoMuX ermöglicht es, Untersuchungen zur Komposition auf eine breite empirische Basis zu stellen und einer ‚empirischen Wortbildungsforschung‘ somit ein Stück weit näher zu kommen. Der Explorer basiert auf einer systematischen Datenerhebung, bei der alle nominalen Komposita automatisch aus dem KoGra-Untersuchungskorpus (KoGra 2022), einem Ausschnitt des Deutschen Referenzkorpus DeReKo (Kupietz u. a. 2010), extrahiert wurden. Diese Datengrundlage ist unseres Wissens nach die erste ihrer Art.

Mit KoMuX wird erstmals eine Untermenge dieses Komposita-Inventars des Deutschen frei zugänglich und systematisch durchsuchbar gemacht, und zwar aus einer Muster-Perspektive (vgl. Stein und Stumpf 2019) heraus: Wir betrachten Komposita als konkrete sprachliche Realisierungen von zugrundeliegenden abstrakten oder lexikalisch-teilspezifizierten Mustern. Diese Muster aus spezifischen Paarungen von Erst- und Zweitgliedern wiederum können z.B. zur Erklärung von beobachtbaren Produktivitätsunterschieden herangezogen (vgl. Hein und Brunner 2020; Brunner u. a. 2021) oder – ganz allgemein – als Grundprinzip verstanden werden, das erklärt, wie die Komposition funktioniert bzw. wie sich das Inventar von Komposita grundsätzlich systematisieren lässt (vgl. Hein ersch. 2023). Der Musteransatz bietet darüber hinaus eine direkte Anschlussfähigkeit an Grammatiktheorien wie die Konstruktionsgrammatik bzw. die Construction Morphology (Booij 2010).

Datengrundlage

Das KoGra-Untersuchungskorpus umfasst ca. 7 Milliarden Tokens und besteht zum größten Teil (~90%) aus Presstexten (zur genauen Zusammensetzung vgl. KoGra 2022; Bubenhofer, Konopka und Schneider 2014). Es wurde mit einem automatischen Werkzeug annotiert, welches die Canoo Language Tools adaptiert, und für jedes Token detaillierte morphologische Informationen liefert, auf deren Basis nominale Komposita extrahiert wurden. KoMuX basiert auf einer Untermenge von 100.000 Komposita-Tokens, die zufällig aus der Gesamtmenge von ca. 489 Millionen Komposita-Tokens gezogen wurden. Daraus ergibt sich die Frequenzliste mit ca. 50.000 Komposita-Types, die durchsucht werden kann.

Die automatischen morphologischen Analysen wurden manuell und semi-automatisch verbessert. Dies umfasste v.a. das Entfernen von falschen Einträgen (Tokens ohne Komposita-Status) sowie Korrekturen von falschen

dings nicht der morphologische Analysierer (vgl. https://dict.leo.org/pages/about/ende/canoonet_de.html).

Bibliographie

Booij, Geert E. 2010. *Construction morphology*. Oxford: Oxford University Press.

Brunner, Annelen, Stefan Engelberg und Katrin Hein. 2021. „The distribution of constituent words in nominal compounds and its impact on semantic interpretation: an empirical study“. *Journal of Word Formation* 1: 7–36.

Bubenhofer, Noah, Marek Konopka und Roman Schneider. 2014. *Präliminarien einer Korpusgrammatik. Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache* 4. Tübingen: Narr.

Elsen, Hilke und Sascha Michel. 2007. „Wortbildung im Sprachgebrauch. Desiderate und Perspektiven einer etablierten Forschungsrichtung“. *Muttersprache* 117: 1–16.

Hein, Katrin. 2023. „Auf dem Weg zu einem Komposita-Konstruktion? – ein empirischer Anwendungsversuch der Construction Morphology auf die Nominalkomposition im Deutschen“. In *Konstruktionsfamilien im Deutschen*, hg. von Fabio Mollica und Sören Stumpf. Tübingen: Stauffenburg.

Hein, Katrin und Annelen Brunner. 2020. „Why do some lexemes combine more frequently than others? – An empirical approach to productivity in German compound formation“. In *Rules, patterns, schemas and analogy. Online Proceedings of the 12th Mediterranean Morphology Meetings (MMM12)* 12: 28–41.

KoGra. 2022. „Korpus des Projekts Korpusgrammatik“. In *Leibniz-Institut für Deutsche Sprache: „Korpusgrammatik“. Grammatisches Informationssystem grammis*. <https://grammis.ids-mannheim.de/korpusgrammatik/6615>.

Kupietz, Marc, Cyril Belica, Holger Keibel und Andreas Witt. 2010. „The German Reference Corpus DeReKo: A primordial sample for Linguistic Research“. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner und Daniel Tapias: 1848–54. Malta: European Language Resources Association (ELRA).

Stein, Stephan und Sören Stumpf. 2019. *Muster in Sprache und Kommunikation. Eine Einführung in Konzepte sprachlicher Vorgeformtheit*. Berlin: Erich Schmidt.

Metaphors of Religion

Gebhard, Henning

henning.gebhard@rub.de
Ruhr Universität Bochum, Deutschland

Jha, Vandana

vandana.jha@kit.edu
Karlsruhe Institute for Technology, Deutschland

Tögel, Philipp

philipp.toegel@kit.edu
Karlsruhe Institute for Technology, Deutschland

Dipper, Stefanie

stefanie.dipper@rub.de
Ruhr Universität Bochum, Deutschland

Elwert, Frederik

frederik.elwert@rub.de
Ruhr Universität Bochum, Deutschland

Tonne, Danah

danah.tonne@kit.edu
Karlsruhe Institute for Technology, Deutschland

Towards a Shared Infrastructure for Metaphor Analysis¹

The collaborative research center (CRC) 1475² studies the role of metaphors in religious meaning-making. In metaphors, meaning is transferred from one semantic domain to another. Metaphors can thus serve as a means to express abstract concepts with reference to more concrete ones closer to human experience. Religion, which cannot directly address its ultimate subject (the transcendent, i.e., gods, otherworlds, etc.), is especially dependent on this procedure. By adopting conceptual metaphor theory (Lakoff and Johnson, 1980; Steen et al., 2010; Nacey et al., 2019) the CRC seeks to more thoroughly understand this process theoretically and grasp it methodologically to research its semantic forms empirically and comparatively. Through its multidisciplinary subprojects the CRC contributes to the historiography of religions and to answering systematic questions in the comparative study of religions. It covers a plethora of religious traditions from across the globe, including Christianity, Islam, Judaism, Zoroastrianism, Jainism, Buddhism, and Daoism. The time frame ranges from 3,000 BCE to the present day, including texts from multiple languages and diverse genres, from Korean Confucian ego-documents to Christian online forums.

Motivated by the challenge of comparability and interoperability between its extremely heterogeneous subprojects, the CRC deliberately puts emphasis on digital methods. Thus, the shared infrastructure, provided by the information infrastructure (INF) project, does not only support the individual research projects, but fosters conceptual integration. Utilizing this infrastructure the subprojects annotate religious texts to make metaphorical language explicit. For this process we adapt the Five Step Method (Steen, 2011) to not only mark the presence of metaphors, but to include complex analysis of the structural functioning of the metaphor and the resulting domain mappings as well. As a standardization measure we append an additional step, where each concept is linked to a conceptual thesaurus.

The Metaphor Workbench

Within the INF project, scholars of religion, computational linguists, and computer scientists jointly establish the digital research infrastructure of the CRC and provide the necessary tools for the subprojects.

Shared text repository

An instance of the KIT Data Manager Base Repo (Jejkal et al., 2014) is used as a research data repository for all CRC subprojects. Existing data from the subprojects is stored as structured data objects including their respective metadata in the form of TEI compliant XML files (Burnard and Bauman, 2010) and is available for further processing, enrichment, analysis, etc. via standardized interfaces. In close collaboration with the subprojects, the INF project is evaluating existing representations of their data, required format conversion, legal aspects and rights management, as well as assisting the provision of required descriptive metadata.

Annotation services

To annotate the metaphors the INF team provides tools for annotation implementing a shared metaphor annotation schema, which is based on the web annotation data model (Young, Sanderson, and Ciccicarese, 2017). Because the available annotation tools (like INCEpTION, CATMA, WissKI etc.) are lacking the possibility to create the complex annotations needed for the CRC's methodology, a new metaphor analysis tool will be developed, which guides and documents the interpretative analysis process. In the future, we will be using NLP expertise present in the CRC's subprojects to integrate methods of (semi-)automatic metaphor detection and analysis.

Furthermore, the INF project is advising the subprojects that aim for additional, project-specific annotation of their data, particularly with regard to the use of existing annotation standards and best practices. All of the annotations are provided in a Web Annotation Protocol Server (Tonne et al., 2019) as RDF triples to foster analysis across the subprojects and to ensure interoperability and reuse.

Conceptual thesaurus

To facilitate comparability of our metaphor annotations across barriers of languages and cultural traditions, the CRC is developing a conceptual thesaurus (CT) as a shared reference system. Its taxonomy is based on the Historical Thesaurus of English (Kay, 2009), albeit extending and adapting it for the languages and topics prevalent in the CRC. Using the SKOS data model (Miles and Bechhofer, 2009) and principles from Linked Open Data (LOD; Berners-Lee, 2006), the CT will provide a language-independent framework for the annotation of domains used in metaphorical mappings. Linking concrete metaphorical expressions with a central semantic resource enables retrieving conceptually related meta-

phors from different corpora, and comparatively studying semantic domains used in metaphors.

Thesaurus of religious metaphors

Linking texts, analysis, and concepts, the resulting annotations will make up the thesaurus of religious metaphors (TRM). The TRM will enable studies of metaphors in a systematic and comparative way by providing a semantically indexed collection of religious metaphors, as well as query and analysis tools.

Conclusions and Future Scope

The CRC's infrastructure – research data repository, annotation services, conceptual thesaurus and thesaurus of religious metaphors – fosters reusability as well as interoperability with external knowledge graphs by focusing on open data principles and will be published under open licenses. In particular, the emerging TRM will act as a unique resource for scholars worldwide studying religious metaphors. The INF project itself is an integral part of the CRC in providing this shared infrastructure and thus enabling comparative studies on an unprecedented scale in the field.

Fußnoten

1. Contributor Roles: Henning Gebhard (Writing – original draft; Writing – review & editing; Data Curation; Software), Vandana Jha (Writing – original draft; Writing – review & editing; Data Curation; Software), Philipp Tögel (Writing – original draft; Writing – review & editing; Data Curation; Software), Stefanie Dipper (Writing – review & editing; Conceptualization; Funding acquisition; Methodology; Supervision), Frederik Elwert (Writing – review & editing; Conceptualization; Funding acquisition; Methodology; Supervision), Danah Tonne (Writing – review & editing; Conceptualization; Funding acquisition; Methodology; Supervision).
2. Funded by the German Research Association (DFG, Project ID 441126958), CRC 1475 “Metaphors of Religion” consists of researchers from the Ruhr University Bochum and the Karlsruhe Institute of Technology in 14 scientific subprojects.

Bibliographie

- Berners-Lee, Tim. 2006. “*Linked Data*.” July 27, 2006. <https://www.w3.org/DesignIssues/LinkedData.html>.
- Burnard, Lou, and Syd Bauman, eds. 2010. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 1.9.1, March 2011. Charlottesville: TEI Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.
- Jejkal, Thomas, Alexander Vondrous, Andreas Kopmann, Rainer Stotzka, and Volker Hartmann. 2014. “*KIT Data Manager: The Repository Architecture Enabling Cross-Disciplinary Research*.” Large-Scale Data Manage-

ment and Analysis (LSDMA) - Big Data in Science. Hrsg.: Ch. Jung, 9.

Kay, Christian, ed. 2009. *Historical Thesaurus of the Oxford English Dictionary: With Additional Material from "A Thesaurus of Old English."* Oxford; New York: Oxford University Press.

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago [u.a.]: Univ. of Chicago Press.

Miles, Alistair, and Sean Bechhofer. 2009. "SKOS Simple Knowledge Organization System Reference." 2009. <https://www.w3.org/TR/skos-reference/>.

Nacey, Susan, Aletta G. Dorst, Tina Krennmayr, and W. Gudrun Reijnierse, eds. 2019. *Metaphor Identification in Multiple Languages: MIPVU around the World*. Converging Evidence in Language and Communication Research 22. Philadelphia: John Benjamins Publishing Company.

Steen, Gerard J. 2011. "From Three Dimensions to Five Steps: The Value of Deliberate Metaphor." *Metaphorik.de* – Online-Journal Zur Metaphorik in Sprache, Literatur, Medien.

Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Vol. 14. Converging Evidence in Language and Communication Research 14. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tonne, Danah, Germaine Götzelmann, Philipp Hegel, Michael Krewet, Julia Hübner, Sibylle Söring, Andreas Löffler, Michael Hitzker, Markus Höfler, and Timo Schmidt. 2019. "Ein Web Annotation Protocol Server Zur Untersuchung Vormoderner Wissensbestände." https://dhd-boas-app.acdh-dev.oeaw.ac.at/pages/show.html?document=TONNE_Danah_Ein_Web_Annotation_Protocol_Server_zur_Untersuchung.xml.

Young, Benjamin, Robert Sanderson, and Paolo Ciccarese. 2017. "Web Annotation Data Model." 2017. <https://www.w3.org/TR/annotation-model/>.

Netzwerk Offenes Mittelalter

Borek, Luise

luise.borek@tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Busch, Hannah

hannah.busch@huygens.knaw.nl
Huygens Instituut

Ketschik, Nora

nora.ketschik@ilw.uni-stuttgart.de
Universität Stuttgart

Die germanistische Mediävistik hat über Jahrzehnte viele digitale Ressourcen aufbereitet, die zu etablierten und zentralen Säulen des Faches gewachsen sind. Das

DFG-Netzwerk *Netzwerk Offenes Mittelalter* widmet sich dem weiteren Erschließungspotenzial dieser heterogenen und historisch gewachsenen Forschungsdaten, ihrer Vernetzung und ihrer qualitativen Verdichtung und exploriert dabei den Einsatz von Linked Open Data (LOD). Das Potenzial und die Grenzen dieser Verfahren werden forschungsorientiert erprobt und methodisch reflektiert.

Unter ‚offenes Mittelalter‘ verstehen wir dabei nicht nur die für LOD ohnehin erforderliche Einhaltung der FAIR-Prinzipien, sondern auch eine disziplinäre wie epochenübergreifende Durchlässigkeit, die den Austausch mit den relevanten Communitys und Gedächtnisinstitutionen, aber auch die Übertragbarkeit der Methoden widerspiegelt.

Das Netzwerk liefert aus einer fachdisziplinären Perspektive heraus Impulse für den Einsatz von LOD in geisteswissenschaftlicher Forschung und befindet sich dabei in Austausch mit der Nationalen Forschungsdateninfrastruktur (NFDI), insbesondere den Konsortien Text+ und NFDI4Culture.

Das Netzwerk setzt sich zusammen aus 18 Wissenschaftler*innen aus Deutschland, Österreich und den Niederlanden. Die individuellen Hintergründe sind vielfältig, sodass neben verschiedener mediävistischer Expertisen auch Informationswissenschaften und Bildungsforschung vertreten sind. Über die einzelnen Forschungsaktivitäten und -projekte sind zudem zentrale Ressourcen der germanistischen Mediävistik unmittelbar eingebunden, etwa der Handschriftencensus, das Mittelhochdeutsche Wörterbuch, die Mittelhochdeutsche Begriffsdatenbank, das Referenzkorpus Mittelhochdeutsch, das Mittelalterblog und diverse einschlägige Editionen (s. KONDE).

Ergänzt und erweitert wird das Netzwerk durch Expert*innen, die als Gäste wertvolle Impulse einbringen und Austausch und Vernetzung vorantreiben. Für diesen wichtigen Aspekt des Community-Building haben wir zudem die Beteiligungsmöglichkeit als „Assoziierte Mitglieder“ geschaffen, um das Netzwerk fruchtbar zu erweitern und der fachlichen und methodischen Diversität gerecht zu werden.

Konzeptionell beleuchtet das Netzwerk die Verfahren aus verschiedenen Perspektiven. Hierzu gehört ein editonsphilologischer Fokus: Als etablierter Standard für die Kodierung von Texten bildet TEI-XML die Grundlage vieler digitaler Editionen und dient als Archivformat (vgl. Wettlaufer 2018). Darin sind Informationen zunächst implizit erfasst, d. h. menschenlesbar, aber nicht ‚semantisch‘ nach Kriterien des Semantic Web (vgl. Hitzler 2021). Hier wurden etwa Möglichkeiten des Transfers eruiert (z. B. XTriples, Addelee 2019), aber auch Herausforderungen in diesem Bereich diskutiert z. B. im Umgang mit Unsicherheiten und Ambiguitäten, wie sie besonders beim Umgang mit historischen Ressourcen häufig auftreten und zurzeit von besonderem Forschungsinteresse sind (Kuczera 2019; Andrews 2021-2026). Eine weitere wichtige Säule für LOD bilden persistente Identifier. Für eine forschungsgetriebene Anwendung, die über basale Metadatenkategorien hinausgeht, ergeben sich verschiedene Herausforderungen. Hier befindet sich das Netzwerk in Austausch mit der GND, beleuchtet Entwicklungen zu Normdatenstandards (z. B. Burrows et al. 2020) und exploriert zudem die Einbindung verschiedener Wikisysteme wie Wikidata und Factgrid.

Eine wichtige Rolle spielt auch das Spannungsfeld von Materialität und Text, in dem nicht nur die Stellvertreterfunktion digitaler Objekte beachtet werden, sondern auch Möglichkeiten zu domänenspezifischen Vokabularen und granularen Identifiern gegeben sein müssen. Die Materialität stellt einen disziplinenübergreifenden ‚Berührungspunkt‘ dar, der sowohl eine Schnittstelle für LOD bilden kann als auch große Herausforderungen und Chancen für eine Erfahrbarkeit im digitalen Raum bietet.

Die skizzierten Verfahren sind nur tragfähig, wenn für sie eine weitreichende Akzeptanz besteht und geeignete Forschungsinfrastrukturen vorhanden sind. Das Netzwerk bemüht sich daher um eine Methodenzusammenchau, liefert Best Practices und erstellt Showcases, die einen ertragreichen Einsatz von LOD illustrieren. Zu den Ergebnissen des Netzwerks gehört auch eine Wissensplattform, die eine Ressourcensammlung bietet, in der einschlägige Ressourcen zu LOD, Tutorials und Projekte gelistet werden, eine domänenspezifische Bibliographie zu LOD in der germanistischen Mediävistik gepflegt wird und die Showcases aus der Forschung der Netzwerkmitglieder präsentiert werden. Flankiert werden diese Ressourcen zudem von der ausführlichen Dokumentation der Aktivitäten des Netzwerks über Blogbeiträge (z. B. Borek et al. 2022), Meldungen auf der Website des Netzwerks und über Präsenz und Vernetzung über Social Media.

Mit unserem Poster informieren wir über unsere bisherigen Ergebnisse und Aktivitäten des *Netzwerks Offenes Mittelalter* und laden ein zu weiterem Austausch, um nachhaltige und innovative Forschungsinfrastrukturen in den digitalen Geisteswissenschaften gemeinsam mitzugestalten.

Bibliographie

Addlesee, Angus. 2019. “Using OntoRefine to Transform Tabular Data into Linked Data”, <https://medium.com/wallscope/using-ontorefine-to-transform-tabular-data-into-linked-data-7277ec8c2c0f> (zugegriffen: 28.07.2022).

Andrews, Tara. 2021-2026. “Re-Evaluating the Eleventh Century with Linked Events and Entities”. ERC-Projekt an der Universität Wien.

Borek, Luise, Katharina Zeppezauer-Wachauer und Nora Ketschik. 2022. “Eindeutig Uneindeutig. Zur Modellierung von Unschärfe in der Mediävistik”, In *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*. <https://mittelalter.hypotheses.org/27658> (zugegriffen: 28.07.2022).

Burrows, Toby, Antoine Brix, Doug Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Synnøve Myking, Kevin Page, Lynn Ransom, Emma Cawfield Thomson, Jouni Tuominen, Hanno Wijsman und Pip Willcox. 2020. “Linked Open Data vocabularies and identifiers for medieval studies.” In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN)*. 2612: 211-218.

CoReMA – Cooking Recipes of the Middle Ages, <https://gams.uni-graz.at/context:corema> (GAMS. 562.10).

FactGrid – a database for historians: <https://blog.factgrid.de/archives/1143> (zugegriffen: 27.07.2022).

FAIR Principles: <https://www.go-fair.org/fair-principles/> (zugegriffen: 28.07.2020).

Hitzler, Pascal. 2021. “A Review of the Semantic Web Field.” In *Communications of the ACM*. 64 (2): 76-83. 10.1145/3397512. <https://cacm.acm.org/magazines/2021/2/250085-a-review-of-the-semantic-web-field/fulltext> (zugegriffen: 28.07.2022).

KONDE – Kompetenznetzwerk Digitale Editionen: <http://www.digitale-edition.at/> (zugegriffen: 21.07.2022).

Kuczera, Andreas, Thorsten Wübbena, Thomas Kolatz (Hg.). 2019. “Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten.” *Wolfenbüttel. (Sonderband der Zeitschrift für digitale Geisteswissenschaften 4)* DOI: 10.17175/sb004.

MHDBDB – Mittelhochdeutsche Begriffsdatenbank: <http://mhdadb.sbg.ac.at> (zugegriffen: 21.07.2022).

Mittelalterblog: <https://mittelalter.hypotheses.org/> (zugegriffen: 21.07.2022).

Netzwerk Offenes Mittelalter (2021-2024). DFG Netzwerk.

OxGarage Conversion: <https://oxgarage.tei-c.org/>

Wettlaufer, Jörg. 2018. “Der nächste Schritt? Semantic Web und digitale Editionen.” In *Digitale Metamorphose: Digital Humanities und Editionswissenschaft (Sonderband der Zeitschrift für digitale Geisteswissenschaften 2)* DOI: 10.17175/sb002_007.

Wikidata: <https://www.wikidata.org> (zugegriffen: 28.07.2022).

XTriples: <https://xtriples.lod.academy/index.html> (28.07.2022).

NFDI4Culture und Text+ – Kartierung einer Zusammenarbeit

Schrade, Torsten

Torsten.Schrade@adwmainz.de

Akademie der Wissenschaften und der Literatur | Mainz

Stein, Regine

regine.stein@sub.uni-goettingen.de

Niedersächsische Staats- und Universitätsbibliothek Göttingen

Tolksdorf, Julia

Julia.Tolksdorf@adwmainz.de

Akademie der Wissenschaften und der Literatur | Mainz

Vater, Christian

Christian.Vater@adwmainz.de

Akademie der Wissenschaften und der Literatur | Mainz

Weimer, Lukas

weimer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek
Göttingen

Einführung

Werden kulturelle Daten nachhaltig gespeichert, sind sie die Basis nicht nur der heutigen, sondern auch zukünftiger Wissenschaftsgenerationen. In der Gegenwart geschieht dies typischerweise digital und auch – in Hinblick auf kulturelle Teilhabe und kollaborative Partizipation – offen (Schöch 2017).

Diesem Grundsatz der Offenheit haben sich die beiden Konsortien der Nationalen Forschungsdateninfrastruktur NFDI4Culture und Text+ – wie auch die gesamte NFDI – verschrieben. Vor dem Hintergrund des gemeinsamen Wissenschaftsbereiches und ihrem Fokus auf Sprach-, Text- und Kulturdaten wird diese Offenheit auch in der intensiven Zusammenarbeit der beiden Konsortien gelebt. Die Zusammenarbeit erfolgt hierbei in einem communitygestützten dynamischen Prozess, in den auch verwandte weitere Konsortien und Konsortialinitiativen eingebettet werden. So können Bedarfe innerhalb der gesamten NFDI mit einer Stimme artikuliert werden. Umso wichtiger ist es, einen Überblick über beteiligte Akteursgruppen zu erhalten und in die (Fach-)Öffentlichkeit kommunizierbar zu machen. Dazu bedürfen die vorhandenen Datenpunkte nicht nur der Vernetzung, sondern auch der Visualisierung. Dieser zweiteiligen ganz praxisorientierten Fragestellung – (1) Wie sammle ich meine Akteursdaten und (2) wie werte ich diese graphisch passend aus? – widmet sich das hier vorgeschlagene Posterprojekt.

Hintergrund: NFDI und das Memorandum of Understanding

Basierend auf einem Bund-Länder-Beschluss 2018 (Bundesanzeiger 2018) hat die Nationale Forschungsdateninfrastruktur (NFDI) zum Ziel, Datenbestände entlang der FAIR-Prinzipien (Wilkinson et al. 2016) zu erschließen und langfristig zu sichern. Sie wird dabei "in einem aus der Wissenschaft getriebenen Prozess als vernetzte Struktur eigeninitiativ agierender Konsortien aufgebaut" (DFG 2020). Zwei dieser bislang 19 geförderten Konsortien sind NFDI4Culture und Text+. Mit den beiden NFDI-Initiativen NFDI4Memory und NFDI4Objects haben sie sich 2019 in einem Memorandum of Understanding (Brünger-Weilandt 2020) zusammengeschlossen, um die Bedarfe der Geistes- und Kulturwissenschaften gemeinsam zu bearbeiten und organisatorische und technische Lösungen für deren Fragestellungen zu finden. Diese Interdisziplinarität bietet die Chance, übergreifende Angebote zu entwickeln und die Vision von Open Humanities voranzutreiben – für die Wissenschaft, GLAM-Einrichtungen und die breite Öffentlichkeit.

Zusammenarbeit von NFDI4Culture und Text+

NFDI4Culture und Text+ vertreten äußerst vielfältige und disziplinär diverse (Fach-) Communities. Gleichzeitig gibt es mit Blick auf die an den beiden Konsortien beteiligten Institutionen und Einzelpersonen interessante Schnittmengen.

Eine genauere Untersuchung von Verbindungen, Rollen und gemeinsamen aber auch unterschiedlichen Handlungsebenen der an den beiden geistes- und kulturwissenschaftlichen NFDI-Konsortien beteiligten Fachcommunities, Institutionen und Einzelpersonen steht bislang noch aus. Gleichzeitig existiert mit der strukturierten Erfassung der NFDI im Rahmen eines Wikidata-Projektes (vgl. https://www.wikidata.org/wiki/WikiProject_NFDI) in Zusammenarbeit zwischen NFDI-Direktorat und -Konsortien eine erste Datenbasis für die Analyse. Hinzu kommen verfügbare Daten aus den beiden Internetportalen von NFDI4Culture und Text+, die zusätzliche Strukturfacetten liefern.

Die Daten von NFDI4Culture können über eine API und einen SPARQL-Endpoint, den Culture Knowledge Graph, abgefragt werden. Die Informationen zu Personen, Institutionen, Projekten, Nachrichten, Veranstaltungen, Forschungsprodukten und Services liegen als Linked Data in den Formaten Turtle, JSON-LD und RDF/XML vor. Text+seitig liegen entsprechende Daten in tabellarischer Form vor, die dann mit den Daten von NFDI4Culture verschnitten werden. Ein konkretes Beispiel für unsere Datenauswertung ist die Identifikation von Akteuren in der NFDI, die (a) gleichzeitig Mitglieder in Text+ und NFDI4Culture sind oder (b) in Sektionen/AGs/Task Forces der NFDI gleichzeitig präsent sind. Zusätzlich sollen auch Zugangspunkte für die Communities visualisiert werden sowie deren Möglichkeiten zur Beteiligung.

Das gemeinsame Poster stellt die grafische Auswertung des gemeinsamen Akteurs-Netzwerkes der beiden Konsortien NFDI4Culture und Text+ vor. Hierbei werden die diagrammatischen Möglichkeiten der sozialen Netzwerkanalyse durchgespielt (vgl. Drucker 2014), die über die tabellierten und 'graphierten' Daten der Akteursmatrix gelegt werden. Dabei werden – wo möglich und praktikabel – Daten aus den jeweiligen konsortialen Wissensgraphen und den Wikidata-Identifikatoren (<http://www.wikidata.org/entity/Q98276929>, <http://www.wikidata.org/entity/Q98271443>) verwendet. Der Arbeitsprozess ist aufgrund der dynamischen Entwicklungen in beiden Konsortien explorativ und dient auch dem Versuch zu fassen, was mit datengetriebenen diagrammatischen Methoden 'neu' erkannt werden kann (vgl. Gold 2012).

Berücksichtigt werden auch gemeinsame Veranstaltungen, gemeinsam genutzte Dienste und/oder Repositorien, gemeinsame Arbeit in NFDI-Sektionen etc.

Bibliographie

Brünger-Weilandt, Sabine, Kai-Christian Bruhn, Alexandra W. Busch, Erhard Hinrichs, Gerald Maier, Johannes Paulmann, Andrea Rapp, Philipp von Rummel, Eva

Schlotheuber, Dörte Schmidt, Torsten Schrade, Holger Simon, Regine Stein und Elke Teich. 2020. "Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies." *Zenodo*. <https://doi.org/10.5281/zenodo.4045000>.

Bundesanzeiger. 2018. *Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018*.

Deutsche Forschungsgemeinschaft DFG. 2020. *Nationale Forschungsdateninfrastruktur – Ausschreibung 2020 für die Förderung von Konsortien (2. Ausschreibungsrunde)*. https://www.dfg.de/foerderung/info_wissenschaft/2020/info_wissenschaft_20_29/index.html.

Drucker, Johanna. 2014. *Graphesis. Visual Forms of Knowledge Production*. Cambridge (MA): Harvard University Press.

Gold, Matthew K. (Hg.). 2012. *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Schöch, Christof. 2017. "Aufbau von Datensammlungen." In *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein, 223–233. Stuttgart: Metzler.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data* 3: 160018. <https://doi.org/10.1038/sdata.2016.18>.

Offene Editionen – Die Task Area Editionen im NFDI-Konsortium Text+

Blumtritt, Jonathan

jonathan.blumtritt@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities

Cugliana, Elisa

elisa.cugliana@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities

Geißler, Nils

nils.geissler@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities

Hegel, Philipp

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

Hensen, Kilian

kilian.hensen@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities

Hörnschemeyer, Jörg

hoernschemeyer@dhi-roma.it
Deutsches Historisches Institut Rom

Kudella, Christoph

kudella@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Lemke, Karoline

karoline.lemke@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Lordick, Harald

lor@steinheim-institut.org
Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte an der Universität Duisburg-Essen

Neuber, Frederike

frederike.neuber@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Neuefeind, Claes

c.neuefeind@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities

Schulz, Daniela

schulz@hab.de
Herzog August Bibliothek Wolfenbüttel

Seltmann, Melanie Elisabeth-H.

melanie.seltmann@tu-darmstadt.de
Universitäts- und Landesbibliothek Darmstadt

Sievers, Martin

martin.sievers@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz

Gengnagel, Tessa

tessa.gengnagel@uni-koeln.de
Universität zu Köln, Cologne Center for eHumanities

Das geisteswissenschaftliche Konsortium Text+ (<https://www.text-plus.org/>) hat im Oktober 2021 seine Arbeit innerhalb der Nationalen Forschungsdateninfrastruktur (NFDI) aufgenommen. Text+ widmet sich text- und sprachbasierten Daten aus den verschiedensten Disziplinen (u. a. Literaturwissenschaft, Sprachwissenschaft, Geschichtswissenschaft, Philosophie) in drei Datendomänen bzw. Task Areas: Lexikalische Ressourcen, Sammlungen und Editionen. Als Vertreter:innen der Task Area Editionen¹ möchten wir mit einem Poster einen Einblick in unsere Arbeit geben, die von einem vielschichtigen Verständnis der ‚Offenheit‘ digitaler Editionen geleitet ist.

Im Kontext digitaler Editionen wird ‚Offenheit‘ meist im Zusammenhang mit Rechtfragen, d. h. Lizenzen und speziell Open Access, diskutiert (Hanneschläger 2019;

Sichani 2017; Dillen und Neyt 2016). Zentral für die Öffnung einer Edition nach außen ist darüber hinaus ihre technische Vernetzbarkeit; Editionen aggregieren einerseits Daten über Schnittstellen, und machen über diese auch ihre eigenen Daten abrufbar (Witt 2018, 255). Prinzipiell können digitale Editionen auch durch die Möglichkeit ihrer weiteren Bearbeitung, Anreicherung oder Erweiterung "offen" bleiben. Darüber hinaus gibt es in jüngster Zeit vermehrt Ansätze, 'Offenheit' im Sinne von 'Zugänglichkeit' aus 'sozialer' Perspektive zu betrachten (Martinez et al. 2019; Rojas-Castro 2020). Eine Übersicht verschiedener Facetten von 'Offenheit' schaffen Jeffrey Pomerantz und Robin Peek (2016), indem sie den Begriff 'open' u. a. in Bezug zu den Themen Zugänglichkeit, Nutzung, Partizipation und Transparenz stellen, die – gemeinsam mit den bereits genannten Facetten von Offenheit – im Rahmen der Task Area Editionen die Zieldimensionen mit vorgeben. Zentrale Maßnahmen der Task Area Editionen zur Förderung der 'Offenheit' digitaler Editionen umfassen u. a.:

1. Empfehlungen für 'FAIRer' Editionen

Ein Ziel von Text+ ist es, die Anwendung der FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016) zu fördern, weshalb die Task Area Leitlinien zur Erstellung und Publikation 'FAIRer' Editionen erarbeitet. Da die Anwendung der FAIR-Prinzipien bisher im Editions-kontext noch nicht tiefergehend diskutiert wurde, zugleich aber wichtige Bereiche wie Auffindbarkeit, Vernetzung, Lizenzierung und Nachnutzung betrifft, wurde eine Kooperation mit dem Rezensionenjournal RIDE des Instituts für Dokumentologie und Editorik (IDE 2014–2022) initiiert. Über einen Call for Reviews (<https://ride.i-d-e.de/reviewers/call-for-reviews/ride-textplus-de/>), der sich neben dem Kriterienkatalog des IDE (Sahle 2014) auch an im Rahmen von Text+ entwickelten FAIR-Kriterien (Gengnagel et al. 2022) orientiert, werden Rezensionen gesammelt, anhand derer evaluiert wird, inwieweit die FAIR-Prinzipien in ihrer aktuellen Formulierung auf digitale Editionen anwendbar sind und wie sie bisher umgesetzt werden. Die ersten Rezensionen aus der Kooperation zwischen Text+ und IDE erscheinen Anfang 2023.

2. Verzeichnis digitaler und gedruckter Editionen

Bei der sog. Registry handelt es sich um ein kuratiertes Verzeichnis von Editionen. Es soll einen strukturierten Zugriff auf die große Zahl an vorhandenen Ressourcen bieten und neben allgemeinen Zugängen nach bestimmten Kriterien (z.B. Sprachen des Edendums, Disziplin-zugehörigkeit, Projektbeteiligte) auch erstmals die FAIRness digitaler Editionen berücksichtigen. Durch den holistischen Nachweis von Editionen unabhängig von ihrer Medienform, Verknüpfungen zu einer – ebenfalls im Kontext von Text+ entstehenden – Software-Registry sowie der Präsentation von Best-Practice-Beispielen

für bestimmte Disziplinen, Genres oder Editionstypen (sog. model editions), geht die Editionen-Registry über bestehende Nachweissysteme (z.B. Franzini 2012–2022, Sahle 2020–2022 und früher) hinaus. Gleichzeitig beschränkt sie sich zunächst vornehmlich auf Editionen, die an Institutionen im deutschen Raum angesiedelt bzw. an denen deutsche Forschungsinstitutionen beteiligt sind, um die Auffindbarkeit und Sichtbarkeit dieser Projekte signifikant zu erhöhen und insbesondere den Zugriff auf deren Forschungsdaten zu befördern.

3. Maßnahmen zur Vernetzung von Editionen

Die Task Area Editionen erkundet Vernetzungspotenziale digitaler Editionen auf Basis von GND-Normdatenannotationen und evaluiert angewandte Praktiken, etwa die Vernetzung durch veröffentlichte BEACON-Daten, entsprechende Schnittstellen und aggregierende (Fach-)Dienste (Lordick und Mache 2018). Diese Praxis stärkt einerseits die FAIRness von (digitalen) Editionen: Sie steigert die Auffindbarkeit durch vernetzte Recherchesysteme, ermöglicht verteilte Datenangebote durch semantische Interoperabilität und verbessert die Datenqualität durch 'Eindeutigkeit'. Sie wirft andererseits die Frage nach projektspezifischen Normdatenbedarfen und damit den Mitwirkungsmöglichkeiten an der GND auf: Die Task Area Editionen arbeitet deshalb mit der entstehenden „Text+ GND-Agentur“ zusammen (Kett et al. 2022).

Die drei genannten Maßnahmen stellen nur einen Ausschnitt der laufenden Arbeiten dar. Für einen möglichst inklusiven Diskurs rund um das Thema „editionsspezifische Forschungsdaten“ sind Austauschformate zur Einbindung der Community geplant – das Poster selbst (mit weiterführenden QR-Codes) gehört dazu. Grundsätzlich muss bei der Umsetzung der skizzierten Maßnahmen berücksichtigt werden, dass jede digitale Edition in einem hohen Maße individuell ist. Dies betrifft zum einen die Materialien der Edition, die Methoden der Erschließung sowie die Publikationswege und zum anderen die Rahmenbedingungen eines Editionsunternehmens (u. a. personelle, zeitliche und finanzielle Ressourcen, vorhandenes Know-How). Diese Individualität wird bei allen geplanten Maßnahmen in Text+ stets berücksichtigt, indem 'Offenheit' als „Skala“ verstanden wird, „auf der Projekte, die offene Methoden anwenden wollen, den für sie jeweils angemessenen Platz finden müssen“ (Hanneschläger 2020, 143).

Fußnoten

1. Contributor Roles: Harald Lordick, Frederike Neuber, Daniela Schulz (Writing – original draft), Philipp Hegel, Kilian Hensen, Jörg Hörnschemeyer, Christoph Kudella, Claes Neufeind, Melanie Elisabeth-H. Seltmann, Martin Sievers (Writing – review & editing)

Bibliographie

Dillen, Wout und Vincent Neyt. 2016. "Scholarly Digital Editing within the Boundaries of Copyright Restrictions." *Digital Scholarship in the Humanities* 31, Nr. 1: 785–796. <https://doi.org/10.1093/llc/fqw011> (zugegriffen: 19. Juli 2022).

Franzini, Greta. 2012–2022. *Catalogue of Digital Editions*. <https://dig-ed-cat.acdh.oeaw.ac.at/> (zugegriffen: 19. Juli 2022).

Gengnagel, Tessa, Frederike Neuber und Daniela Schulz. 2022. *FAIR Principles in Digital Scholarly Editions*. <https://ride.i-d-e.de/fair-criteria-editions/> (zugegriffen: 19. Juli 2022).

Hannessschläger, Vanessa. 2020. „Forschung öffnen: Möglichkeiten, Potentiale und Grenzen von Open Science am Beispiel der offenen Datenbank ‚Handke: in Zungen‘.“ In *Digital Humanities Austria 2018. Empowering Researchers*, hg. v. Marlene Ernst, Peter Hinkelmanns, Lina Maria Zangerl, Katharina Zeppezauer-Wachauer und Verena M. Höller, 140–144. Wien: Austrian Academy of Sciences Press.

Hannessschläger, Vanessa. 2019–2020. "Common Creativity International: CC-licensing and Other Options for TEI-based Digital Editions in an International Context." In *Journal of the Text Encoding Initiative* 11. <https://doi.org/10.4000/jtei.2610> (zugegriffen: 19. Juli 2022).

Institut für Dokumentologie und Editorik. 2014–2022. *RIDE – A Review Journal for Digital Editions and Resources*. <https://ride.i-d-e.de/> (zugegriffen: 19. Juli 2022).

Kett, Jürgen, Christoph Kudella, Andrea Rapp, Regine Stein und Thorsten Trippel. 2022. „Text+ und die GND – Community-Hub und Wissensgraph.“ *Zeitschrift für Bibliothekswesen und Bibliographie* 69, Nr. 1-2: 37–47. <https://doi.org/10.3196/1864295020691262> (zugegriffen: 19. Juli 2022).

Lordick, Harald und Beata Mache. 2018. „Annotationen anhand der Gemeinsamen Normdatei aus einer anwendungsorientierten Perspektive historischer Forschung.“ *Digital Humanities im deutschsprachigen Raum (DHd2018)*, Köln. <https://doi.org/10.5281/zenodo.1188230> (zugegriffen: 27. Juli 2022).

Martinez, Merisa, Wout Dillen, Elli Bleeker, Anna-Maria Sichani und Aodhán Kelly. 2019. "Refining our Conceptions of 'Access' in Digital Scholarly Editing: Reflections on a Qualitative Survey on Inclusive Design and Dissemination." In *Variants* 14: 41–74. <https://doi.org/10.4000/variants.1070> (zugegriffen: 19. Juli 2022).

Pomerantz, Jeffrey und Robin Peek. 2016. "Fifty shades of open." In *First Monday* 4. <http://firstmonday.org/ojs/index.php/fm/article/view/6360/5460> (zugegriffen: 19. Juli 2022).

Rojas Castro, Antonio. 2020. "FAIR enough? Building DH Resources in an Unequal World." *Digital Humanities Kolloquium der BBAW*, 7. August 2020. <https://vimeo.com/445147368> (zugegriffen: 19. Juli 2022).

Sahle, Patrick. 2014. *Kriterienkatalog für die Besprechung digitaler Editionen*. <https://www.i-d-e.de/publikationen/weitereschritten/kriterien-version-1-1/> (zugegriffen: 19. Juli 2022).

Sahle, Patrick. 2020ff. *A Catalogue of Digital Scholarly Editions*, Version 4. <https://www.digitale-edition.de>

Sichani, Anna-Maria. 2017. "Beyond Open Access. (Re)use, Impact and the Ethos of Openness in Digital Editing." In *Advances in Digital Scholarly Editing*, hg. von Boot, Peter, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Andreas Mertgens, Anna-Maria Sichani, Elena Spadini und Dirk Van Hulle, 439–448. Leiden: Siestone Press.

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." *Scientific Data* 3: 160018.10.1038/sdata.2016.18.

Witt, Jeffrey C. 2018. "Digital Scholarly Editions and API Consuming Applications." In: *Digital Scholarly Editions as Interfaces* 12, hrsg. von Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber und Gerlinde Schneider, 219–247. Norderstedt: Books on Demand.

Onboard onto DraCor. Prototyping Workflows to Homogenize Drama Corpora for an Open Infrastructure

Börner, Ingo

ingo.boerner@uni-potsdam.de
Universität Potsdam

Fischer, Frank

fr.fischer@fu-berlin.de
Freie Universität Berlin

Giovannini, Luca

giovannini@uni-potsdam.de
Universität Potsdam

Lu, Christopher

christopher.lu@balliol.ox.ac.uk
University of Oxford

Milling, Carsten

milling@uni-potsdam.de
Universität Potsdam

Skorinkin, Daniil

daniil.skorinkin@uni-potsdam.de
Universität Potsdam

Sluyter-Gäthje, Henny

henny.sluyter-gaethje@uni-potsdam.de
Universität Potsdam

Trilcke, Peer

trilcke@uni-potsdam.de
Universität Potsdam

Approaches to Corpus Homogenization

Comparative endeavors in Computational Literary Studies typically require corpora which are both diverse, i.e., including texts in different languages and from different sources, and homogenized, i.e., formal and structural consistent. One way to tackle this issue is to establish upstream internal guidelines, such as the ones developed within the ELTeC initiative (Schöch et al. 2021).¹ In the following, we report on our approach to homogenizing corpora for DraCor.²

DraCor, based on the concept of Programmable Corpora (Fischer et al. 2019), is an open platform as well as a growing network for hosting, accessing, and analyzing theater plays. DraCor relies on the general TEI model for dramatic texts, with minimal enhancements, and thus facilitates contributions by external scholars who want to onboard their corpora onto its ecosystem. Once integrated, corpora can benefit from the platform's APIs and services, ranging from the computation of network metrics via various extraction functions to SPARQL queries.

Typically, corpora for DraCor are not built from scratch, but are created either by aggregating formally heterogeneous texts from different sources or by transforming existing corpora. Unlike in ELTeC, the homogenization of texts for DraCor usually does not stand at the beginning of the corpus creation process, but is rather an intervention in existing corpora which are sometimes subject to amendment and growth, hence ›living‹. This approach poses a number of challenges, for which we are currently prototyping several workflows. Here, we present the pipelines for mounting to DraCor two new corpora: the English-language *EarlyPrint Drama Corpus* (EPDraCor) and the *Ukrainian Drama Corpus* (UDraCor).³

Corpus Onboarding

From a technical point of view, onboarding corpora onto DraCor is a series of automated and manual transformations of the source data, which depend crucially on the format and markup of the files. Texts from a single, homogeneous collection with pre-existing markup and metadata will require different workflows and pipelines than those coming, for example, from a variety of raw text sources.

This heterogeneous point of departure is what shapes our onboarding approach. Consequently, we are developing a modular workflow made up of a set of demand-dependent components. In addition to guideline-based manual revisions (e.g. pre-structuring texts with Markdown), we use XSLT scripts for automated transformations. Edits specific to theater plays, such as the task of speaker identification, are supported by an Oxygen frame-

work;⁴ we are furthermore experimenting with task-specific GUI applications based on react.js.⁵ The correction and enrichment of metadata, such as the addition of Wikidata ID, is organized semi-automatically via OpenRefine.

A particular challenge is posed by living corpora. Here, the manual transformations performed during onboarding should be reapplicable in case of edits to the source data. Accordingly, we implemented routines for a ›backward compatibility‹ of the markup: the changes made by us during onboarding can later be applied again to a newer version of the source files.⁶

EPDraCor and UDraCor

To develop our workflows and pipelines, two corpora with very different requirements are currently in the process of onboarding. While UDraCor originates from a growing collection of heterogeneous sources, EPDraCor is based on semantically rich TEI files from the Early Print project.⁷ The onboarding of EPDraCor starts with enhancing and correcting the original markup in our copy of the source corpus, accompanied by collecting LOD metadata from additional sources. Then, we combine the enhanced sources with their metadata and use XSLT to transform them, so that the TEI fulfills the requirements of the DraCor platform.

Due to the heterogeneity of the sources, a more case-specific solution must be found for UDraCor in this initial step. Here, the conversion is a semi-automatic procedure with heavy use of string patterns and regex. At the same time, UDraCor takes a community-based corpus-building approach by inviting scholars specializing in Ukrainian studies to work on both technical and content-related tasks. This work on UDraCor once again shows how the technical task of corpus building and community activities are crucially intertwined.

Fußnoten

1. See <https://distantreading.github.io/ELTeC> and <https://distantreading.github.io/Schema/eltec-1.html>.

2. <https://dracor.org>. In the context of CLS INFRA (<https://clsinfra.io>), DraCor has received funding from the European Union's Horizon 2020 program (grant agreement No. 101004984).

3. Both corpora are still a work in progress. For review, the two corpora can (as public alpha) be accessed in the corresponding GitHub repositories <https://github.com/dracor-org/epdracor> and <https://github.com/dracor-org/udracor>. Both corpora will be published as public beta in the context of DHd2023.

4. <https://github.com/dracor-org/dracor-oxygen-framework>.

5. See e.g. our prototype of a Who-Is-Identification-Tool <https://github.com/dracor-org/epdracor-whois> and its interface <https://dracor-org.github.io/epdracor-whois>.

6. For this, see our prototype script in the EPDraCor repository: <https://github.com/dracor-org/epdracor>.

7. <https://earlyprint.org>.

Bibliographie

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama". In *Proceedings of DH2019: "Complexities"*. Utrecht: Utrecht University. <https://doi.org/10.5281/zenodo.4284002>.

Mueller, Martin, and Joseph Loewenstein (eds.). n.d. "Early Print Library". Accessed August 3, 2022. <https://earlyprint.org>.

Schöch, Christof, Roxana Patras, Diana Santos, and Tomaz Erjavec. 2021. "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives". *Modern Languages Open* 1: 25. <https://doi.org/10.3828/mlo.v0i0.364>.

Open Archives VR. Ein 3D-Modell des Theodor-Fontane- Archivs als interaktiver Erlebnis- und Kommunikationsraum

Brandes, Vanessa

vanessa.brandes@fontane-gesellschaft.de
Theodor Fontane Gesellschaft e.V.

Busch, Anna

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

Trilcke, Peer

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

Zimmermann, Ronny

ronny.zimmermann@uni-potsdam.de
Theodor Fontane Gesellschaft e.V.; Theodor-Fontane-Archiv | Universität Potsdam

Offene Archive in virtuellen Räumen

Literaturarchive – wie literarische Gedächtniseinrichtungen im Allgemeinen – stehen heute mehr denn je vor der Herausforderung, ihre gesellschaftlichen Aufga-

ben der partizipativen Vermittlung von und der diskursiven Verständigung über Literatur, Kultur und Geschichte und also ihr Selbstverständnis als gegenwartsorientierte Akteure des kulturellen Gedächtnisses vor dem Hintergrund des medialen Wandels neu zu bestimmen (Wettmann 2018), z.B. im Kontext aktueller Konzeptualisierungen von „Virtual Heritage“ (Champion 2021). Auf die erste Phase der digitalen Öffnung von Literaturarchiven – die Transformation zu offenen Datenräumen – folgt derzeit eine zweite, in der sich die Gedächtniseinrichtungen selbst in den virtuellen Raum übersetzen, um weitere Möglichkeiten der Zugänglichkeit zu eröffnen (Bekele und Champion 2019).

Forciert durch die Pandemie bedeutet dies zunehmend auch, translokale Formen der Präsenz im Archiv zu erproben und damit das Archiv als Ort (Cunningham 2017) neu zu denken. Gepaart mit den Digitalisierungs- und Bereitstellungsbestrebungen von offenen Kulturdaten und -objekten im digitalen Raum, bieten sich so neue Experimentierräume für digitale Interaktionen (vgl. museum4punkt0 2022). Dabei sind die Akteure auf Technologien angewiesen, die Forschungs- und Vermittlungsziele offenlegen und derart kommunizieren, dass über eine museale Präsentation hinaus eine Interaktion mit Kulturobjekten und -praktiken im digitalen Raum möglich wird. Das kann durch virtuelle Lern- oder Spielumgebungen geschehen, die interaktive Erzählräume schaffen. Datenvisualisierung und Datendesign erfolgen beispielsweise durch virtuelle und räumliche Darstellungen (vgl. Glinka u.a. 2020) sowie zunehmend durch simulierte Umgebungen und virtuelle Welten, etwa 3D-Rundgänge (vgl. z.B. Österreichisches Staatsarchiv o.J.).

Das Kooperationsprojekt „FontaneVR“ der Theodor Fontane Gesellschaft e.V. und des Theodor-Fontane-Archivs der Universität Potsdam hat auf diese Überlegungen aufbauend im Jahr 2022 einen Prototyp einer digitalen 3D-Ausstellungs- und Interaktionsumgebung konzipiert und entwickelt. Dafür ist eine VR-basierte Begegnungsstätte gestaltet worden, die der Villa Quandt, dem Sitz des Theodor-Fontane-Archivs, nachempfunden ist, diese aber auch erweitert und hybridisiert. Für diesen virtuellen Ort werden begleitend digitale Vermittlungsformate (Literatur- und Bildungsevents) sowie interaktive Ausstellungsformate entwickelt, die sich an ein diverses Publikum richten und, erreichbar über Webbrowser, individuell oder kollektiv besucht werden können.

„FontaneVR“: Umsetzung und Konzept

Das Projekt „FontaneVR“¹ verfolgt zwei Schwerpunkte: Zum einen den *Bau eines 3D-Modells* der Villa Quandt, samt realistischer Nachbildung der Umgebung, photographischer Erfassung der Architektur und maßstabsgetreuer, realistischer, grafischer 3D-Darstellung der Räumlichkeiten und Ausstellungsobjekte. Zum anderen die Konzeption verschiedener *Vermittlungsformate*, die im Rahmen des 3D-Modells virtuelle Begehung, Exploration und Begegnung erlauben.

3D-Modell

Der detailgetreuen Modellierung der Villa Quandt (siehe Abb.1), ihrer Innenräume und ausgewählter Objekte (siehe Abb. 2) liegt ein objekt- und materialangemessener 3D-Digitalisierungsprozess nach dem Prinzip der Photogrammetrie (siehe Abb. 3.) sowie eine web-basierte 3D-Präsentation über Mozilla Hubs zugrunde. Während ein fotorealistisches Modell des Archiv-Gebäudes die Grundlage bildet, werden einzelne Räume zugleich gezielt »fiktionalisiert«, d.h. mit Möglichkeiten etwa zur musealen Präsentation und Kontextualisierung von Archivobjekten ausgestattet, die in der realen Villa Quandt nicht gegeben sind (Abb. 3). Damit ist das 3D-Modell nicht nur eine Abbildung des Archivortes, sondern eine Erweiterung, die VR bewusst zum Aufbau zusätzlicher archivarischer Funktionen nutzt.²



Abb. 1: prototypische Außenansicht des 3D-Modells der Villa Quandt



Abb. 2: Innenraum der Villa Quandt (3D-Referenzmodell)

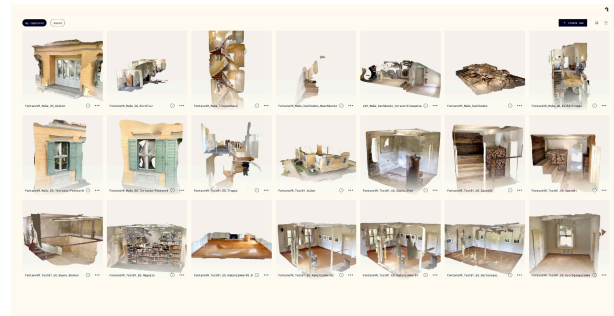


Abb. 3: Verschiedene photogrammetrische Erfassungen von Außen- und Innenräumen der Villa Quandt



Abb. 4: »Fiktionalisierte« Innenräume für museale Inszenierungen mit 3D-Ausstellungsobjekten

Vermittlungsformate

Als konkrete Vermittlungsformate sind eine Reihe von Implementierungen und eventbasierte Nutzungen der virtuellen Begegnungsstätte geplant, die zuvor kuratorisch konzipiert und entwickelt wurden.

Ästhetische Erfahrung

Die literarische Ausstellung: In den virtuellen Räumen werden verschiedene Ausstellungen konzipiert, die es ermöglichen, die Materialität des Fontane-Nachlasses (z.B. Handschriften-Digitalisate) zu erfahren, die Aktualität Fontanes zu entdecken (z.B. durch Tafeln zu neuen Forschungspositionen) oder die Geschichte der Fontane Gesellschaft und des Archivs kennenzulernen. Die Ausstellungen können dabei zeitungebunden durch interessierte Nutzer*innen besucht werden; darüber hinaus werden von professionellen Guides durchgeführte Führungen durch die Ausstellung angeboten.

Information und Unterhaltung

Die literarische Veranstaltung: In Form von Vorträgen, Lesungen, Vorträgen, Werkstattgespräche oder Diskussionen werden diskursive Formen der Vermittlung von Literatur und Geschichte angeboten, zu denen Interessierte sich ortsungebunden in der virtuellen Begegnungsstätte versammeln können.

Gamification

Das *literarische Spiel*: Insbesondere – aber keineswegs nur – für die jüngere Generation (etwa Schülerinnen und Schüler) wird die virtuelle Begegnungsstätte zur Spielstätte transformiert, in der sich Such- und Erkundungsspiele rund um das literarische Werk und den handschriftlichen Nachlass Fontanes durchführen lassen.

Fußnoten

1. „FontaneVR“ wird im Rahmen von „dive in. Programm für digitale Interaktion“ der Kulturstiftung des Bundes entwickelt und durch die Beauftragte der Bundesregierung für Kultur und Medien (BKM) im Programm NEUSTART KULTUR gefördert. Weitere Informationen zum Projekt finden sich hier: <https://www.fontanearchiv.de/fontanevr>.
2. Direkt zur Anwendung gelangt man hier: <https://www.fontanearchiv.de/fontanevr/vr>.

Bibliographie

Bekele, Mafkereseb Kassahun und Erik Champion. 2019 „A Comparison of Immersive Realities and Interaction Methods: Cultural Learning in Virtual Heritage“. *Frontiers in Robotics and AI* 6. <https://doi.org/10.3389/frobt.2019.00091>

Champion, Erik (Hg). 2021. *Virtual Heritage. A Guide*. London: Ubiquity Press. <https://doi.org/10.5334/bck>

Cunningham, Adrian. „Archives as a Place“. In *Currents of Archival Thinking*, hg. von Heather MacNeil, Terry Eastwood, Second edition, 53–79. Santa Barbara, California: Libraries Unlimited, 2017.

Glinka, Katrin, Patrick Tobias Fischer, Claudia Müller-Birn, Silke Krohn. (2020): „Investigating Modes of Activity and Guidance for Mediating Museum Exhibits in Mixed Reality“. In *Kultur und Informatik: Extended Reality*, hg. von Johann Habakuk Israel, Christian Kassung, Jürgen Sieck. Berlin: vwh. <https://doi.org/10.48550/arXiv.2106.13494>

museum4punkt0. 2022. [Website]. <https://www.museum4punkt0.de/> (zugegriffen: 1. August 2022).

Österreichisches Staatsarchiv. o.J. 99 Dokumente [3D-Rundgang]. <https://oe99.staatsarchiv.at/tour-durch-das-archiv/> (zugegriffen: 1. August 2022).

Wettmann, Andrea. 2018. „Die Archive und der ‘Digital Turn’. Eine Standortbestimmung“. In *Kooperative Informationsinfrastrukturen als Chance und Herausforderung*, hg. von Achim Bonte, Juliane Rehnolt, 361–371. Berlin: De Gruyter. <https://doi.org/10.1515/9783110587524-038>

Opening a Journal. Erfahrungen bei der Gründung des Journal of Computational Literary Studies

Gius, Evelyn

evelyn.gius@tu-darmstadt.de
Technische Universität Darmstadt

Schöch, Christof

schoech@uni-trier.de
Universität Trier

Trilcke, Peer

trilcke@uni-potsdam.de
Universität Potsdam

Gerstorfer, Dominik

dominik.gerstorfer@tu-darmstadt.de
Technische Universität Darmstadt

Guhr, Svenja

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

Ripoll, Elodie

ripoll@uni-trier.de
Universität Trier

Sluyter-Gäthje, Henny

sluytergaeth@uni-potsdam.de
Universität Potsdam

1. Einleitung

Die Open-Access-Transformation des wissenschaftlichen Publikationssystems hat nicht nur eine Rekonzeptionalisierung des Zugangs zu Erkenntnissen sowie eine Neuausrichtung der Finanzierung wissenschaftlichen Publizierens eingeleitet; mit ihr gehen auch neue institutionelle Organisationsformen und technische Gestaltungsspielräume einher (u.a. Wissenschaftsrat 2021). Neben die Openness als Zugang treten daher weitere Szenarien der Öffnung. Gerade dort, wo Open-Access-Publikationsmedien ohne die Beteiligung etablierter Verlage gegründet werden, können die Rollen, die Logiken und die Konventionen des wissenschaftlichen

Publizierens derzeit neu verhandelt und womöglich sogar neu erfunden werden.

Im Folgenden berichten wir aus dem Work in Progress der Gründung eines verlagsunabhängigen Open-Access-Journals in den Digital Humanities, des *Journal of Computational Literary Studies* (JCLS).¹ Der erste Call for Papers von JCLS wurde im Herbst 2021 veröffentlicht. Mit seinem ersten Rolling Issue im Herbst 2022 wird JCLS den Publikationsbetrieb aufnehmen. Anlässlich dieser Gründung möchten wir das Journal zum einen vorstellen (Abschnitt 2); zum anderen möchten wir – ausgehend von den Erfahrungen der Gründungsphase des Journals – drei Felder umreißen, auf denen sich aus unserer Sicht aktuell Entwicklungs- und Öffnungspotenziale im Publikationssystem bieten. Dabei handelt es sich um das Feld der Community (Abschnitt 3.1), das Feld des Reviews (Abschnitt 3.2) und das Feld des Workflows (Abschnitt 3.3). Damit wollen wir auch zu einem breiten Austausch über die konkrete Ausgestaltung von unabhängigen Open-Access-Journals anregen.

2. Über JCLS

JCLS ist dauerhaft als internationales Golden-Open-Access-Journal ohne Gebühren für Schreibende oder Lesende angelegt. Es bietet eine Publikationsplattform für Arbeiten zur Entwicklung, Anwendung und Kritik von computergestützten Ansätzen in den Literaturwissenschaften. Die Gründung der Zeitschrift erfolgte zu einem Zeitpunkt, an dem die Computational Literary Studies (CLS) im Rahmen der zunehmenden Ausdifferenzierung der Digital Humanities eine Sichtbarkeit erlangt haben. Die Zeitschrift will Forschung fördern, die das Spektrum computergestützter Methoden zur Analyse literarischer Texte und ihrer (kulturellen, sozialen, historischen, performativen) Kontexte erweitert. Sie bietet ein Forum, um den Aufbau literarischer Korpora, die Identifizierung von Besonderheiten literarischer Texte, die Domänenanpassung von Methoden, die Operationalisierung von Konzepten, die Annotation von Texten, die Evaluation von Messverfahren, die Interpretierbarkeit von Ergebnissen und deren Reproduzierbarkeit zu behandeln. JCLS will schließlich auch die Debattierbarkeit der Kernkonzepte der CLS, Computationalität und Literarizität, adressieren.

Institutionell ist JCLS als verlagsunabhängiges Journal ausgerichtet, das in Kooperation von drei Professuren (in der Rolle der Herausgeber:innen) mit der Universitäts- und Landesbibliothek Darmstadt (ULB) als Infrastrukturpartner betrieben wird. Die ULB sorgt für die nachhaltige Verfügbarkeit der publizierten Artikel (in PDF, XML, HTML und LaTeX) und stellt über eine Kooperation mit der gemeinnützigen Open Library of Humanities² (OLH) das auf Python basierende Redaktionsmanagement- und Publikationssystem Janeway³ zur Verfügung. Dieses wird als Open-Source-Software von der OLH entwickelt (Eve und Byers 2018). Auch die OLH agiert als Infrastrukturpartner, nicht als Verlag.

3. Entwicklungspotenziale

3.1 Community

Die digitale Transformation von Öffentlichkeiten ermöglicht auch ein breiteres Verständnis der Akteursrolle von Publikationsmedien. Während traditionelle Publikationsorgane sich vornehmlich als Publikationsdienstleister verstehen, begreifen sich Publikationsmedien derzeit zunehmend als Community-Hubs, die auch jenseits der Publikation von wissenschaftlichen Inhalten mit ihrer Community interagieren und etwa eventbasiert Räume für die Community schaffen. Openness erweist sich hier als Praxis der Öffnung von diversifizierten Kommunikationsräumen. Mit der Gründung von JCLS wurde entsprechend (neben einem Twitter-Account) zugleich die *Annual Conference of Computational Literary Studies* ins Leben gerufen, die erstmals im Juni 2022 ausgerichtet wurde.⁴

3.2 Review-Verfahren

Es gibt zunehmend – auch in den Digital Humanities (vgl. Burghardt et al. 2022) – Forderungen nach einer konsequenten Umstellung der wissenschaftlichen Qualitätssicherung auf Open Peer Review (vgl. Ross-Hellauer 2017). Dennoch sprechen insbesondere internationale Anerkennungs- und Gratifikationsmechanismen derzeit noch für ein Double Blind Peer Review, das als maßgeblich für die Akzeptanz von Publikationsorganen gilt. JCLS hat sich daher, einem Majoritätstvotum des Editorial Boards folgend, für ein Double Blind Peer Review-Verfahren entschieden. Zugleich integrieren wir im Rahmen der Conference-Paper-Issues eine Phase des offenen, kollaborativen Reviews in den Qualitätssicherungsprozess, die Ideen des Open Peer Review aufgreift.

3.3 Redaktionsworkflow

Die technische Realisierung von Redaktionsworkflows basiert bis heute meist auf einer Logik des Up- und Downloads von Dateien, was u.a. Probleme mit der Versionierung, Hürden für die kollaborative Textbearbeitung und vermeidbaren Mehraufwand mit sich bringt. Mit JCLS haben wir erste Schritte in Richtung eines online-basierten Workflows eingeleitet, bei dem die Texte in einem webbasierten LaTeX-Editor mit Git-Anbindung auf der Grundlage eines gezielt entwickelten und zur Nachnutzung zur Verfügung gestellten LaTeX-Templates⁵ kollaborativ verfasst werden können. Auch den Reviewer:innen wird die Möglichkeit gegeben, anonym Änderungsvorschläge direkt in diesem Online-Dokument vorzunehmen. Projektiert wird derzeit eine Weiterentwicklung dieses Workflows, durch die eine direkte, API-basierte Kommunikation zwischen Janeway und dem LaTeX-Editor möglich werden soll.

Fußnoten

1. Vgl. <https://jcls.io/>.
2. Vgl. <https://www.openlibhums.org>.
3. Vgl. <https://janeway.systems/>.
4. Vgl. <https://jcls.io/site/conference/>.
5. Vgl. <https://github.com/Journal-of-CLS/JCLS-Template/>.

Bibliographie

Burghardt, Manuel et al. 2022. "Offen für alle(s)? Open Identities im Reviewprozess der DHd-Konferenz". In *DHd2022: Kulturen des digitalen Gedächtnisses Konferenzabstracts*. Potsdam, 21-24. DOI: <https://doi.org/10.5281/zenodo.6328145>

Eve, Martin Paul und Andy Byers. 2018. "Jane-way: A Scholarly Communications Platform". In *Insights: The UKSG Journal*. <https://eprints.bbk.ac.uk/id/eprint/22452/> (zugegriffen: 19. Juli 2022).

Ross-Hellauer, Tony. 2017. "What is open peer review? A systematic review." In *F1000Res* (6:588). DOI: <https://doi.org/10.12688/f1000research.11369.2> (zugegriffen: 19. Juli 2022).

Wissenschaftsrat. "Empfehlungen zur Transformation des wissenschaftlichen Publizierens zu Open Access." Köln: Wissenschaftsrat, 2022. <https://www.wissenschaftsrat.de/download/2022/9477-22.pdf>.

Open Jean Paul. Funktionen und Potentiale offener Editionsdaten

Neuber, Frederike

neuber@bbaw.de
Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Lecroq, Axelle

axelle.lecroq@bbaw.de
Berlin-Brandenburgische Akademie der
Wissenschaften, Deutschland

Jean Paul (1763–1825) zählt zu den bedeutendsten Schriftstellern der deutschen Literatur um 1800 und war ein überaus produktiver und geistreicher Briefeschreiber, der mit bekannten Persönlichkeiten wie Heinrich Jacobi, Caroline und Johann Gottfried Herder, Charlotte von Kalb und Rahel Levin Varnhagen korrespondierte. Die Briefe Jean Pauls erschienen bereits Mitte des 20. Jahrhunderts in der Historisch-kritischen Ausgabe (Bernd 1952–1964); Anfang des 21. Jahrhunderts folgten die Briefe an Jean Paul (Begemann et al. 2003–2017), eben-

falls im Druck. Seit 2018 ist Jean Pauls Briefkosmos auf dem Weg in die digitale Welt: Die 5562 Von-Briefe, die zunächst buchzentriert retrodigitalisiert¹ und anschließend briefzentriert retrokonvertiert wurden,² sind seitdem auf *Jean Paul – Sämtliche Briefe digital* verfügbar (Miller et al. 2018–2022). Daneben sind derzeit (Dezember 2022) bereits 1479 Dokumente des sich noch im Aufbau befindenden ‚born digital‘-Korpus der Umfeldbriefe erschienen,³ das die Korrespondenz von Familie, Freundinnen und Kolleginnen des Schriftstellers umfasst.⁴

Aus methodisch-technischer Perspektive, setzt die Edition mit der Verwendung von XML/TEI und dem Basisformat des Deutschen Textarchivs (DTA 2011–2020) sowie der Anreicherung mit Normdaten (GND, GeoNames) auf Standards. Im Zeichen von ‚Open Data‘ erscheinen die XML/TEI-Dokumente der Briefe unter Creative Commons-Lizenz (CC-BY-SA 4.0), und zwar in drei Publikationsmodi, die verschiedene Funktionen hinsichtlich ihrer Nutzung erfüllen:

(1) Zur *Datenlektüre* bzw. zum Abgleich zwischen einem Brieftext in HTML und den zugrundeliegenden Daten kann man jedes Dokument innerhalb der digitalen Edition einzeln als XML/TEI herunterladen. Die editorische Arbeit bzw. das ‚Wissen‘ in den Daten wird so transparent, nachvollziehbar und vollumfänglich zugänglich gemacht, da die Komplexität der Kodierung, wie in den meisten digitalen Editionen, nicht vollständig im User Interface abgebildet wird (Neuber 2023).

(2) Zur *maschinellen Interaktion* über technische Schnittstellen (Witt 2018) bietet die digitale Edition derzeit verschiedene BEACON-APIs und eine CMIF-API, wodurch u. a. die *Deutsche Biographie* und *correspSearch* (Dumont et al. 2021) Informationen der Edition aggregieren.⁵ Gleichzeitig bezieht die digitale Edition über die erfassten Normdaten auch selbst Informationen von Schnittstellen, beispielsweise Koordinaten von *GeoNames* zur Generierung von Karten und Bild-URLs mit Portraits von *Wikimedia Commons*.

(3) Zur *Nachnutzung* bzw. *Re-kontextualisierung* der Datensätze stellt die Edition die Brieftexte als Paket auf GitHub und Zenodo bereit (Neuber 2022a), wodurch die Daten archiviert und versioniert sowie mit einer DOI zitierbar sind. Durch diese Art der Datenpublikation wurde das Korpus der Briefe von Jean Paul bereits mehrfach jenseits der Edition in anderen Kontexten nachgenutzt: im *Digitalen Wörterbuch der Deutschen Sprache* als historisches Korpus (2022), im *CorpusExplorer* als korpuslinguistische Ressource⁶ (Oliver 2018) und auf Twitter als Bot @jeanpaultoday⁷ (Neuber 2022b). So wird durch die offene Bereitstellung der Daten Forschung jenseits der Edition gefördert und die Brieftexte einem gänzlich neuen Publikum zur Verfügung gestellt.

Der Beitrag, der die Publikationsmodi der Jean Paul Briefedition und ihre jeweilige Funktion illustriert, ist für die DHd-Konferenz höchst relevant, da ‚Open Data‘ im Editions-kontext immer noch eher die Ausnahme als die Regel ist. Aus Greta Franzinis Editionenkatalog (2016–2022) geht hervor, dass von 320 digitalen Editionen lediglich ~27% CC-Lizenzen verwenden, ~23% TEI-Daten zum Download bereitstellen und ~5% APIs anbieten. Die Zahlen sind bedauerlich, da Daten das primäre Forschungsergebnis digitaler Editionen sind:

[...] [I]n digital editions the encoded texts themselves are the most important long-term outcome of the project, while their initial presentation within a particular application should be considered only a single perspective on the data. Any given view will be far from unique or canonical, as different usage scenarios call for different presentations (Turska et al. 2017, #4).

Im Kontext der Jean Paul-Edition gelten die Daten den Herausgeberinnen als *Primärpublikation*, die in der digitalen Edition im Web ihre Kernpräsentation, nicht aber ihre einzige (Re-)Präsentationsform finden müssen. Die Maßnahmen für offene Daten zielen daher auf ein Höchstmaß an Transparenz, Interoperabilität und Nutzbarkeit, um einer Nachnutzung durch Mensch und Maschine gerecht zu werden (Baillot und Busch 2021).⁸

Fußnoten

1. Siehe für Band 1 der Briefe von Jean Paul im Deutschen Textarchiv https://www.deutschestextarchiv.de/jeanpaul_briefe01.1956 (zugegriffen: 26. Juli 2022).
2. Zum DFG-Projekt der Digitalisierung der Briefe von Jean Paul siehe <https://www.jeanpaul-edition.de/von-briefe-digital.html> (zugegriffen: 26. Juli 2022).
3. Zum DFG-Projekt der Erschließung der Umfeldbriefe siehe <https://www.jeanpaul-edition.de/umfeldbriefe.html> (zugegriffen: 26. Juli 2022).
4. Perspektivisch soll die digitale Edition um das noch zu retrokonvertierende Korpus der Briefe an Jean Paul ergänzt und damit vervollständigt werden.
5. Für Informationen zu den Schnittstellen siehe <https://www.jeanpaul-edition.de/daten.html> (zugegriffen: 26. Juli 2022).
6. Siehe <https://notes.jan-oliver-ruediger.de/korpora/> (zugegriffen: 26. Juli 2022).
7. Siehe <https://twitter.com/jeanpaultoday> (zugegriffen: 26. Juli 2022).
8. Neben den genannten Aspekten wird das Poster die nächsten Schritte hinsichtlich der weiteren Pflege und Verbesserung der Daten und ihrer Publikation aufzeigen, die u.a. die XML-Schemata und die API betreffen.

Bibliographie

- Baillot, Anne und Anna Busch. 2021. "Editing for Man and Machine. Digital Scholarly Editions and their Users" In *Variants* 15-16. <https://doi.org/10.4000/variants.1220> (zugegriffen: 26. Juli 2022).
- Begemann, Christian, Markus Bernauer und Norbert Miller, Hg. 2003-2017. *Jean Pauls Sämtliche Werke. Vierte Abteilung: Briefe an Jean Paul*. Berlin-Brandenburgischen Akademie der Wissenschaften.
- Berend, Eduard, Hg. 1952-1964. *Jean Pauls Sämtliche Werke. Vierte Abteilung: Briefe an Jean Paul*. Berlin-Brandenburgischen Akademie der Wissenschaften.
- Bernauer, Markus, Norbert Miller und Frederike Neuber, Hg. 2018-2022. *Jean Paul – Sämtliche Briefe digital*. Berlin-Brandenburgischen Akademie der Wissenschaften. <https://www.jeanpaul-edition.de> (zugegriffen: 26. Juli 2022).

Deutsches Textarchiv, Hg. 2011-2020. *DTABf. Deutsches Textarchiv – Basisformat*. Berlin-Brandenburgischen Akademie der Wissenschaften. <http://deutsches-textarchiv.de/doku/basisformat> (zugegriffen: 26. Juli 2022).

Digitales Wörterbuch der deutschen Sprache, Hg. 2022. *Jean Paul Briefe (Textkorpus)*. https://www.dwds.de/d/korpora/jean_paul (zugegriffen: 26. Juli 2022).

Dumont, Stefan, Sascha Grabsch und Jonas Müller-Laackman. 2021. *correspSearch – Briefeditionen vernetzen (2.0.0)*. Berlin-Brandenburgische Akademie der Wissenschaften. <https://correspSearch.net> (zugegriffen: 26. Juli 2022).

Franzini, Greta. 2012-2022. *Catalogue of Digital Editions*. <https://dig-ed-cat.acdh.oeaw.ac.at/> (zugegriffen: 26. Juli 2022).

Neuber, Frederike, Hg. 2022a. *telota/jean_paul_briefe: Daten der Briefe von Jean Paul und der Briefe aus seinem Umfeld (v.6.0)* [Zenodo-Repositorium]. <https://doi.org/10.5281/zenodo.6892400> (zugegriffen: 26. Juli 2022).

Neuber, Frederike. 2022b. *jeanpaultoday* [GitHub-Repositorium]. <https://github.com/FrederikeNeuber/jeanpaultoday> (zugegriffen: 26. Juli 2022).

Neuber, Frederike. 2023 [im Erscheinen]. "Der digitale Editionstext. Technologische Schichten, 'editorischer Kerntext' und Rezeption 2.0." In *Der Text und seine (Re-)Produktion (editio Beihefte)*, hg. von Niklas Fröhlich, Bastian Politycki, Dirk Schäfer, Annkathrin Sonder. Berlin/Boston: De Gruyter.

Rüdiger, Jan Oliver. 2018. *CorpusExplorer*. Universität Kassel / Universität Siegen. <http://www.CorpusExplorer.de> (zugegriffen: 26. Juli 2022).

Turska, Magdalena, James Cummings und Sebastian Rahtz. 2017. "Challenging the Myth of Presentation in Digital Editions." In *Journal of the Text Encoding Initiative* 9. <https://doi.org/10.4000/jtei.1453> (zugegriffen: 26. Juli 2022).

Witt, Jeffrey C. 2018. "Digital Scholarly Editions and API Consuming Applications." In: *Digital Scholarly Editions as Interfaces 12*, Hg. von Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber und Gerlinde Schneider, 219-247. Norderstedt: Books on Demand.

OWIDplusLIVE – Tagesaktuelle N-Gramm-Analysen

Rüdiger, Jan Oliver

ruediger@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Deutschland

Wolfer, Sascha

wolfer@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Deutschland

Cotgrove, Louis

cotgrove@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Deutschland

Schon bald nachdem die ersten Coronavirus-Infektionsfälle auch in Deutschland bestätigt wurden, deutete sich an, dass die gesellschaftlichen Auswirkungen der Pandemie immens sein würden. Es war daher teilweise vorauszusehen, dass die Pandemie auch ihren Niederschlag in der Sprache finden würde. Und doch ist erstaunlich, wie weitreichend und tiefgreifend das Pandemiegeschehen und die gesellschaftlich-politischen Reaktionen Einfluss auf unseren Sprachgebrauch übten und üben, insbesondere auf der Ebene des Wortschatzes. Wir stellen zwei Ressourcen (OWIDplusLIVE und das zugrundeliegende Live-RSS-Korpus) vor, die einen explorativen Zugang zur Erforschung dieses Einflusses bieten. Zudem soll der sprachwissenschaftlichen Forschungsgemeinschaft ein Instrument an die Hand gegeben werden, auch andere sprachliche Entwicklungen in der Zukunft möglichst unmittelbar zu entdecken und anhand von Frequenzverläufen nachzuzeichnen. Das folgende Beispiel (Abb. 1) zeigt vier nacheinander gestellte Suchabfragen zu den Bi-Grammen: *zweite* (in blau), *dritte* (grün), *vierte* (gelb) und *fünfte Welle* (rot) [Stand: 26. September 2022].

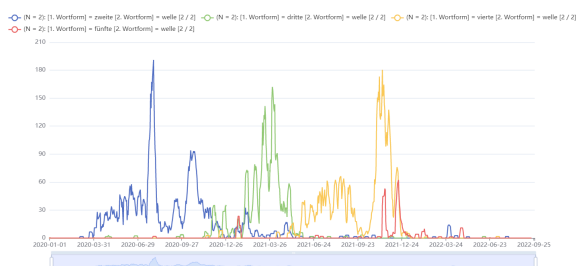


Abb. 1

Das zugrundeliegende Korpus besteht aus Titeln und kurzen Einführungstexten (sog. RSS-Feeds) zu Artikeln aus (derzeit) 13 deutschsprachigen Online-Quellen (Details zu den Quellen und zur Quellenauswahl siehe Vorprojekt: Wolfer u. a. 2020). Das Korpus wird seit dem 01.01.2020 täglich erhoben und umfasste am 26. September 2022 ca. 84,1 Millionen Token. Die Daten sind auch in Form von täglichen Unigramm- (inkl. Wortarten-Tagging) und Bigramm-Frequenzlisten frei auf OWIDplus (www.owid.de/plus/covidplus2020) verfügbar.

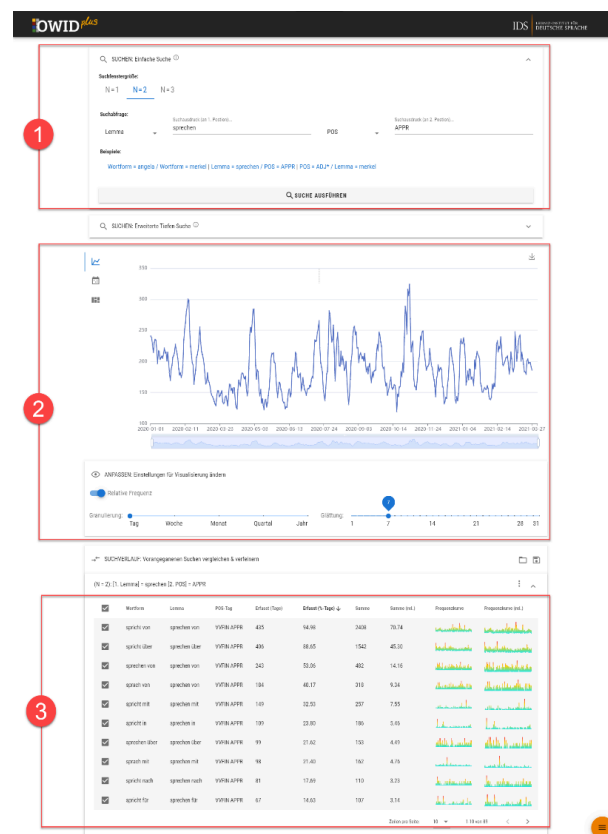


Abb. 2

OWIDplusLIVE wurde mit dem Ziel entwickelt, eine flexible und performante Lösung unabhängig vom Gegenstand (COVID-19) zu bieten. Damit löst dieses Tool den zuvor erstellten Prototypen 'cOWIDplus Viewer' (Wolfer u. a. 2020) ab. OWIDplusLIVE professionalisiert den Prototypen in folgenden fünf Bereichen: (1) zusätzliche Annotations-Layer, konkret: Lemma und Wortart (Part-of-Speech, POS), (2) größere N-Gramme (aktuell Tri-Gramme, aber auch die Möglichkeit N-Gramme größer 3 zu erfassen) sowie (3) die Möglichkeit zusätzlicher Visualisierungen. Dafür (4) wurden sowohl die webbasierte Oberfläche als auch das dahinterliegende Daten-Backend von Grund auf neu entwickelt. Die bestehende Feed-Verarbeitungspipeline konnte ohne größere Änderungen übernommen werden. Zentral für den Ansatz hinter OWIDplusLIVE ist die (5) gezielte Verzahnung von Technologien (Falk u. a. 2020; Banon u. a. 2022; You u. a. 2022), die es ermöglichen, die Anwendung einfach mit neuen Daten (und ggf. Analysemöglichkeiten) zu erweitern, die Berechnungen über mehrere Server zu verteilen, sowie Anfragen so effizient wie möglich zu verarbeiten. Alle im Projekt entwickelten Komponenten (API und Web-Frontend) stehen kostenfrei als OpenSource (unter der AGPL-3.0 Lizenz) zur Verfügung - siehe: <https://github.com/notesjor/IDS.OWID.Plus.Live>

Die Abfrage durch die Nutzer*innen erfolgt über eine webbasierte Oberfläche. Ein Großteil der Berechnungen und Visualisierungen findet im Browser der Nutzer*innen statt. OWIDplusLIVE ist verfügbar unter <https://www.owid.de/plus/live-2021>. Die Oberfläche ist in drei Segmente eingeteilt, die im Folgenden benannt und weiter unten er-

klärt werden (siehe Abb. 2): (1) Der Abfragebereich. (2) Ein Bereich mit drei unterschiedlichen Visualisierungen. (3) Sowie die Detailansicht.

Q. SUCHEN: Einfache Suche

Suchfenstergröße: N=1 N=2 N=3

Suchabfrage: Suchabdruck (in 1. Position): ... Suchabdruck (in 2. Position): ...

Lemma: sprechen POS: APPR

Beispiele: Wortform = angela / Wortform = merkel / Lemma = sprechen / POS = APPR / POS = ADJ* / Lemma = merkel

SUCHE AUSFÜHREN

Abb. 3

Abb. 3 zeigt den Abfragebereich mit einer einfachen Suche nach Bi-Grammen auf unterschiedlichen Layern. Auf eine komplexe Such-Syntax wurde bewusst verzichtet. Platzhalter wie ‚?‘ und ‚*‘ sind jedoch möglich. Zuerst (1) wurde die Suchfenstergröße N=2 (Bi-Gramm) gewählt. An der ersten Position des Bi-Gramms wird auf dem Layer ‚Lemma‘ (2) nach ‚sprechen‘ (3) gesucht. An der zweiten Position wird auf dem POS-Layer (4) nach APPR (5) gesucht (APPR steht für die Wortart ‚Präposition; Zirkumposition links‘). Diese Abfrage ergibt somit Bi-Gramme wie „sprechen mit“, „spricht über“, „sprachen aufgrund“ usw.

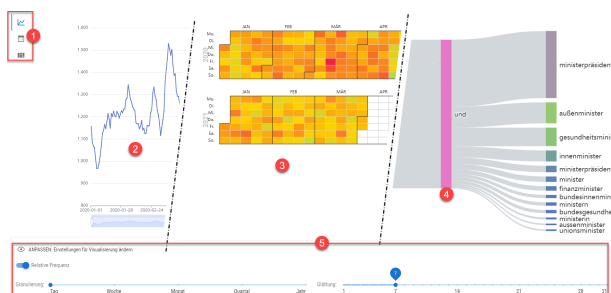


Abb. 4

Abb. 4 zeigt, kompakt zusammengeschnitten, die aktuell verfügbaren Visualisierungen. Diese können links (siehe Abb. 4 – Markierung 1) gewählt werden. Zur Verfügung steht ein tagesbasierter Frequenzverlauf (2 – siehe auch Abb. 1), eine Kalenderansicht (3) und ein Sankey-Diagramm (4). Die Visualisierungen können über den unteren Bereich (5) angepasst werden. Es ist z. B. möglich, absolute und relative Frequenzen auszuwerten, eine Granulierung (Auswertung pro Tag, Woche, Monat, Quartal und Jahr) und davon abhängig eine Glättung zu wählen.

– SUCHVERLAUF: Vorangegangenen Suchen vergleichen & verfeinern

(N = 2): (1. Lemma) = sprechen (2. POS) = APPR

Wortform	Lemma	POS-Tag	Erfasst (Tag)	Erfasst (%-Tag)	Sememe	Sememe (rel.)	Frequenzkurve	Frequenzkurve (rel.)
spricht von	sprechen von	VVFVN APPR	425	94.98	2408	70.74		
spricht über	sprechen über	VVFVN APPR	406	88.65	1542	45.30		
sprechen von	sprechen von	VVFVN APPR	243	53.06	482	14.16		
sprach von	sprechen von	VVFVN APPR	184	40.17	318	9.34		

Abb. 5

Der Auszug der Detail-Ergebnisse im Suchverlauf (siehe Abb. 5) ermöglicht es, eine Teilmenge von Ergebnissen auszuwählen (1). Die gesamten Daten einer einzelnen Suchabfrage können über das Dreipunkt-Menü (siehe Bereich 2) als JSON, TSV und URL exportiert werden, um die Daten weiterzugeben bzw. auch um die Daten mit anderen Programmen auszuwerten und zu visualisieren. Außerdem ist es möglich, den gesamten Suchverlauf (siehe Bereich 3), also alle Suchabfragen, als JSON zu exportieren und einen gespeicherten Suchverlauf wiederherzustellen.

OWIDplusLIVE stellt bereits jetzt eine Ressource für die tagesaktuelle Analyse sprachlicher Daten in RSS-Newsfeeds deutscher Online-Presse dar. Trotzdem gibt es an einigen Stellen Potential zur Weiterentwicklung. So könnten die analysierten Zeitabschnitte noch flexibler gestaltet werden, um auch Entwicklungen zu erfassen, die kleinteiliger als ein Tag (z. B. für die Analyse von Social-Media-Sprachdaten) oder grobkörniger als ein Jahr (z. B. für diachrone Analysen) sind. Außerdem sind zusätzliche Visualisierungen denkbar, die unterschiedliche Blickwinkel auf die Daten ermöglichen würden.

Bibliographie

Banon, Shay und ‚Elastic NV contributors‘. 2022. Elasticsearch. <https://www.elastic.co/de/elasticsearch/> (zugegriffen: 28. Juli 2022).

Falk, Warren und ‚RocksDB contributors‘. 2020. RocksDB. <https://github.com/elastic/elasticsearch-net> (zugegriffen: 28. Juli 2022).

Wolfer, Sascha; Koplenig, Alexander; Michaelis, Frank und Müller-Spitzer, Carolin. 2020. Tracking and analyzing recent developments in German-language online press in the face of the coronavirus crisis cOWIDplus Analysis and cOWIDplus Viewer. In *International Journal of Corpus Linguistics*. <https://doi.org/10.1075/ijcl.20078.wol> (zugegriffen: 10. Oktober 2022).

You, Evan und ‚Vue.js contributors‘. 2022. Vue.js. JavaScript. <https://vuejs.org/> (zugegriffen: 28. Juli 2022).

Projektvorstellung – Sprachanfragen. Empirisch gestützte Erforschung von Zweifelsfällen

Lang, Christian

lang@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Deutschland

Tu, Ngoc Duyen Tanja

tu@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Deutschland

Schneider, Roman

schneider@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Volodina, Anna

volodina@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Einführung

Das im Januar 2022 gestartete Projekt „Sprachanfragen“ (<https://www.ids-mannheim.de/gr/projekte2/sprachanfragen/>) verfolgt das Ziel, Sprachanfragedaten – also Daten, die im Rahmen von verschiedenen Sprachberatungsszenarien entstehen, wie beispielsweise (1) – zu erfassen, aufzubereiten und ein wissenschaftsöffentliches Monitorkorpus aus ihnen zu erstellen. Dazukommend wird eine Rechterschnittstelle entwickelt, mit der die Sprachanfragen systematisch wissenschaftlich analysierbar gemacht werden.

(1) „[Frage:] Heißt es „Dramaform“ oder „Dramenform“ [...]?“

[Antwort:] In allgemeinsprachlichen Wörterbüchern ist diese Zusammensetzung nicht erfasst. Im allgemeinen Schreibgebrauch wird – wie eine Internetrecherche ergab – die Form mit Fugen-en vorgezogen.“

Sprachanfragen bieten einen authentischen Einblick in Probleme und Themen, die Sprecher:innen außerhalb der linguistisch-fachwissenschaftlichen Gemeinschaft beschäftigen. Wie Breindl (2016, 86f.) ausführt, bietet eine systematische Auswertung der Sprachberatungspraxis eine wertvolle Grundlage für die Erforschung einer großen Bandbreite verschiedener Fragestellungen. So können diese Daten u. a. dazu benutzt werden, um (i) Zweifelsfälle zu analysieren, wodurch Normierungslücken aufgedeckt werden können, und um (ii) Sprachwandelphänomene nachzuvollziehen. Ebenfalls können Sprachanfragen herangezogen werden, um (iii) Strategien zu erforschen, wie fachspezifische Inhalte von Nicht-Fachpersonen erfragt werden. Dadurch können bspw. die Zugangswege zu grammatischen und orthographischen Inhalten in einem webbasierten Informationssystem optimiert werden. Eine mögliche Optimierung wäre, Sprachanfragen automatisch in Form eines Chatbots zu beantworten.

Das Poster gibt einen Überblick über das Projekt, zeigt erste Ergebnisse und bietet einen Ausblick auf Überlegungen zur Konzeption eines Chatbots zur automatisierten Beantwortung von Sprachanfragen.

Datengrundlage

Das Monitorkorpus wird zum einen aus ~50.000 Sprachanfragen, die an den Sprachberatungsservice des WAHRIG-Verlags per E-Mail geschickt wurden, aufgebaut. Diese decken einen Zeitraum von 1999 bis 2018 ab. Die zugehörigen Antworten werden ebenfalls in das Korpus aufgenommen. Zum anderen wird das Korpus

kontinuierlich mit Sprachanfragen erweitert, die im Leibniz-Institut für Deutsche Sprache eingehen. Um mehr Daten für das Trainieren eines Chatbots zu generieren, werden darüber hinaus Sprachanfragen aus Online-Quellen, wie z.B. gutefrage.net, extrahiert.

In einem ersten Schritt werden die Daten aufwendig vorverarbeitet. Dabei werden sie anonymisiert, um den Datenschutz zu gewährleisten und das Korpus wissenschaftsöffentlich zur Verfügung stellen zu können. Für die Anonymisierung ist die Nutzung eines Named-Entity-Erkennters, wie in anderen Arbeiten geschehen (vgl. u.a. Bleicken et al., 2016; Kleinberg et al., 2017), nicht optimal, da u. a. Namen ebenfalls Teil der Fragestellung sein können (vgl. (2)). Somit müssen automatisierte Lösungswege gefunden werden, um primär tatsächlich personenbezogene Daten zu ersetzen und die anschließende manuelle Nachkorrektur maßgeblich zu erleichtern.

(2) „[...] Der Genitiv des Wortes „Paulus“ [...] sollte wie lauten: „Pauli“ oder „Paulus“? [...]“

Darüber hinaus werden die Sprachanfragen nach orthographischen und terminologischen Kriterien strukturiert, indem sie mit grammatischen Termini (z. B. „Dativ“, „Fugen-s“, „Getrenntschreibung“) annotiert werden. Basis dafür ist die terminologische Ressource der Abteilung Grammatik des Leibniz-Instituts für Deutsche Sprache, die sogenannte Wissenschaftliche Terminologie (WT, <https://grammis.ids-mannheim.de/terminologie/>). Diese beinhaltet ~6.000 Termini aus der Domäne Deutsche Grammatik (vgl. u.a. Suchowolec et al., 2019). Berücksichtigt werden Uni- (z.B. „Substantivierung“), Bi- (z.B. „indirekte Rede“) und Trigramme (z.B. „negationsinduzierend additive Konnektoren“). Mit Hilfe eines Pattern Matchings werden vorkommende Termini in den lemmatisierten Sprachanfragen automatisiert detektiert. Über exakte Treffer hinaus werden bei der Annotation von Uni-grammen auch Teiltreffer am Anfang oder am Ende eines Lemmas aufgenommen, bspw. „Genitivbezug“, „Dativform“, „Muss-Komma“. Somit werden auch Ausdrücke erfasst, die einen Terminus als Erst- oder Zweitglied beinhalten, als Ganzes jedoch nicht als Termini in der WT auftreten.

Um zu evaluieren, wie gut die Automatisierung der beiden Vorverarbeitungsschritte funktioniert, wird ein Subkorpus aus 1.000 zufällig extrahierten Sprachanfragen erstellt. Dieses wird manuell anonymisiert sowie terminologisch annotiert und als Goldstandard bei der Auswertung der automatischen Methoden herangezogen.

Ausblick: automatisierte Beantwortung von Sprachanfragen

Eine weiterführende, zukünftige Zielsetzung ist zudem, bei ausreichender Größe des Monitorkorpus, einen Chatbot zur automatischen Beantwortung von Sprachanfragen zu entwickeln. Dafür werden die Sprachanfragen nach den zugeordneten Termini gruppiert und ein Modell je Gruppe trainiert. Als Baseline wird ebenfalls ein regelbasierter Chatbot implementiert. Denkbar wäre auch eine Kombination aus regelbasiertem und trainiertem Chatbot. Das Ziel ist es, mit einem solchen Sys-

tem eine nicht-kommerzielle und offene (im Sinne von Veröffentlichung des Quellcodes) Alternative zu anderen Online-Grammatik- und Rechtschreibhilfertools (z. B. Deepkomma, Duden-Mentor, LanguageTool oder Studi-Kompass) zu schaffen, die durch nahtlose Anknüpfung an die umfassenden sprachwissenschaftlichen Ressourcen des hauseigenen wissenschaftlichen Informationssystem zur deutschen Grammatik grammis (<https://grammis.ids-mannheim.de/>) umfangreiche Materialien zum weiterführenden Selbststudium auf verschiedenen Komplexitätsstufen bietet.

Im Fokus der automatischen Beantwortung soll also nicht nur die Korrektur, sondern es sollen auch die sprachwissenschaftlichen Hintergründe einer Frage stehen. Zum Beispiel sollen im Fall der folgenden authentischen Sprachanfrage: „Was ist korrekt: „Haushalthilfe“ oder „Haushaltshilfe“, „Haushaltspflege“ oder „Haushaltspflege“?“ über die Angabe der korrekten Variante hinaus das zugrundeliegende Phänomen (Fugenelemente) benannt und entsprechende Artikel aus grammis verlinkt werden.

Bibliographie

Bibliographisches Institut GmbH. 2022. „Duden Mentor.“ <https://mentor.duden.de/>.

Bleicken, Julian, Thomas Hanke, Ute Salden, und Sven Wagner. 2016. „Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data.“ In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 16)*, 3303-3306. Portorož, Slovenia.

Breindl, Eva. 2016. „Sprachberatung im interaktiven Web“. In *Die Kodifizierung der Sprache. Strukturen, Funktionen, Konsequenzen*, herausgegeben von Wolf-Peter Klein und Sven Staffeldt, 85-109. WespA – Würzburger elektronische sprachwissenschaftliche Arbeiten 17. Würzburg.

gutefrage.net GmbH. o.J. „guteFrage.“ <https://www.gutefrage.net/>

Kleinberg, Bennett, Maximilian Mozes, Yaloe van der Toolen, und Bruno Verschuere. 2017. „NETANOS - Named entity-based Text Anonymization for Open Science.“ Preprint. Open Science Framework. <https://doi.org/10.31219/osf.io/w9nhb>.

LanguageTool GmbH. o. J. „LanguageTool.“ <https://languetool.org/de>.

Mannheim: Leibniz-Institut für Deutsche Sprache. o. J. „Grammatisches Informationssystem ‚grammis‘.“ <http://grammis.ids-mannheim.de>.

Suchowolec, Karolina, Christian Lang, und Roman Schneider. 2019. „An empirically validated, onomasiologically structured, and linguistically motivated online terminology. Re-designing scientific resources on German grammar.“ *International Journal on Digital Libraries* 20: 253-268.

Uniprof LLP. 2016-2022. „Studi-Kompass.“ <https://studi-kompass.com/generatoren/online-rechtschreibpruefung>.

Wefelscheid, Cornelius. o. J. „DeepKomma.“ <https://deepkomma.de>.

Schlendern im Digitalen Museum

Hall, Mark

mark.hall@work.room3b.eu

The Open University, Vereinigtes Königreich

Walsh, David

david.walsh@edgehill.ac.uk

Edge Hill University, Vereinigtes Königreich

Die Digitalisierung unseres Kulturguts hat zu riesigen Sammlungen geführt, welche über Suchsysteme öffentlich verfügbar sind. Im Sinne der „Open Culture“ ist Verfügbarkeit aber nicht genug (Walker 2022), da die weiße Suchmaske für nicht-Expert:inn:en eine signifikante Hürde für den Zugriff darstellt (Belkin, Oddy, and Brooks 1982; Whitelaw 2015). Nicht-Expert:inn:en fehlt oft das notwendige Wissen und das spezifische Suchziel (Mayr et al. 2016), um erfolgreich Suchschlüsselwörter für die Suche zu entwickeln. Für digitale Sammlungen ist daher eine Bounce-rate von über 50% normal (Hall et al. 2012; Walsh et al. 2020).

Rich Prospect Browsing (Ruecker, Radzikowska, and Sinclair 2011) und Generous Interfaces (Whitelaw 2015) bieten den weniger erfahrenen Nutzer:inne:n Interfaces an, die einen sanfteren Einstieg in digitale Sammlungen ermöglichen. Beide Ansätze versuchen einen Überblick über die Sammlung zu geben, bevor sie zu einzelnen Objekten hineinzoomen (Shneiderman et al. 1998).

Bestehende Ansätze wie das Coins interface (Gortana et al. 2018), Curator Table interface (Google Arts & Culture 2022), oder das Museum of the World (The British Museum and Google Cultural Institute 2022) nutzen Visualisierungen um einen Überblick über die gesamte Sammlung zu geben, die dann erforscht werden kann. Die Visualisierung bieten jedoch wenig bis keine Hinweise darauf gibt, welche Themen die Sammlung abdeckt und wo in der Visualisierung diese zu finden sind. Wegen der Größe der Sammlungen sind die einzelnen Objekte in der Visualisierung am Anfang auch oft wenig größer als ein paar farbige Pixel.

In einem physischen Museum gibt die Gebäude- und Raumstruktur einen Rahmen um das Museum zu erforschen. Eine derartige Struktur ist auch im digitalen Museum notwendig, fehlt aber in den meisten digitalen Sammlungen und den Kulturorganisationen fehlt die Kapazität um eine zu entwickeln. In Hall and Walsh (2021) haben wir mit der Digital Museum Map einen automatischen Kuratierungsalgorithmus entwickelt, der, basierend auf den Objektmetadaten und einer Reihe von Heuristiken, eine navigierbare Struktur entwirft, die aus hierarchisch strukturierten Gruppen von zwischen 25 und 100 Objekten besteht. Die Qualität der Struktur ist niedriger als bei einer händischen Kuratierung, erlaubt es aber große Sammlungen schnell zu kuratieren. Der Algorithmus arrangiert die navigierbare Struktur dann in Räume, welche in Stockwerken organisiert sind. Dazu wird eine greedy Heuristik genutzt, um die automatisch kuratierte Struktur, basierend auf thematischen Ähnlich-

keiten, kohärent im Stockwerksgrundriss zu positionieren. Der Grundriss wird vom Museum bereitgestellt und erlaubt es das Layout den Museumsinteressen anzupassen.

Das Digitale Museum

Das Digitale Museum nutzt auf der Landeseite (Abb. 1) keine Visualisierung um die gesamte Sammlung anzuzeigen, sondern bietet vier verschiedene Einstiegspunkte in die Sammlung an: ein Objekt des Tages, eine Auswahl der größten Teilsammlungen, eine zufällige Auswahl an Objekten, und ein Link in die gesamte Sammlung. Folgt der/die Nutzer:in dem Link zu einer Teilsammlung oder in die gesamte Sammlung, dann wird das Navigationsinterface in Abbildung 2 angezeigt. Links können die einzelnen Stockwerke ausgewählt werden und rechts werden die Räume auf dem Museumsgrundriss angezeigt. Nach der Auswahl eines Raumes, werden die Objekte des Raumes angezeigt (Abb. 3). Auf diese Weise kann das gesamte Museum durch Browsing erkundet werden.

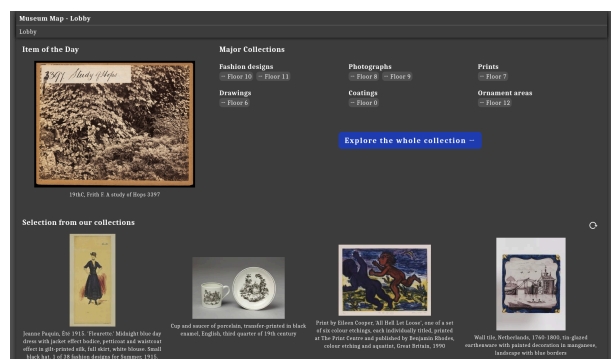


Abbildung 1: Die Landeseite bietet Einstiegspunkte in das Digitale Museum: auf der Objektebene, auf der Teilsammlungsebene, und auf der Ebene der gesamten Sammlung. Alle Bilder © Victoria & Albert Museum, 2022.

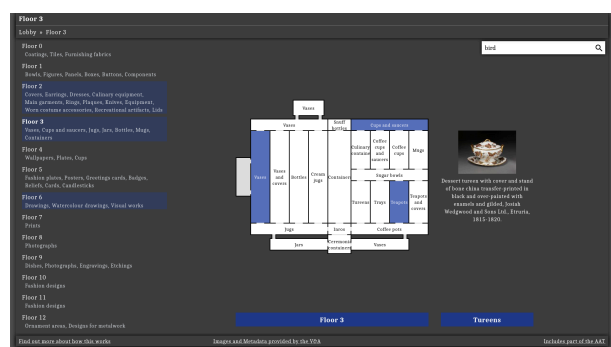


Abbildung 2: Das Stockwerksinterface erlaubt freies Entdecken im Digitalen Museum und kann mit anderen Informationen überlagert werden, wie hier mit Suchergebnissen.

Ein Vorteil des digitalen Mediums ist, dass zusätzliche Informationen über die Stockwerksvisualisierung gelegt werden können. Browsing wird zwar bevorzugt, aber eine Suchfunktion ist auch vorhanden. Sucht der/die Nutzer:in nach etwas, werden die Stockwerke und Räume,

die passende Objekte beinhalten, visuell hervorgehoben. Ebenso werden in der Raumsansicht die passenden Objekte hervorgehoben. Dadurch kann gezielt gesucht werden, ohne den Kontext, in dem die Objekte stehen, zu verlieren.

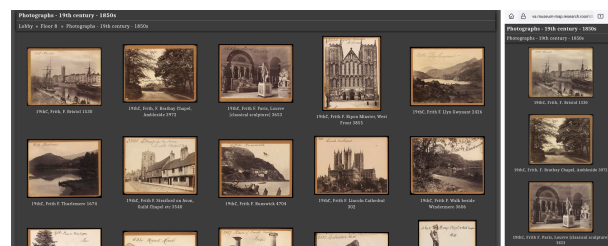


Abbildung 3: Ein Raum im Museum, links in der Desktopansicht, rechts die mobile Version.

Software und Ausblick

Die Digital Museum Map bietet einen einfachen Einstieg, um Sammlungen der interessierten Öffentlichkeit in einem Digitalen Museum zugänglich zu machen. Daher ist sie konfigurierbar und als Open-Source Software verfügbar (<https://github.com/scmmmh/museum-map/>). Zur Zeit bestehen Demoverversionen der Digital Museum Map für einen Teil der *Victoria & Albert* Sammlung (<https://va.museum-map.research.room3b.eu/>) und für die *People Past and Present* Sammlung der Stadt Durham (<https://ppp.museum-map.research.room3b.eu/>).

Weiterentwicklungen werden sich auf zwei Bereiche konzentrieren. Empfehlungen für ähnliche Objekte, um eine horizontale Navigation zwischen Objekten zu ermöglichen. Eine offene Frage ist was für Empfehlungen Nutzer:innen wollen: mehr ähnliche Objekte oder eine diversere Empfehlung. Der zweite Bereich ist die Integration digitaler Museumsführer, die von Kurator:inn:en und auch Nutzer:inn:en erstellt werden und einen alternativen Zugang bieten.

Bibliographie

Belkin, Nicholas J, Robert N Oddy, and Helen M Brooks. 1982. "ASK for Information Retrieval: Part I. Background and Theory." *Journal of Documentation* 38 (2): 61-71.

Google Arts & Culture. 2022. "Experiments: Curator Table." July 22, 2022. <https://artsexperiments.withgoogle.com/curatortable/>.

Gortana, Flavio, Franziska von Tenspolde, Daniela Guhlmann, and Marian Dörk. 2018. "Off the Grid: Visualizing a Numismatic Collection as Dynamic Piles and Streams." *Open Library of Humanities* 4 (2): 1-25.

Hall, Mark Michael, Oier Lopez de Lacalle, Aitor Soroa, Paul D Clough, and Eneko Agirre. 2012. "Enabling the Discovery of Digital Cultural Heritage Objects through Wikipedia." In *Proceedings of the LaTeCH Workshop Held at EACL 2012*.

Hall, Mark Michael, and David Walsh. 2021. "Exploring Digital Cultural Heritage through Browsing." In *Information Organization in Digital Humanities: A Global Perspective*, edited by Koraljka Golub and Ying-Hsang Liu. Taylor & Francis.

Mayr, Eva, Paolo Federico, Silvia Miksch, Günther Schreder, Michael Smuc, and Florian Windhager. 2016. "Visualization of Cultural Heritage Data for Casual Users." In *IEEE VIS Workshop on Visualization for the Digital Humanities*. Vol. 1.

Ruecker, Stan, Milena Radzikowska, and Stéfan Sinclair. 2011. *Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-Prospect Browsing*. Routledge.

Shneiderman, Ben, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 1998. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson.

The British Museum and Google Cultural Institute. 2022. "The Museum of the World." July 22, 2022. <https://britishmuseum.withgoogle.com/>.

Walsh, David, Mark Michael Hall, Paul Clough, and Jonathan Foster. 2020. "Characterising Online Museum Users: A Study of the National Museums Liverpool Museum Website." *International Journal on Digital Libraries* 21(1): 75–87. <https://doi.org/10.1007/s00799-018-0248-8>.

Walker, William S. 2022. "History Museums: Enhancing Audience Engagement through Digital Technologies." *Handbook of Digital Public History*: 165.

Whitelaw, Mitchell. 2015. "Generous Interfaces for Digital Cultural Collections." *Digital Humanities Quarterly* 9 (1). <http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html>.

Sharing the CROWN – Von Sammlungsdaten zu Linked Open Research Data

Grießer, Martina

martina.griesser@khm.at
Kunsthistorisches Museum, Wien

Hanzer, Helene

helene.hanzer@khm.at
Kunsthistorisches Museum, Wien

Kirchweger, Franz

franz.kirchweger@khm.at
Kunsthistorisches Museum, Wien

Kloser, Peter

peter.kloser@khm.at
Kunsthistorisches Museum, Wien

Lamers, Teresa

teresa.lamers@khm.at
Kunsthistorisches Museum, Wien

Pollin, Christopher

christopher.pollin@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz

Scholger, Martina

martina.scholger@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz

Steiner, Elisabeth

elisabeth.steiner@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz

Vasold, Gunter

gunter.vasold@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz

Die Reichskrone des Heiligen Römischen Reiches ist eines der wichtigsten Symbole europäischer Geschichte. Heute ist sie Teil der Sammlungen des Kunsthistorischen Museums (KHM) in Wien. Im Zuge des vom KHM initiierten CROWN-Projekts¹ wird eine umfassende Analyse der Reichskrone durchgeführt. Ziel dieser Analyse ist es nicht nur den konservatorischen Status des Objektes zu bestimmen, sondern auch die Diskussion um Entstehungszeit und Entstehungsort voranzubringen. Dazu werden alle Bestandteile der Krone, die Platten, das Stirnkreuz, der Bügel, die Edelsteine und Zierelemente etc. aus unterschiedlicher Perspektive - naturwissenschaftlich, konservatorisch, (kunst)historisch - analysiert. Dieses interdisziplinäre Vorhaben läuft noch bis 2024.

Im Rahmen des CROWN-Projekts wird ein einzelnes Objekt - die Reichskrone - im Vergleich zu einigen wenigen ausgewählten Vergleichsobjekten eingehend und möglichst umfassend beschrieben und analysiert. Die Forschungsdaten, die sich aus der Anwendung naturwissenschaftlicher Analysetechniken zur Untersuchung der Herstellungstechniken und der verwendeten Materialien ergeben, werden mit The Museum System (TMS)² erfasst. TMS ist eine weit verbreitete, aber proprietäre Softwarelösung, die für Museen entwickelt wurde. Sie bietet eine relationale Datenbank, die für die Inventarisierung, Dokumentation und Verwaltung von Sammlungen verwendet wird. Die Verwendung von TMS ist nicht nur auf vorhandene Ressourcen zurückzuführen, sondern hat pragmatische Gründe: Die Etablierung eines weiteren Systems zur Erfassung kann im jetzigen Projektkontext nicht erfolgen. Bei den im CROWN-Projekt erfassten Daten handelt es sich nicht um Inventardaten, also deskriptive Daten, die Objekte in einer Sammlung beschreiben und wofür TMS primär entwickelt wurde, sondern um hochspezifische Forschungsdaten. Ziel dieses Beitrags ist es, einen Workflow zur Überführung der in TMS erfassten Daten zu beschreiben, an dessen Ende eine hochstrukturierte, den FAIR-Kriterien entsprechende und Linked-Open-Data-fähige (LOD) Ressource steht.

LOD beschreibt die freie und offene Zurverfügungstellung von strukturierten Daten über das Web. Oft wird der Technologiestack des Semantic Web verwendet. Fundamentale Grundlage ist dabei das Resource Description Framework (RDF)³, das ein standardisiertes und graphenbasiertes Modell zur Darstellung und zum Austausch von Daten und deren Semantik.

Die verschiedenen Forschungsfragen können nur im Rahmen einer interdisziplinären Untersuchung der gemeinsamen Auswertung und Interpretation des Datenmaterials, zu dem auch historische Bild- und Schriftquellen gehören, beantwortet werden. Diese wiederum, d.h. die Verbindung von wissenschaftlichen Messungen und Quellen, geht weit über die übliche Beschreibung von Objekten in Sammlungen hinaus. Da den Autor*innen für dieses Vorhaben kein geeigneter Standard bekannt ist, wird ein Semantic-Web-Ansatz verfolgt. Dieser beinhaltet die Entwicklung einer domänenspezifischen Ontologie und deren Anbindung an die Top-Level-Ontologien CIDOC-CRM⁴ und Basic Formal Ontology (BFO)⁵. Im Ontology Engineering Prozess wird auf etablierte Ontology Design Patterns⁶ und Modellierungswerkzeuge zurückgegriffen. Die Domäne umfasst die formale Abbildung der im CROWN-Projekt erfassten Daten, also die Zusammenführung technologischer Untersuchungen und naturwissenschaftlicher Analysen mit Ergebnissen der historischen bzw. kunsthistorischen Forschung. Folgendes Beispiel soll die Domäne veranschaulichen:

Eine "Hochfassung für Perle mit Einsteckstiften" ist eine Komponente an einer Platte der Reichskrone. Dabei handelt es sich um eine 30 x 30 x 20 mm große Fassung aus Gold. Fünf Punkte auf dieser Fassung werden im Zuge des CROWN-Projektes mit drei Analyseverfahren untersucht: 3D Mikroskopie, Röntgenfluoreszenzanalyse (XRF) und Multispectral Imaging (MSI). Jede Analyse umfasst eine Beschreibung des Messvorgangs, Messdaten in tabellarischer Form als CSV oder XLSX, sowie Ergebnisse in Form von Diagrammen als bspw. BMP, sowie eine (kon-)textuelle Interpretation bzw. Schlussfolgerung der Messung durch die Fachwissenschaftler*in. Zusätzlich gibt es ggf. eine historische Quelle, etwa eine überlieferte Beauftragung eines Goldschmieds, in der weitere Informationen enthalten sind, oder eine kunsthistorische Quelle in Form eines Gemäldes in dem die Fassungen zu einem bestimmten Zeitpunkt abgebildet sind. Ziel der domänenspezifischen Ontologie ist es, genau diese Informationen formal zusammenzuführen und als Forschungsressource nachnutzbar zu machen.

Der dabei entwickelte Best-Practice-Workflow kann von anderen Forschungsvorhaben in Museen, die TMS nutzen, nachgenutzt werden. Dieser beinhaltet das Ontology Engineering, also die Umsetzung der Domäne als Ontologie und den Export und die Transformation von TMS nach LOD mittels Python. Python erlaubt weiters ein Semantic Enrichment, also die Normalisierung von Entitäten mittels Reconciliation durch Wikidata, sowie andere kontrollierte Vokabularien (z. B. Getty⁷ oder GND⁸). Am Ende dieser Überlegungen steht nicht eine Software, sondern ein beschriebener Workflow, in dem unterschiedliche Komponenten modular austauschbar sind. Das heißt konkret: statt TMS könnte auch ein anderes System zur Verwaltung und Erfassung von Inventardaten stehen, oder auch eine andere Programmierspra-

che verwendet werden als Python, mit der man in der Lage ist, programmatisch über einen Export oder eine Schnittstelle Daten nach RDF zu überführen. Der Fokus liegt auf der Interoperabilität verschiedener Werkzeuge und der Gestaltung dieser Schnittstellen. Die Autor*innen sind sich bewusst, dass ein Modell, das dieses Vorhaben auf generischer Ebene löst (Top Level Ontology), von zentralem Interesse ist. Das gegenständliche Projekt ist als eine prototypische Vorarbeit zu einem solchen Vorhaben zu verstehen.

Fußnoten

1. CROWN. Untersuchungen zu Materialität, Technologie und Erhaltungszustand der Wiener Reichskrone. <https://www.khm.at/erfahren/forschung/forschungsprojekte/restauratorisch-konservatorische-projekte/crown-untersuchungen-zu-materialitaet-technologie-und-erhaltungszustand-der-wiener-reichskrone/>, 20.06.2022.
2. The Museum System, <https://www.gallerysystem-s.com/solutions/tms-classic/>, 20.06.2022.
3. RDF 1.1 Concepts and Abstract Syntax, <https://www.w3.org/TR/rdf11-concepts/>, 13.12.2022.
4. CIDOC Conceptual Reference Model, <https://www.cidoc-crm.org/>, 20.06.2022.
5. Basic Formal Ontology, <https://basic-formal-ontology.org/>, 20.06.2022.
6. Ontology Design Patterns, <http://ontologydesignpatterns.org/wiki/>, 13.12.2022.
7. Getty Vocabularies, <https://www.getty.edu/research/tools/vocabularies>, 13.12.2022.
8. Gemeinsame Normdatei, <https://www.dnb.de>, 13.12.2022.

Bibliographie

Boettiger, Carl. "rdflib: A high level wrapper around the redland package for common rdf applications" (Version 0.1.0). Zenodo, 2018, <https://doi.org/10.5281/zenodo.1098478>.

Bruseker, George, Nicola Carboni, and Anaïs Guillem. "Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM". In *Heritage and Archaeology in the Digital Age. Quantitative Methods in the Humanities and Social Sciences*, edited by M. Vincent, V. López-Menchero Bendicho, M. Ioannides, T. Levy, 93-131. Cham: Springer, 2017, https://doi.org/10.1007/978-3-319-65370-9_6.

Cavalcante, J. "reconciler 0.2.1". GitHub, 2021, <https://github.com/jvfe/reconciler>.

Delpauch, Antonin. *Running a Reconciliation Service for Wikidata*. In *Proceedings of the 1st Wikidata Workshop (Wikidata 2020)*, edited by Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, Denny Vrandečić. Virtual Conference, November 2-6, 2020, <http://ceur-ws.org/Vol-2773/paper-17.pdf>.

Doerr, Martin, Richard Light, and Gerald Hiebel. "Implementing the CIDOC Conceptual Reference Model in RDF". Version 1.0. 2020, <https://cidoc-crm.org/sites/default/files/Implementen>

ting%20the%20CIDOC
%20Conceptual%20Reference%20Model%20in
%20RDF.pdf.

Hitzler, Pascal, "A review of the semantic web field", Communications of the ACM. 64(2), (February 2021): 76-83, <https://doi.org/10.1145/3397512>.

Kamrin, Janice, and Jennie Choi, "Taking Advantage of TMS", CIPEG Journal: Ancient Egyptian & Sudanese Collections and Museums, No. 3, (2019): 17-25, <https://doi.org/10.11588/cipeg.2019.3.66053>.

Otte, J. Neil, John Beverley, and Alan Ruttenberg, "BFO: Basic Formal Ontology", Applied Ontology, 17 (1), (2022): 17-43, <https://doi.org/10.3233/ao-220262>.

Vlachidis, A., A. Antoniou, A. Bikakis, and M.M. Terras. "Semantic metadata enrichment and data augmentation of small museum collections following the FAIR principles". In Information and Knowledge Organisation in Digital Humanities, edited by Koraljka Golub and Ying-Hsang Liu, London & New York: Routledge 2021. 106-129.

Wilkinson, Mark D. et al., "The FAIR Guiding Principles for scientific data management and stewardship", Sci Data 3:160018. (2016), <https://doi.org/10.1038/sdata.2016.18>.

Standoff-Tools – Generische Dienste für die automatische Annotation von XML-Dokumenten mit Plain-Text-Werkzeugen

Lück, Christian

christian.lueck@uni-muenster.de

Westfälische Wilhelmsuniversität Münster, Deutschland

TEI-XML ist eine exzellente Technologie für digitale Editionen. Aber für die computationale Analyse ist ein Korpus von XML-Dokumenten keine günstige Grundlage, weil Algorithmen erheblich komplexer werden, wenn statt linearem plain text ein XML-Baum durchlaufen werden muss. Zwar ist die Extraktion von plain text aus XML nicht aufwendig, aber für die Rückführung von Analyse-Ergebnissen in den XML-Baum gibt es bislang keine allgemeine Lösung. Dennoch ist eine solche in vielen Fällen wünschenswert, weil dann Analyse-Ergebnisse und die Struktur des Dokuments in einer allgemeinen Abfragesprache wie XQuery ausgewertet werden können. Es existieren Insellösungen zum Enrichment von TEI-XML, die auf bestimmte Tools ausgerichtet sind, etwa Spacy (Andorfer und Schlögl 2021, Meyer 2022), oder auf bestimmte Anwendungsfälle, etwa Alliterationen (Consalvi und Fumagalli 2022). Wünschenswert wäre jedoch, ganz allgemein Analyse-Tools für plain text entwickeln zu können und sie auch zum Enrichment von XML einsetzen

zu können. Genau dies ermöglichen die hier vorgestellten *StandOff Tools*.

Komponenten für Annotationspipelines

Die *StandOff Tools* bestehen aus zwei Komponenten:¹ dem Extraktor *E* und dem Internalizer *I*. In einer Annotationspipeline ist zwischen *E* und *I* ein anwendungsspezifischer Plain-Text-Tagger *T* geschaltet. (Abb. 1) *E* extrahiert aus dem XML-Quelldokument plain text, der an *T* geleitet wird. *I* nimmt von *T* gelieferte Referenzierungen von Plain-Text-Fragmenten entgegen und bringt sie als Inline-Markup stets so in den XML-Baum ein, dass wohlgeformtes XML herauskommt.

E und *I* sind insofern generische Komponenten, als dass ein beliebiger Tagger in eine Pipeline zum Enrichment von XML eingesetzt werden kann, solange er folgende Anforderungen erfüllt: a) Er muss plain text verarbeiten, b) er muss darin Textpassagen (ranges) per *character offsets* referenzieren, genauer: nach den in RFC 5147, Abschnitt 2.1.1 und 2.2.2 beschriebenen Regeln für *character ranges*.² Die vom Tagger ausgegebenen Ranges dürfen einander umfassen oder überlappen, so dass auch komplexe Strukturen in der XML-Quelle ausgezeichnet werden können. Die Ranges werden von *I* so zerschnitten, dass Überlappungen mit anderen Ranges und dem internen Markup der XML-Quelle aufgelöst werden (Splitting).

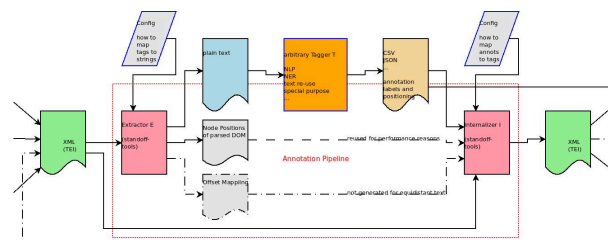


Abb. 1: Schema einer Annotationspipeline für TEI XML

Ansätze

Die Komponenten *E* und *I* müssen so aufeinander abgestimmt sein, dass *I* hinreichend Informationen hat, um die Annotationen von *T* in das XML-Dokument einzubauen. Neben Ansätzen, die ein Sprachmodell, und sei es auch nur eine Tokenisierung, einführen und insofern nicht generisch sind (Schopper 2021, Andorfer und Schlögl 2021), lassen sich in der DH-Software-Entwicklung zwei Ansätze beobachten.

Nach dem ersten extrahiert *E* die Textknoten und sammelt dabei Informationen darüber, wo Elemente anfangen und aufhören. Dies ist auf DOM-Ebene möglich. *I* kann dann Blätter des DOM-Baums so ersetzen, dass an ihrer Stelle neue Teilbäume die (zerschnittenen) Annotationen darstellen. Diesen Ansatz verfolgen die Lösung von Meyer (2022) und der Prototyp von Lassner (2021).

Sein Vorteil ist, dass gängige XML-Parser eingesetzt werden können. Sein Nachteil ist, dass manche Merkmale der serialisierten XML-Quelle nicht bewahrt werden können, weil sie nicht zur DOM-Spezifikation gehören: *character* und *entity references* sowie *whitespace* in Tags.

Die *StandOff Tools* verfolgen einen anderen Ansatz: Hier speichert *E* Informationen über die Position, die jedes Zeichen des extrahierten Textes im XML-Quelldokument gehabt hat (*offset mapping*). DOM-Manipulationen erfolgen nicht. Ein Parser ermittelt Positionsdaten aus dem XML-Quelldokument. Die Extraktion erfolgt dann im Wesentlichen durch Kopieren von Zeichenketten aus dem serialisiert vorliegenden XML-Quelldokument. *I* kann aufgrund des *offset mapping* die Positionsdaten des Taggers auf die XML-Quelle beziehen. Wie im ersten Ansatz zerschneidet *I* die Annotationen. Das Splitting, das Herzstück des Algorithmus, basiert auf einer Auswertung der Positionsdaten (offsets), die der XML-Parser für das interne Markup und der Tagger *T* für die Annotationen liefert. Anschließend werden aufgrund der Positionsdaten Portionen der XML-Quelle in die Ausgabe kopiert und neue Tags dazwischen gesetzt. Dieser Ansatz hat den Nachteil, dass kein üblicher XML-Parser eingesetzt werden kann. Aber er reproduziert diejenigen Merkmale der XML-Quelle, welche im ersten Ansatz verloren gehen.

Einsatz für manuelle Annotationen

I kann auch allein betrieben werden, um manuelle Annotationen, die Passagen eines XML-Dokuments per character offsets referenzieren, zu internalisieren und zu anschließend zu visualisieren oder auszuwerten. Solche Annotationen können z.B. als Web Annotations (OA-Ontology) realisiert werden. Wenn dabei allerdings wie in CATMA eine schlichte Extraktion von Textknoten aus HTML oder XML erfolgt, gehen dabei Informationen verloren, welche für eine Internalisierung der Annotationen erforderlich sind. Insofern bleibt die Internalisierung von CATMA-Annotationen in TEI noch Desiderat (Cayless 2019). Dennoch weisen die verschiedenen neuen Entwicklungsansätze einen Weg abseits vom nicht-funktionalen Standoff-Konzept der TEI (Banski 2010), auf welchem der Mehrwert sowohl von Standoff-Annotationen als auch von TEI realisierbar ist.

Fußnoten

1. Die Komponenten liegen derzeit als Kommandozeilen-Programme vor. Eine Implementierung als Web-services ist in Planung, wobei sich die Modellierung des Informationsflusses komplexer gestaltet, weil der Extraktor einen mehrfachen Output hat.
2. Derzeit steht als Eingabe-Format für den Internalizer CSV zur Verfügung. Eine Implementierung der Syntax von RFC 5147 ist geplant.

Bibliographie

Andorfer, Peter und Schlögl, Matthias. 2021. acdh-spacytei. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". Aufgerufen am: 3.8.2022. Handle: hdl.handle.net/11471/562.50.2.

Banski, Piotr. 2010. "Why TEI stand-off annotation doesn't quite work and why you might want to use it nevertheless." Balisage: The Markup Conference 2010, 10.4242/BalisageVol5.Banski01

Cayless, Hugh. 2019. "Implementing TEI Standoff Annotation in the browser." Proceedings of Balisage: The Markup Conference 2019, no. 23, 10.4242/BalisageVol23.Cayless01

Consalvi, A. und Fumagalli, S. 2022. "Alliteration: automatic identification and encoding." 2022 TEI Conference. Conference abstract, Session 1A-2, (im Erscheinen)

Lassner, David. "The Standoff Converter. A standoff-based approach to work on TEI documents in Python that connects the world of digital philology with NLP." 2021 TEI Conference and Members' Meeting.

Meier, Wolfgang. 2022. "Names sell. Named Entity Recognition in TEI Publisher", <https://e-editiones.org/blog/> (zugegriffen: 3. August 2022)

Schopper, Daniel. 2021. xsl-tokenizer. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". Aufgerufen am: 3.8.2022. Handle: hdl.handle.net/11471/562.50.216.

Urheberrechtlich geschützte Texte nachnutzen – Der XSample-Workflow

Andresen, Melanie

melanie.andresen@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Gaertner, Markus

markus.gaertner@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Jacke, Janina

janina.jacke@uni-goettingen.de
Universität Stuttgart, Deutschland

Ketschik, Nora

nora.ketschik@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pichler, Axel

axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart, Deutschland

Eines der Hindernisse, die der freien Weitergabe von Forschungsdaten im Sinne der Open-Science-Bewegung im Wege stehen, ist das Urheberrecht (UrhG). Dieses erschwert die Einhaltung der guten wissenschaftlichen Praxis und der FAIR-Prinzipien (Wilkinson et al. 2016) insbesondere bei Forschung zu zeitgenössischen Texten. Für urheberrechtlich geschützte Texte besteht bisher nur die Möglichkeit, sog. abgeleitete Textformate (Schöch et al. 2020) für „non-consumptive access“ (Organisciak und Downie 2021) zu veröffentlichen, etwa in Form von Frequenzlisten. Geisteswissenschaftliche Forschungsprojekte sind allerdings häufig auf die Verfügbarkeit von textuellem Kontext angewiesen, der erst eine angemessene Interpretation der Daten erlaubt. Um die Nutzbarkeit von Textdaten in dieser Hinsicht zu unterstützen, wurde im Projekt XSample¹ ein Workflow entwickelt, der auf das Recht zur Weitergabe von Textauszügen aufbaut, und dabei ermöglicht, die Auszüge auf das eigene Forschungsanliegen hin zu optimieren.

Rechtslage

Im deutschen UrhG unterliegen Texte bis 70 Jahre nach dem Tod der Autor*innen dem urheberrechtlichen Schutz, der die Vervielfältigung und die öffentliche Zugänglichmachung von Texten erheblich einschränkt. Hiervon sind vor allem zeitgenössische Texte betroffen, die aus diesem Grund womöglich gar nicht erst als Gegenstand von Forschungsprojekten in Erwägung gezogen werden. Seit 2018 ist zumindest die Verwendung von urheberrechtlich geschützten Texten zu Zwecken des Text- und Dataminings durch § 60d UrhG legitimiert (vgl. Raue 2021). Die Nachnutzung der Daten nach Abschluss eines Projektes ist aber weiterhin nur unklar geregelt (vgl. Kleinkopf et al. 2021, Andresen et al. 2022).

Der XSample-Workflow kombiniert den § 60d UrhG mit § 60c Abs. 1 Nr. 1 UrhG, der es erlaubt, bis zu 15 Prozent von Werken und auch vollständige Werke geringen Umfangs zu Zwecken der nicht-kommerziellen wissenschaftlichen Forschung zu vervielfältigen und an bestimmt abgegrenzte Personenkreise für deren eigene wissenschaftliche Forschung weiterzugeben.

Weitergabe von Auszügen im XSample-Workflow

Die Weitergabe von nur 15 Prozent eines Textes erscheint auf den ersten Blick nicht hinreichend. Um die Nützlichkeit dieser Textauszüge zu maximieren, wird im XSample-Workflow über eine Benutzeroberfläche eine gezielte Textauswahl unterstützt. So können Forschende die Textauswahl genau auf die eigenen Forschungsanliegen zuschneiden. Dafür werden auch im Korpus enthaltene Annotationen berücksichtigt, sodass bei Interesse an einem bestimmten Phänomen systematisch die

mit den relevanten Kategorien annotierten Textstellen extrahiert werden können (vgl. Gärtner 2020). Die Möglichkeiten und Grenzen des Auszugskonzepts sind am Beispiel zweier Anwendungsfälle aus der Literaturwissenschaft und Linguistik erprobt worden (vgl. Andresen et al. 2022). Sie zeigen beispielsweise, dass viele Forschungsfragen davon profitieren, wenn abgeleitete Textformate, die quantitative Analysen auf dem Gesamtkorpus ermöglichen, und Auszüge, die die qualitative Interpretation erlauben, kombiniert werden.

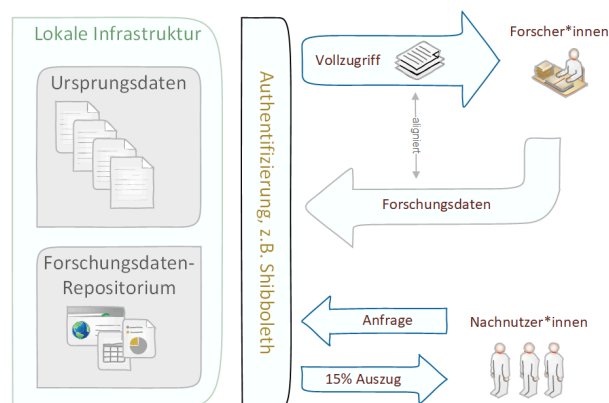


Abb. 1: Der XSample-Workflow im Überblick.

Abbildung 1 fasst das Auszugskonzept zusammen: Forscher*innen übermitteln ihre (annotierten) Forschungsdaten an Forschungsinfrastruktureinrichtungen (z. B. wissenschaftliche Bibliotheken), die diese verwalten. Nachnutzer*innen können Zugriffsanfragen an die Infrastrukturbetreiber stellen, um Auszüge aus den Forschungsdaten zu erhalten.

Der in Abbildung 1 beschriebene Workflow wurde im Rahmen des XSample-Projekts prototypisch in Form eines Webservices implementiert.² Die urheberrechtlich geschützten Korpusdaten liegen hierbei sicher in einem Dataverse³-Repositorium. Öffentlich auffindbare Metadaten ermöglichen Nutzer*innen den Einstieg in den XSample-Workflow und leiten auf einen gesonderten Server weiter, wo die eigentliche Auszugserstellung stattfindet. Hierfür stehen verschiedene Verfahren zur Auswahl, die dabei helfen können, die Nützlichkeit der Auszugsdaten für die eigenen Bedarfe zu erhöhen. Für die anfragebasierte Auszugserzeugung setzt unsere Implementierung auf ein Java-Framework für Korpusanfragen (Gärtner 2020), wodurch gezielt nach bestimmten annotierten Phänomenen gesucht werden kann, die dann als Grundlage für die individuelle Auswahl der Auszüge dienen. Während des gesamten Prozesses haben Nutzer*innen keinen direkten Zugriff auf die geschützten Daten und erhalten vor dem Download ihres individuellen Auszugs lediglich grafische Veranschaulichungen der Auszugskomposition basierend auf dem gewählten Verfahren und/oder den Suchergebnissen (vgl. Andresen et al. 2022).

Fazit

„Open Access“ ist und bleibt die Zielvorstellung für offene und reproduzierbare Forschung auch in den digitalen Geisteswissenschaften. Das hier vorgestellte Auszugskonzept stellt demgegenüber eine „Behelfslösung“ dar, um die Nachnutzung urheberrechtlich geschützter Daten zu ermöglichen und der Tendenz entgegenzuwirken, diese Texte per se aus Forschungsprojekten auszuschließen (vgl. Gärtner et al. 2021). Die Lösung ist an der Forschungspraxis der digitalen Literatur- bzw. Geisteswissenschaften ausgerichtet, für die andere „verfremdende“ Verfahren wie abgeleitete Textformate (Schöch et al. 2020, Organisciak und Downie 2021) nur eingeschränkt nachnutzbar sind. Das Auszugskonzept ermöglicht hingegen eine größere Nähe zum Text, indem die ursprüngliche Textgestalt beibehalten wird, die für die Interpretation der Daten häufig unabdingbar ist. Darüber hinaus wird der rechtliche Rahmen durch die individuelle Auszugsgenerierung optimal ausgeschöpft und den individuellen Forschungsinteressen angepasst.

Fußnoten

1. <https://www.izus.uni-stuttgart.de/fokus/fdm-projekte/xsample/> (Stand: 14.12.2022).
2. Langfristig soll dieser Dienst an der Universitätsbibliothek Stuttgart angeboten werden. Die Software für den Prototypen ist unter <https://github.com/ICARUS-tooling/xsample-server> open-source verfügbar und kann somit auch von anderen Einrichtungen genutzt werden, um eine eigene Instanz des XSample-Servers zu betreiben.
3. Dataverse ist eine weit verbreitete Open-Source-Repositoryssoftware, die über eine für XSample notwendige feingranulare Rechteverwaltung verfügt und eine simple Schnittstelle zur Integration externer Services in die Benutzeroberfläche bietet.

Bibliographie

Andresen, Melanie, Markus Gärtner, Sybille Hermann, Janina Jacke, Nora Ketschik, Felicitas Kleinkopf, Jonas Kuhn und Axel Pichler. 2022. „Vorzüge von Auszügen – Urheberrechtlich geschützte Texte in den digitalen Geisteswissenschaften (nach-)nutzen.“ *Zeitschrift für digitale Geisteswissenschaften*. <https://doi.org/10.17175/2022.007>.

Gärtner, Markus. 2020. „The Corpus Query Middleware of Tomorrow – A Proposal for a Hybrid Corpus Query Architecture.“ In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, 31–39. DOI: <https://www.aclweb.org/anthology/2020.cmlc-1.5>.

Gärtner, Markus, Felicitas Kleinkopf, Melanie Andresen, Sybille Hermann. 2021. „Corpus Reusability and Copyright – Challenges and Opportunities.“ In *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, 10–19. DOI: <https://doi.org/10.14618/ids-pub-10467>.

Kleinkopf, Felicitas, Janina Jacke und Markus Gärtner. 2021. „Text- und Data-Mining – Urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora bei computergestützten Verfahren und digitalen Ressourcen.“ *MMR Zeitschrift für IT-Recht und Recht der Digitalisierung* 24, H. 3: 196–200. DOI: <http://dx.doi.org/10.18419/opus-11445>.

Organisciak, Peter und J. Stephen Downie. 2021. „Research access to in-copyright texts in the humanities.“ In *Information and Knowledge Organisation in Digital Humanities*, hg. von Koraljka Golub und Ying-Hsang Liu, 157–177. Digital Research in the Arts and Humanities. London, New York: Routledge.

Raue, Benjamin. 2021. „Die Freistellung von Datenanalysen durch die neuen Text und Data Mining-Schranken.“ *Zeitschrift für Urheber- und Medienrecht* 56, H. 10: 793–802.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen.“ *ZfdG* 5. DOI: <https://doi.org/10.17175/2020.006>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg et al. 2016. „The FAIR Guiding Principles for scientific data management and stewardship.“ *Sci Data* 3, H. 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>.

Vom Finden, Filtern und Auswerten der relevanten Daten im digitalen Nachlass von Friedrich Kittler im Deutschen Literaturarchiv Marbach

Holz, Alex

alex.holz@dla-marbach.de

Deutsches Literaturarchiv Marbach, Deutschland

Çakir, Dîlan Canan

dilan.cakir@dla-marbach.de

Deutsches Literaturarchiv Marbach, Deutschland

Auf unserem Poster zeigen wir, wie wir von einem digitalen Nachlass mit 3,3 Millionen Dateien zu einem Data-Set mit etwa 30.000 Dateien gelangt sind, mit dem wir sinnvoll zum Born-digital-Nachlass des Literaturwissenschaftlers und Medientheoretikers Friedrich Kittler (1943–2011), der im Deutschen Literaturarchiv Marbach (DLA) aufbewahrt wird, forschen können und dürfen. Mit

unserem Fallbeispiel wollen wir zeigen, wie umfangreich ein Born-digital-Nachlass im DLA sein kann, wie man mit diesem technisch, konservatorisch und rechtlich umgeht und welches Potential darin liegt. Anhand unseres kuratierten Arbeitskorpus können wir sodann beispielsweise an der statistischen Auswertung und Visualisierung der (technischen¹) Metadaten arbeiten und damit zur Erschließung des Nachlasses beitragen.²

Digitale Vor- und Nachlässe im DLA

Mit Kittlers Nachlass kam 2011 der bisher umfangreichste und komplexeste digitale Nachlass ins DLA. Immer noch gehört dieser mit 762 Datenträgern (648 Disketten, 100 optische Medien, 1 USB Speicher, 12 Festplatten) und einer Größe von 1,6 TB zum quantitativ größten digitalen Bestand im DLA. Berücksichtigt man Kittler nicht, umfassen die gesamten digitalen Vor- und Nachlässe von etwa 70 Bestandsbildner*innen im DLA (also im Archiv, nicht in der Bibliothek) derzeit mit etwa 1600 Datenträgern insgesamt knapp 5,2 TB.

In den Jahren 2012 bis 2013 galt es, diesen neuen Bestand analog zu sichten, zu sortieren und zu verzeichnen. Dieser Workflow wurde bereits dokumentiert und publiziert (Enge/Kramski 2017). Die PCs, Laptops und physischen Datenträger zu sortieren, reicht (anders als beispielsweise das Sortieren von Papiermanuskripten) allerdings nicht aus, um den Nachlass zu katalogisieren. Im DLA haben wir für digitale Vor- und Nachlässe eigens sogenannte „Digital Curator[s]“, die man auch Data Librarian oder „Datenarchäolog[innen]“ nennen könnte (Bülow/Kramski 2011: 159), die seit Mitte der 2000er Jahre immer wichtiger werden (Jaillant 2022: 8). Diese bereiten ‚rohe‘ digitale Nachlässe für verschiedene Anliegen auf. Bei digitalen Archivobjekten „muss die Lese- und Interpretationsfähigkeit [nämlich] zuerst aufwendig hergestellt werden.“ (Bülow/Kramski 2011: 159) Zudem ist eine manuelle Sichtung der Dateien bei einem kleineren digitalen Nachlass zwar vorstellbar, für Kittlers Nachlass allerdings wegen des Umfangs und der Art der Dateien eher unpraktisch.

Kittlers digitaler Nachlass: Von 3,3 Millionen zu ca. 30.000 Dateien

Bei erst kürzlich verstorbenen Autor*innen, wie Kittler, muss immer mitbedacht werden, dass selten der ganze Nachlass benutzbar ist und als Ganzes gar nicht oder zu einem späteren Zeitpunkt erforscht werden kann – bei Born-digital-Nachlässen spricht man so auch von „dark archives“, denn selten kann alles (online) leicht zugänglich gemacht werden, obwohl es bereits digital vorliegt. „Nobody would reasonably claim that all born-digital data should be unlocked and openly accessible. Yet, it is important to recognize that ‚dark‘ archives contain vast

amounts of data essential to scholars [...]“ (Jaillant 2022: 7).

Der aktuelle Kittler-Bestand umfasst ca. 3,3 Millionen Dateien. Der erste Schritt besteht in dem ‚Identifizieren und Aussortieren‘ von hunderttausenden nicht-unikaler³ Dateien, also Dateien, die nicht von Kittler selbst erstellt wurden. Dies erfolgte automatisiert vor allem über einen Abgleich mit der National Software Reference Library (NSRL) des NIST⁴. Für solche Vorgänge sind auch KI-gestützte Methoden denkbar; allerdings werden solche derzeit eher für „low-level tasks“ herangezogen, etwa bei der Identifizierung von sensiblen persönlichen Informationen (Jaillant 2022: 14). Damit gelangten wir zu etwa 2,25 Millionen Dateien.

In einem nächsten Schritt wurde die Menge auf die Dateien begrenzt, die von Seiten der Nachlassverwaltung begutachtet und mit einem Status versehen wurden (freigegeben, vorläufig gesperrt, gesperrt). Unser Arbeitskorpus ist also als Momentaufnahme zu sehen, da er nur auf den bereits bewerteten Dateien basiert. Ohne die bislang unbekannten bzw. noch nicht bewerteten Dateien kommen wir auf 219.989 Dateien.

Aus dieser Menge wurden zuletzt nun die für die Forschung vollständig freigegebenen Dateien (Metadaten und Inhalt) herausgefiltert, die zudem von Kittler erstellt wurden. Wir landeten bei etwa 30.000 Dateien, also etwa 0,88% von den 3,3 Millionen Dateien.

Bei den meisten Dateien in unserem Arbeitskorpus handelt es sich um Textdateien in unterschiedlichsten Formaten. Kittler hat „nicht nur Texte, Bilder und Videos hinterlassen, sondern auch Relikte seiner programmierenden Tätigkeiten.“ (Enge/Kramski 2017). Er interessierte sich neben der Literatur, Musik und Philosophie vor allem für Technologien, Medien und das Programmieren (Winthrop-Young, 2017: 210). „[W]ie ungezählte Teenager in dieser Zeit“ habe er vor allem mit Codes gespielt und autodidaktisch gelernt (Pias 2014: 39–44). Das spiegelt sich in gewisser Weise auch in seinem Nachlass wider. So befinden sich im Arbeitskorpus beispielsweise auch mit Kommentaren versehene Dateien der Programmiersprache C/C++ (insbesondere *.c und *.h-Dateien).

Fußnoten

1. Wir orientieren uns bei den Metadaten an der u.a. von Jenn Riley (2017) beschriebenen Kategorisierung. Die Technischen Metadaten sind eine Untergruppe der administrativen Metadaten. Es geht also um Informationen zu den Dateien selbst, nicht ihren Inhalt (deskriptive Metadaten). Neben FileSize und MimeType sind hier für uns besonders die Zeitstempel oder auch Informationen aus den Dateipfaden wichtig.
2. Dies ist Teil des Projekts „Archivierung, Erschließung und Erforschung von Born-digitals“ des Forschungsverbundes Marbach, Weimar, Wolfenbüttel, <https://www.mww-forschung.de/born-digitals> (letzter Zugriff 08.07.2022).
3. Als unikale Dateien bezeichnen wir Dateien, die von dem/der Bestandsbildner*in erstellt wurden und die außerhalb dieses digitalen Vor- oder Nachlasses nicht verfügbar sind.

4. National Software Reference Library (NSRL) des National Institute of Standards and Technology. <https://www.nist.gov/itl/ssd/software-quality-group/national-software-reference-library-nsrl> (letzter Zugriff am 22.07.2022).

Bibliographie

Bülow, Ulrich von/Kramski, Heinz Werner. „Es füllt sich der Speicher mit köstlicher Habe“ – Erfahrungen mit digitalen Archivmaterialien im Deutschen Literaturarchiv Marbach.“ In: Neues Erbe: Aspekte, Perspektiven und Konsequenzen der digitalen Überlieferung, hg. v. Caroline Y., Hauser Robert, Robertson-von Trotha, 141–62. Kulturelle Überlieferung – digital 1. Karlsruhe: KIT Scientific Publishing, 2011.

Enge, Jürgen/Kramski, Heinz Werner. „Friedrich Kittler’s Digital Legacy – PART I – Challenges, Insights and Problem-Solving Approaches in the Editing of Complex Digital Data Collections.“ Digital Humanities Quarterly 11, Nr. 2 (22. Mai 2017). <http://www.digitalhumanities.org/dhq/vol/11/2/000307/000307.html>.

Jaillant, Lise. „Introduction.“ In: Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections, hg. v. Jaillant, Lise, 7–28. Bielefeld, Germany: Bielefeld University Press / transcript Verlag, 2022.

Pias, Claus. „Friedrich Kittler und der ‚Mißbrauch von Heeresgerät‘. Zur Situation eines Denkbildes 1964–1984–2014.“ Merkur. Deutsche Zeitschrift für europäisches Denken 69, Nr. 791 (2015): 31–43.

Riley, Jenn. Understanding metadata. What is metadata and what is it for? Baltimore 2017.

Was heißt eigentlich ‚offen‘? Eine korpuslinguistische Untersuchung am Beispiel des bibliothekarischen Diskurses der SLUB Dresden

Meier-Vieracker, Simon

simon.meier-vieracker@tu-dresden.de
TU Dresden, Deutschland

Weigelt, Lucie

lucie.weigelt@tu-dresden.de
TU Dresden, Deutschland

Dutschke, René

rene.dutschke@tu-dresden.de
TU Dresden, Deutschland

Lasch, Alexander

alexander.lasch@tu-dresden.de
TU Dresden, Deutschland

Scherbaum, Stefan

stefan.scherbaum@tu-dresden.de
TU Dresden, Deutschland

Seemann, Sophia

sophia_marie.seemann@tu-dresden.de
TU Dresden, Deutschland

Pfeifer, Ulrike

ulrike_marie.pfeifer1@tu-dresden.de
TU Dresden, Deutschland

Hintergrund

Forschung zum Thema Open Science fokussiert typischerweise technische oder auch regulatorische und wissenschaftspolitische Aspekte (Vicente-Saez und Martinez-Fuentes 2018). Der Begriff der Openness bzw. der Offenheit selbst, der im Diskurs um digitale Forschung als zentrales Schlagwort fungiert, wird dabei eher selten in seiner Semantik reflektiert. Dabei führt der Begriff der Offenheit reichhaltige, auch alltagssprachlich verankerte Assoziationspotenziale mit sich, aus denen sich die ausgesprochen optimistische, vielleicht sogar utopisch-überhöhende Rahmung von Open Science (Dickel und Franzen 2015; Tkacz 2012) und Open Humanities maßgeblich speist.

Im interdisziplinären Projekt „Digitalisierung als Disruption von Wissenssystemen – Opening Knowledge“ (Dia-Disk; Laufzeit 10/2021 – 03/2025) im Rahmen des EXU-Verbundes „Disruption and Societal Change“ an der TU Dresden fragen wir nach den disruptiven Auswirkungen der Digitalisierung in den für die Wissensgesellschaft zentralen Institutionen Universität, Bibliothek und Schule. Wir gehen davon aus, dass sich gerade im Prinzip der Offenheit und seiner diskursiven Verhandlungen Deutungsmuster manifestieren, welche die Disruptionen, die digitale Technologien für die traditionellen Routinen in Bildung und Wissenschaft mit sich bringen, positiv als Chancen, Transformationen usw. rahmen (Koch, Nanz, und Pause 2016, 19). Diese Deutungsmuster nehmen wir in einem der linguistischen Arbeitspakete des Projektes, das sich den Open-Science-Aktivitäten der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB) widmet, aus einer korpuslinguistischen Perspektive in den Blick. Wir untersuchen, welche komplexen diskurssemantischen Profilierungen des Begriffs der Offenheit sich in diesem konkreten Diskurs nachweisen lassen, die seinen Schlagwortcharakter

(Schröter 2011) grundieren. Damit soll eine Schärfung des Begriffs ermöglicht werden, die es erlaubt, die oft strikt nach Pro und Kontra geführten Debatten über Openness differenzierter zu führen, nicht zuletzt in wissenschaftspolitischen Kontexten.

Daten und Methoden

Grundlage unserer Analysen bilden verschiedene Publikationen (Forschungsbeiträge, Geschäftsberichte, aber auch Blogbeiträge und Tweets) sowie Dokumente (Drittmittelanträge und Strategiepapiere) aus dem Umfeld der SLUB, die dezidiert aus der Perspektive der Institution formuliert sind. Wie prominent im Strategiepapier SLUB 2025 festgehalten ist, positioniert sich die SLUB proaktiv in der deutschsprachigen Bibliotheks- und Wissenschaftslandschaft als „Motor für offene Wissenschaft und Gesellschaft“ (Bonte und Muschalek 2019). Sie nutzt in ihren Selbstdarstellungen strategisch die vielfältigen, zumeist positiven Konnotationen des Begriffs(feldes) der Offenheit, weshalb sich diese für unsere Fragestellung besonders gut eignen.

Das laufend zu erweiternde Korpus (1,59 Millionen Tokens, Stand Juli 2022) stellen wir im Projekt in morphosyntaktisch annotierter Form über verschiedene digitale Analyseumgebungen wie die Korpusanalyseplattform CQPweb (Hardie 2012) und die kollaborative Annotationsumgebung INCEPTION (Castilho u. a. 2018) zur Verfügung. Für die Analyse nutzen wir zum einen korpuslinguistische Verfahren der datengeleiteten diskurssemantischen Analyse wie Kollokations- und Ngram-Analysen (Bubenhofer 2017), aber auch Word Embeddings als Verfahren der distributionellen Semantik mit der Software word2vec (Mikolov u. a. 2013; Kozłowski, Taddy, und Evans 2019), um die Gebrauchsprofile einschlägiger Lexeme zu erschließen. Zum anderen wählen wir mit der Frame-Semantik (Ziem 2020) einen stärker theoriegeleiteten Ansatz. Dabei handelt es sich um eine semantische Theorie, die die Bedeutung von sprachlichen Ausdrücken in Bezug auf das in untereinander vernetzten Frames organisierte Weltwissen der Sprechenden beschreibt. Frames sind dabei schematische Repräsentationen von Situationen oder Konstellationen, vor deren Hintergrund dann sprachliche Ausdrücke verstanden werden (Busse 2012). Ausgehend von der lexikographischen Ressource FrameNet (<https://framenet.icsi.berkeley.edu/>) annotieren wir im Textmaterial die semantischen Valenzen einschlägiger Ausdrücke und der durch sie evozierten Frames, so dass die unterschiedlichen Lesarten etwa von *offen* und die assoziierten Formulierungsmuster präzise erfasst werden können.

Erste Ergebnisse

Erste Ergebnisse zeigen, dass im untersuchten Diskurs das Begriffsfeld der Offenheit systematisch zwischen einer eher technisch auf digitale Daten und ihre Zugänglichkeit bezogenen Lesart, einer auch politisch aufgeladenen, auf Partizipation und Inklusion abzielenden Lesart sowie einer auf Offenheit als epistemische Tu-

gend abzielenden Lesart changiert. Typische Kollokate (Assoziationsmaß Log Dice) von *offen* sind etwa *Daten*, *Schnittstelle* und *Standards* im Sinne der technischen Lesart, *Kreativraum* und *Austausch* im Sinne der partizipativen Lesart, aber auch Abstrakta wie *Wissen* und *Neugier*. Paarformeln und Aufzählungen wie *offene und freie Wissensgesellschaft* oder *Offenheit, intellektuelle Freiheit und Redlichkeit* zeigen, dass das Begriffsfeld der Offenheit mit anderen Schlagwörtern in Interaktion gebracht und so zusätzlich semantisch als ethisch gehaltvolle Zielnorm aufgeladen wird. Auch die Berechnung von Word Embeddings weist in diese Richtung. Als sog. Nearest Neighbors von *offen*, die im untersuchten Diskurs also semantische Ähnlichkeit aufweisen, werden *nachhaltig*, *transparent*, *interdisziplinär*, *nachnutzbar*, aber auch *kreativ*, *vernetzt*, *modern* und *innovativ* ausgegeben (Abb. 1); zu *Offenheit* werden *Transparenz*, *Verbreitung*, *barrierefrei*, aber auch *demokratisch* ausgegeben (die Visualisierung des Modells ist unter <https://kurzelinks.de/dia-disk> öffentlich zugänglich; auch auf Teile der Korpora kann auf Anfrage Zugriff gewährt werden).

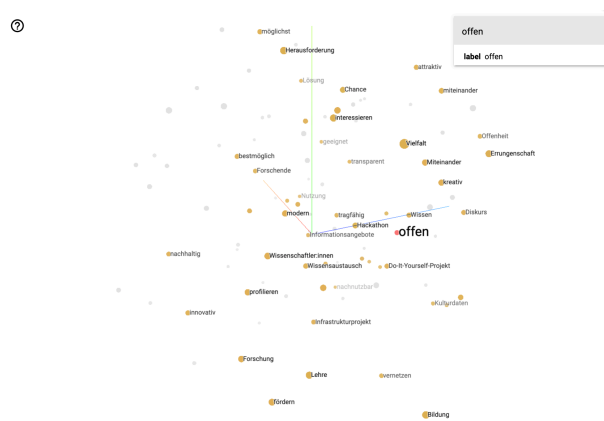


Abb. 1

Für die framesemantische Annotation bietet sich u.a. der Frame Openness an, der auf das Moment der Zugänglichkeit abzielt und sowohl wörtliche als auch metaphorische Verwendungen zulässt. Dabei zeigt sich in einer Pilotierung am Beispiel der Forschungsbeiträge im Korpus (51 Texte), dass die Frame-Elemente THEME (für wen ist etwas offen?) und BARRIER (was verhindert potentiell den Zugang?) nur in 11% bzw. 3% der insgesamt 72 Fälle explizit besetzt werden, so dass der Begriff der Offenheit vage gehalten und so seine freien Assoziationspotenziale besonders gut entfalten kann (Abb. 2).

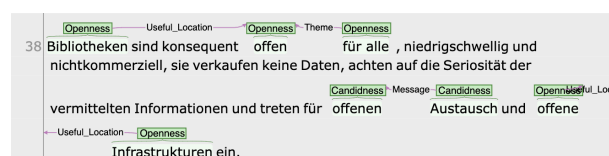


Abb. 2

Ausblick

Unsere Ergebnisse versprechen einen detaillierten und empirisch gestützten Blick auf das für die Digital Humanities so bedeutsame Prinzip der Offenheit weniger in seinen technischen als in seinen diskursiven Aspekten. Im Projekt werden unsere linguistischen Analysen ergänzt werden durch psychologische, experimentelle Erhebungen individueller Konstruktsysteme (Kelly 2005) von Akteur:innen im Feld der Open Science, die wir mit unseren Befunden zu diskursiven Deutungsmustern triangulieren werden. Zudem werden wir die Korpusgrundlage auf Publikationen anderer Akteur:innen im bibliothekarischen Diskurs ausweiten.

Bibliographie

Bonte, Achim und Antonie Muschalek, Hrsg. 2019. *SLUB 2025. Wissen teilen - Menschen verbinden. Strategie der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden*. Dresden: SLUB. <https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-357501>.

Bubenhofer, Noah. 2017. „Kollokationen, n-Gramme, Mehrworteinheiten“. In *Handbuch Sprache in Politik und Gesellschaft*, hg. von Kersten Sven Roth, Martin Wengeler und Alexander Ziem, 69–93. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110296310-004>.

Busse, Dietrich. 2012. „Frame-Semantik. Ein Kompendium“, Berlin u.a.: De Gruyter.

Castilho, Richard Eckart de, Jan-Christoph Klie, Naveen Kumar, Beto Boullosa und Iryna Gurevych. 2018. „INCEPTION - Corpus-based Data Science from Scratch“. In *Digital Infrastructures for Research (DI4R)* 2018. <http://tubiblio.ulb.tu-darmstadt.de/106982/>.

Dickel, Sascha und Martina Franzen. 2015. „Digitale Inklusion: Zur Sozialen Öffnung Des Wissenschaftssystems“. *Zeitschrift für Soziologie* 44 (5): 330–47. <https://doi.org/10.1515/zfsoz-2015-0503>.

Hardie, Andrew. 2012. „CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool“. *International Journal of Corpus Linguistics* 17 (3): 380–409. <https://doi.org/10.1075/ijcl.17.3.04har>.

Kelly, George A. 2005. „A Brief Introduction to Personal Construct Theory“. In *International Handbook of Personal Construct Psychology*, herausgegeben von Fay Fransella, 3–20. Chichester, UK: Wiley. <https://doi.org/10.1002/0470013370.ch1>.

Koch, Lars, Tobias Nanz und Johannes Pause. 2016. „Imaginationen der Störung: ein Konzept“. *Behemoth* 9 (1): 6–23. <https://doi.org/10.6094/BEHEMOTH.2016.9.1.885>.

Kozlowski, Austin C., Matt Taddy und James A. Evans. 2019. „The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings“. *American Sociological Review* 84 (5): 905–49. <https://doi.org/10.1177/0003122419877135>.

Mikolov, Tomas, Kai Chen, Greg Corrado und Jeffrey Dean. 2013. „Efficient Estimation of Word Representations in Vector Space“. arXiv:1301.3781 [cs], Januar. <http://arxiv.org/abs/1301.3781>.

Schröter, Melani. 2011. „Schlagwörter im politischen Diskurs“. *Mitteilungen des Deutschen Ger-*

manistenverbandes 58 (3): 249–57. <https://doi.org/10.14220/mdge.2011.58.3.249>.

Tkacz, Nathaniel. 2012. „From Open Source to Open Government: A Critique of Open Politics“. In *ephemera* 12(4): 386–405.

Vicente-Saez, Ruben und Clara Martinez-Fuentes. 2018. „Open Science Now: A Systematic Literature Review for an Integrated Definition“. *Journal of Business Research* 88 (Juli): 428–36. <https://doi.org/10.1016/j.jbusres.2017.12.043>.

Ziem, Alexander. 2020. „Wortbedeutungen als Frames: ein Rahmenmodell zur Analyse lexikalischer Bedeutungen“. In *Semantiktheorien II: Analysen von Wort- und Satzbedeutungen im Vergleich*, hg. von Jörg Hagemann und Sven Staffeldt, 27–56. Tübingen: Stauffenburg.

Weißbuch Digitale Edition

Galka, Selina

selina.galka@uni-graz.at
Karl-Franzens-Universität Graz, Österreich

Klug, Helmut W.

helmutwklug@gmail.com
Karl-Franzens-Universität Graz, Österreich

Die Digitale Edition stellt ein äußerst wichtiges Forschungsfeld innerhalb der Digital Humanities dar. Es gibt eine Reihe an Publikationen und Ressourcen, die sich sowohl einführend als auch theoriebildend diesem Thema widmen. So verfasste z.B. Patrick Sahle ein umfangreiches Standardwerk dazu (2013: *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.*), bzw. 2019 den einschlägigen Artikel „What is a scholarly digital edition?“ (2019). Ein weiteres Standardwerk, *Digital scholarly editing: theories, models and methods*, wurde 2015 von Elena Pierazzo publiziert; von ihr und James Driscoll wurde 2016 der Sammelband *Digital Scholarly Editing. Theory, Practice and Future Perspectives* herausgegeben. Daneben erschienen unzählige Artikel in einschlägigen Zeitschriften, wie z. B. zum Aspekt der Zukunft von Digitalen Editionen von Elena Pierazzo aus dem Jahr 2019, Digitale Edition im Zusammenhang mit rechtlichen Aspekten von Wout Dillen und Vincent Neyt 2016 oder zum Thema Kommentar in Digitalen Editionen (Bleier/Klug 2020) oder bestimmten Textsorten wie Briefen oder Tagebüchern (Dumont 2019). In diesem Zusammenhang müssen natürlich auch die Leistungen des Instituts für Dokumentologie und Editorik (<https://www.i-d-e.de/>) genannt werden, die mit dem Rezensionsorgan R.I.D.E. und den dafür geschaffenen Rezensionskriterien (Sahle 2014), wie auch einer mittlerweile umfangreichen Buchreihe (SIDE) federführend an der Theoriebildung und Weiterentwicklung Digitaler Edition arbeiten. Unterschiedliche Ausbildungsangebote wie das bereits 2017 ausgelaufene Digital Scholarly Editions Initial Training Network DiXiT (<https://dixit.uni-koeln.de/>) runden dieses Bild ab.

Anfang 2021 wurde als Abschluss des vom österreichischen Bundesministerium für Bildung, Wissenschaft und Forschung geförderten Projekts "KONDE - Kompetenznetzwerk Digitale Edition" eine als Nachschlagewerk konzipierte Publikation, das sogenannte *Weißbuch* zum Thema Digitale Edition veröffentlicht: <https://www.digitale-edition.at>. Im HRSM-Projekt wurde im Rahmen von thematisch einschlägigen Arbeitsgruppen (z.B. Transkription und Textauszeichnung, Netzwerkanalyse und Datamining, Archivierung oder Quellendigitalisierung) Inhalte und Lemmata des *Weißbuchs* erarbeitet. Die Weißbucheinträge wurden von den Autorinnen und Autoren mithilfe eines vorgefertigten Google-Doc-Templates, das eine überschaubare Menge an relevanten Formatierungen enthielt, erstellt. Nach der Redaktion durch das Projektteam wurden die Dokumente nach einem Export aus Google Drive mittels XSLT in das für das Projekt erarbeitete TEI/XML-Datenmodell transformiert und in das Geisteswissenschaftliche Asset Management System (GAMS) eingespeist, welches am Institut Zentrum für Informationsmodellierung in Graz entwickelt wurde (und wird). Die TEI-Dokumente stehen unter einem PID (Persistent Identifier) langzeitarchiviert und stabil referenzierbar zur Verfügung und wurden mit umfassenden Metadaten versehen. Die HTML-Ansicht des *Weißbuchs* wird direkt aus den Daten mittels XSLT generiert.

Das *Weißbuch* enthält neben 25 Portraits von Editionsprojekten über 200 Beiträge, verfasst von 50 Autorinnen und Autoren aus 10 österreichischen Forschungseinrichtungen, die Begriffe aus 12 Themenbereichen primär für Einsteiger in diese Thematik aufbereiten:

1. Editionswissenschaft im Allgemeinen: Erklärung allgemeiner Begriffe und Konzepte (Analysemethoden, Digitale Edition, Editionstext usw.), Editionstypen, Darstellung allgemeiner Themen wie Interpretation, Kollationierung, Normalisierung, Paläographie
2. Digitale Editionswissenschaft: Erklärung allgemeiner Begriffe und Konzepte (Ontologie, XML, Usability usw.), Analysemethoden, Apparat, diverse Editionstypen, Benutzerinnen und Benutzer Digitaler Editionen, Diplomatische Transkription, Editionstypographie, Elemente digitaler Editionen, FAIR-Prinzipien, Informationsarchitektur, Kataloge Digitaler Editionen, Kommentar, Lagenvisualisierung, Linked (Open) Data, Persistent Identifier, Social Edition, Textkritik in digitalen Editionen, Zielgruppen digitaler Editionen, Zitierbarkeit digitaler Ressourcen
3. Digitalisierung: Digitalisierungsdienste, Checkliste Digitalisierung, Transkriptionswerkzeuge, Kosten, Crowdsourcing, HTR, OCR
4. Metadaten: Metadatenformate (CIDOC CRM, METS, PREMIS), Metadaten Harvesting
5. Annotation und Modellierung: Datenmodellierung und -modelle, Modellierungsstandards, Schemata, Normdaten, Semantic Web, Markup und Markup Sprachen
6. NLP: Artikel über computergestützte natürliche Sprachverarbeitung (Distant Reading, Historische Korpora, Lemmatisierung, Named Entity Recognition, Tagger, Tagsets)
7. Schnittstellen: Design und Komponenten (Barrierefreies Webdesign, Benutzerinnen und Benutzer Digitaler Editionen, Editor-testing, Informationsarchitektur, Interface, Usability, Zitiervorschlag)
8. Datenanalyse: Analysemethoden, Datamining, Datenvisualisierung, Dramennetzwerkanalyse, Visualisierungstools
9. Archivierung: Archivierungsstrategien, digitale Nachhaltigkeit, österreichische Archivanbieter, Metadaten, Versionierung
10. Software und Softwareentwicklung: Software zur Erstellung digitaler Editionen und theoretische Aspekte der Softwareentwicklung
11. Rechtliche Aspekte in Bezug auf Digitalisierung, Bereitstellung von Digitalen Editionen, Urheberrecht, Lizenzierung und Lizenzmodelle
12. Institutionen: einschlägige Beschreibungen der am Projekt beteiligten Institutionen mit einer Darstellung der digital-editorischen Schwerpunkte.

Als Begleitmaterialien stehen unter anderem eine umfangreiche, öffentlich verfügbare Zotero-Bibliothek zur Verfügung, die weit mehr Literatur zum Thema beinhaltet, als in den Artikeln zitiert ist. Außerdem gibt es eine nach Anwendungsbereich systematisierte Sammlung von Tools, Ressourcen und Standards, die für das digitale Edieren relevant sein können.

Die einzelnen Einträge des *Weißbuchs* bieten neben der Erläuterung des jeweiligen Begriffes in der Regel Links zu verwandten Lemmata innerhalb des *Weißbuchs* bzw. zu Einträgen in den relevanten digitalen Lexika zum Thema Digitale Edition (*Edlex: Editionslexikon*, *Lexicon of Scholarly Editing*, *Parvum Lexicon Stemmatologicum*), einschlägiger Software oder prototypischen Projekten. Zusätzlich werden Literaturhinweise angeboten, die mit einer umfangreichen Zotero-Bibliothek verlinkt sind, die Literatur zur Digitalen Edition auflistet. Die Artikel werden sowohl über eine thematische Gliederung als auch über einen Index erschlossen; eine Volltextsuche ist ebenfalls implementiert.

Neben der Darstellung auf der Website als HTML werden die Weißbucheinträge auch als TEI/XML- bzw. als PDF-Download zur Verfügung gestellt. Ein Netzdiagramm veranschaulicht die Beziehungen zwischen den einzelnen Artikeln und die Häufigkeit, mit der ein Stichwort genannt wird. Inhaltlich sind die Artikel breit angelegt; sie richten sich primär an interessierte Laien, sie können aber auch für Fachleute aus der Forschung als Einstieg in die Thematik oder pointierte Zusammenfassung dienen.

Das *Weißbuch Digitale Edition* dient als Einstiegswerk in unterschiedlichste Aspekte des Forschungsfeldes, wobei bei jedem Weißbuchartikel weiterführende Literatur angegeben wird, damit sich Benutzerinnen und Benutzer selbstständig vertiefend in die Materie einlesen können. Es zeichnet sich außerdem durch eine starke hypertextuelle Vernetzung der Artikel untereinander aus, die einen explorativen Zugang für alle Nutzerinnen und Nutzer ermöglicht.

Nach der Publikation des *Weißbuchs* Anfang 2021 konnten die Herausgeber im Sommer 2022 eine Förderung für eine Aktualisierung (Überarbeitung und Ergänzung bestehender Artikel) bzw. für die Erweiterung (Verfassen neuer Artikel, Vorstellung weiterer einschlägiger Projekte) des *Weißbuchs Digitale Edition* einwerben: Ein *Call for Contribution* erging an die deutschsprachige

Community digital Edierender, der ca. 60 Einreichungen erbracht hat. Nach einem Redaktionsworkflow, der auch ein *peer-review* der neuen Artikel beinhaltet, ist die Aktualisierung der bestehenden Ressource für Frühjahr 2024 geplant. Die Umsetzung dieses Kleinprojekts wird zeigen, ob ein derartiges Update-System in regelmäßigen Abständen durchgeführt werden kann, sodass der möglicherweise schnellen Überalterung der Inhalte entgegen gewirkt wird.

Bibliographie

Bleier, Roman und Helmut W. Klug. 2020. "Funktion und Umfang des Kommentars in Digitalen Editionen mittelalterlicher Texte: Eine Bestandsaufnahme." In *Annotieren, Kommentieren, Erläutern. Aspekte des Medienwandels*, hg. von Lukas Wolfgang von Elke Richter. Berlin/Boston: de Gruyter, 97–112.

Digital Scholarly Editing. Theory, Practice and Future Perspectives. Ed. by Matthew James Driscoll and Elena Pierazzo. Open Book Publishers: 2016. <https://doi.org/10.11647/OBP.009> (zugegriffen am 14. Dezember 2022).

Dillen, Wout und Vincent Neyt. 2016. "Digital Scholarly Editing within the Boundaries of Copyright Restrictions." *Digital Scholarship in the Humanities* 31.4: 785–96. <https://doi.org/10.1093/llc/fqw011> (zugegriffen 26. Juli 2022).

Dumont, Stefan. 2020. "Kommentieren in digitalen Brief- und Tagebuch-Editionen." In *Annotieren, Kommentieren, Erläutern. Aspekte des Medienwandels*, hg. von Lukas Wolfgang von Elke Richter. Berlin/Boston: de Gruyter 2020, 175–193.

Edlex: Editionslexikon. <https://edlex.de> (zugegriffen 26. Juli 2022).

European Society for Textual Scholarship. Lexicon of Scholarly Editing. <https://lexiconse.uantwerpen.be> (zugegriffen 26. Juli 2022).

Parvum Lexicon Stemmatalogicum. <https://wiki.helsinki.fi/display/stemmatology/Parvum+lexicon+stemmatologicum> (zugegriffen 26. Juli 2022).

Pierazzo, Elena. 2015. *Digital scholarly editing: theories, models and methods*. Farnham. <http://hal.univ-grenoble-alpes.fr/hal-01182162> (zugegriffen 26. Juli 2022).

Pierazzo, Elena. 2019. "What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter." *International Journal of Digital Humanities* 1.2: 209–20. <https://doi.org/10.1007/s42803-019-00019-3> (zugegriffen 26. Juli 2022).

Sahle, Patrick. 2013. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*. 3 Bände. Norderstedt: Books on demand.

Sahle, Patrick. 2014. „Criteria for Reviewing Scholarly Digital Editions, version 1.1 | Institut für Dokumentologie und Editorik“, 2014. <http://www.i-d-e.de/publikationen/weitereschriften/criteria-for-reviewing-scholarly-digital-editions-version-1-1/> (zugegriffen am 14. Dezember 2022).

Sahle, Patrick. 2016. "What Is a Scholarly Digital Edition?" In *Digital Scholarly Editing. Theory, Practice and Fu-*

ture Perspectives, ed. by Matthew Driscoll and Elena Pierazzo. Cambridge: Open Book Publishers, 19–39.

Sahle, Patrick. 2017. "Digitale Edition." In *Digital Humanities. Eine Einführung*. Hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler 2017, 234–254.

Weißbuch Digitale Edition. Hg. von Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". <https://hdl.handle.net/11471/562.50> (zugegriffen am 14. Dezember 2022).

Zotero Bibliografie Kompetenznetzwerk Digitale Edition. <https://www.zotero.org/groups/konde> (zugegriffen am 14. Dezember 2022).

Wie die OPERAS-Projekte PRISM und TRIPLE Open Humanities unterstützen können

Piel, Patrick

piel@maxweberstiftung.de
Max Weber Stiftung, Deutschland

Töpfer, Marlene

toepfer@maxweberstiftung.de
Max Weber Stiftung, Deutschland

Günther, Johanna

Guenther@MaxWeberStiftung.de
Max Weber Stiftung, Deutschland

OPERAS¹, Open Scholarly Communication in the European Research Area for Social Sciences and Humanities, ist eine verteilte europäische Forschungsinfrastruktur für die offene wissenschaftliche Kommunikation in den Sozial- und Geisteswissenschaften und arbeitet an innovativen, auf Nutzer:innen oder Institutionen zugeschnittenen Services. (Maryl et al. 2020) Als solche begleitet OPERAS die Entwicklung dieser Angebote sowohl auf den nationalen Ebenen der sogenannten Core-Mitglieder in Gestalt der National Nodes, die den Kontakt zum eigenen nationalen Bezugsrahmen herstellen, als auch auf der europäischen Ebene über die OPERAS eingegliederten Special Interest Groups (SIGs)², in denen sich die OPERAS-Community über zentrale Aspekte wie Best Practices im Bereich Open Access und Fragen von Multilingualism für OPERAS austauschen kann. Das BMBF-Projekt OPERAS-GER hat unter dem Dach der Max Weber Stiftung die Rolle eines National Node für Deutschland inne und trägt somit als nationales Projekt zur Vernetzung von OPERAS mit der deutschen Wissenschaftslandschaft bei. Durch den Aufbau der nationalen

Kontaktstelle wird ein Beitrag zur nachhaltigen Verknüpfung europäischer und nationaler Forschungsinfrastrukturen geleistet. Als europäische Infrastruktur ist OPERAS als Ganzes seit 2021 Teil der ESFRI Roadmap.³ (Hrušák et al. 2021, 189)

Die Services und der Aufbau von OPERAS richten sich zum einen nach den Bedarfen des akademischen Umfeldes, zum anderen nach den Problemen und Hindernissen bei der Umsetzung einer offenen Wissenskultur in den Sozial- und Geisteswissenschaften. Zwei wesentliche Probleme sind dabei die Fragen nach der Qualitätssicherung von Publikationen, die Open Access publiziert werden und die Problematik der Auffindbarkeit bzw. des Zugriffs auf Open-Access-Publikationen über disziplinäre und sprachliche Grenzen hinweg (Bennett, 2013, 169; Guédon 2019, 30-33). Diese Herausforderungen werden von OPERAS angenommen und durch die Entwicklung der Services GoTriple⁴, im Projekt TRIPLE, Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration und PRISM⁵, Peer Review Information Service for Monographs, beide von OPERAS entwickelt, zugleich als Chance zur Innovation begriffen (Balula and Leão 2021, 96). PRISM, basierend auf DOAB⁶, Directory of Open Access Books, bietet Zugang zu Peer-Reviews von Open-Access-Publikationen. GoTriple, der innovative Discovery-Service von OPERAS, soll die Sichtbarkeit von Forschenden und Open-Access-Literatur verbessern. Außerdem entwickelt OPERAS drei weitere Services, wie den Metrik-Service, der Statistiken über die Nutzung von Publikationen aus dem Open Access, zunächst basierend auf DOAB, bietet⁷. Das Publikationsservice-Portal wiederum soll zukünftig unterschiedliche Publikationsmodelle von OPERAS-Partnern für User übersichtlich aufführen. Schließlich unterstützt COESO⁸, Collaborative Engagement on Societal Issues, die Etablierung von Citizen Science in den Sozial- und Geisteswissenschaften mit der Entwicklung von 10 Pilotprojekten und der VERA-Plattform⁹, Virtual Ecosystem for Research Activation.

Die OPERAS-Services PRISM und GoTriple nehmen das Problem der Qualitätssicherung und der Auffindbarkeit ins Visier. Für die Wissenschaft ist insbesondere das Vertrauen in die Zertifizierung von Forschungen von zentraler Bedeutung, Open-Access-Publikationen haben in diesem Zusammenhang den Nachteil, dass manchen Forschenden die Hürden zur Publikation zu niedrig erscheinen und zugleich die Peer-Review-Verfahren bei Open-Access-Publikationen nicht transparent erscheinen. Zudem werden Open-Access-Formate noch immer nicht als Standard von manchen Herausgebern und Verlagen begriffen.

Zur Sicherstellung von Peer-Review-Standards für Open-Access-Monographien wird daher PRISM entwickelt und über DOAB bereitgestellt. Der Release einer ersten Vollversion ist für Herbst 2022 vorgesehen und wird zukünftig über DOAB und die European Open Science Cloud verfügbar werden. Damit werden Peer-Reviews von Open-Access-Monographien für Nutzer zentral zugänglich gemacht werden, um so die notwendige Transparenz über die Ergebnisse der Evaluation von Publikationen herzustellen, wobei Einsicht in die durchlaufenen Peer-Reviews gewährt werden kann. PRISM bietet eine leichte Integration in Bibliothekskataloge und unter-

stützt Metadatenformate wie MARC21, MARCXML, CSV, RIS und ONIX XML mit OAI-PMH Harvesting. Die Peer-Reviews sind dabei direkt Teil der Metadaten und frei verfügbar.

GoTriple als ein weiterer Service bietet eine ganze Reihe nützlicher Features zur Arbeit mit Open-Access-Publikationen an und stellt sich dabei dem Problem, dass die Auffindbarkeit von Open-Access-Veröffentlichungen durch die institutionellen und sprachlichen Grenzen und durch die Pluralität von Plattformen negativ beeinflusst werden kann. Daher wird die Plattform GoTriple als gebündelter Zugriffspunkt auf Publikationen und Daten zu Forschungen entwickelt. Als Discovery Service verfügt GoTriple über eine gleichzeitige mehrsprachige Suchfunktion in 11 europäischen Sprachen - Englisch, Französisch, Spanisch, Portugiesisch, Deutsch, Italienisch, Polnisch, Griechisch und Kroatisch, sowie Slowenisch und Ukrainisch. Auf diese wird bei einer Suchanfrage zu einem Thema direkt der gesamte unterstützte Raum, über die unterschiedlichen Sprachen hinweg, erfasst. Die Plattform wird von Huma-Num¹⁰ bereitgestellt, eine Vollversion soll 2023 erscheinen. Neben der multilingualen Suche bietet GoTriple weitere integrierte Features an, die den Discovery Service ergänzen: So können auch Daten, die Profile von Forschenden (es gibt hier die Möglichkeit eigene Nutzerprofile anzulegen, sowie über das Trustbuilding System zu vernetzen) und Projekte gesucht werden. Diese Möglichkeiten der Vernetzung mit anderen Forscher:innen und Projekten über GoTriple werden durch das Annotationstool Pundit zur gemeinsamen Ergebnissicherung und ein Crowdfunding-Angebot über die Webseite WeMakelt ergänzt. Zum aktuellen Zeitpunkt speist sich GoTriple aus den folgenden Repositorien und Datenquellen: DOAJ (Directory of Open Access Journals), EKT (Greek National Documentation Centre), OpenAire (Open Access Infrastructure for Research in Europe)¹¹, Isidore und CORDIS (Operational Potential of Ecosystem Research Applications).

Fußnoten

1. OPERAS - Open Scholarly Communication in the European Research Area for Social Sciences and Humanities: Eine der bedeutenden europäischen Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften, denn auf europäischer Ebene sind, auf Geistes- und Sozialwissenschaften spezialisierte Infrastrukturen noch immer eine Seltenheit und zugleich auch eine besondere Herausforderung, wegen der institutionellen und sprachlichen Fragmentierung der vielen eigenen Forschungstraditionen, Institutionen und Kommunikationswege. Die Europäische Kommission definiert Forschungsinfrastrukturen als Einrichtungen, die Dienste und Instrumente anbieten, die der Community der Forschenden für ihre wissenschaftliche Arbeit zur Verfügung gestellt werden können. Diese Einrichtungen können dabei virtuell, alleinstehend oder verteilt aufgebaut sein. Die ESFRI Roadmap (2021) weist unter Anderem OPERAS als eine der wenigen wichtigen Forschungsinfrastrukturen in diesem Bereich aus.
2. OPERAS Mitglieder sind in diesen Special Interest Groups aktiv und wirken dabei an der Umsetzung und

dem Konzept von OPERAS und den entwickelten Services mit. Die Special Interest Groups in OPERAS sind: Advocacy, Best Practices, Common Standards and FAIR Principles, Multilingualism, Open Access Business Models, Open Access Books Network und Tools and Platforms.

3. Die regelmäßig aktualisierte ESFRI Roadmap dient dem Ausbau bzw. der Umsetzung europäischer Forschungsinfrastrukturen und soll so den Forschungsstandort Europa stärken.

4. Die Plattform, <https://www.gotriple.eu/>, des Projektes Triple - Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration, bietet Forschenden und Nutzer:innen die Möglichkeit wissenschaftliche Open-Access-Publikationen zu suchen und zu finden. Ferner bietet es Einrichtungen

5. Peer Review Information Service for Monographs: Dient der Qualitätssicherung von Open-Access-Monographien, indem Informationen zu dem Peer Reviews der Bücher sichtbar und zusammen mit den Metadaten über DOAB, das Directory of Open Access Books, weltweit abrufbar gemacht werden.

6. Das Directory of Open Access Books bietet einen indexierten Zugang zu Open-Access-Büchern.

7. Dieser Metrics Service ist, wie auch PRISM, ein Angebot das auch jenseits von DOAB nutzbar und integrierbar ist. Es richtet sich zugleich an bibliothekarische Einrichtungen, sowie Verlage und implizit auch an Forschende.

8. Collaborative Engagement on Societal Issues ist ein Teilprojekt von OPERAS und konzentriert sich auf die Unterstützung von Citizen Science in den Geistes- und Sozialwissenschaften.

9. Virtual Ecosystem for Research Activation ist eine von COESO entwickelte Plattform, die den Austausch zwischen Forschenden und Bürger:innen für partizipative Wissenschaft in den Geistes- und Sozialwissenschaften ermöglichen soll, so können Nichtwissenschaftler:innen hier beispielsweise nach möglichen akademischen Partnern für die Realisierung von Projekten suchen.

10. Huma-num, bietet Lösungen für Daten in den Geistes- und Sozialwissenschaften an. Es ist ein wichtiger Akteur im Bereich der Infrastrukturen in Europa und an den französischen Anteilen der beiden ERICs DARIAH und CLARIN beteiligt. Huma-num ist wie OPERAS ebenfalls in die Landschaft der europäischen Forschungsinfrastrukturen über die ESFRI Roadmap eingebunden: <https://www.huma-num.fr/>

11. Ein europäisches Forschungsinformationssystem für die Verknüpfung von Forschungsergebnissen (Metadaten aus Repositorien, Journals und Infrastrukturen).

nifer Hansen, Robert Kiley, Anne Kitson, Wim van der Stelt, Kamilla Markram, und Mark Patterson. European Commission, Directorate-General for Research and Innovation. 2019. "Future of Scholarly Publishing and Scholarly Communication. Report of the Expert Group to the European Commission." <https://data.europa.eu/doi/10.2777/836532> (zugegriffen: 03. August 2022).

Hrušák, Jan, Maddalena Donzelli, Marina Carpineti und Petra Dell'Arme. 2021. "Roadmap 2021. Strategy Report on Research Infrastructures." <https://roadmap2021.esfri.eu/media/1295/esfri-roadmap-2021.pdf> (zugegriffen: 03. August 2022).

Maryl, Maciej, Marta Błaszczńska, Agnieszka Szulińska und Paweł Rams. 2020. "The case for an inclusive scholarly communication infrastructure for social sciences and humanities" In *F1000Research* 9:1265. <https://f1000research.com/articles/9-1265/v1#> (zugegriffen: 03. August 2022)

Bibliographie

Balula, Ana und Delfim Leão. 2021. „Multilingualism within Scholarly Communication in SSH – a literature review.“ In *JLIS.it*. 12(2): 88-98. <https://jlis.it/index.php/jlis/article/view/6> (zugegriffen: 03. August 2022).

Bennett, Karen. 2013. „English as a Lingua Franca in Academia Combating Epistemicide through Translator Training.“ In *Interpreter and Translator Trainer* 7: 169–93.

Guédon, Jean-Claude, Michael Jubb, Bianca Kramer, Mikael Laakso, Birgit Schmidt, Elena Šimukovič, Jen-

Index der Autorinnen und Autoren

Achmann, Michael	286
Aehnlich, Barbara	370
Akhlaq, Sara	67
Albani, Benedetta	151
Althage, Melanie	383
Altmann, Friederike	45
Andresen, Melanie	395, 422
Anokhina, Alexandra	151
Aoun, Sandy	364
Atzenhofer-Baumgartner, Florian	364
Baillet, Anne	33
Balck, Sandra	378
Bamberg, Claudia	244, 329
Baresch, Ariadne	81
Barth, Florian	45, 160
Bateman, John	138
Baum, Constanze	22
Beck, Samuel	52
Bell, Peter	242
Berger, Claudia	67
Berndt, Axel	146
Bernhart, Toni	58
Biemann, Chris	227, 252, 388
Biermann, Johannes	45
Bigalke, Jan	360
Bläß, Sandra	88, 355
Blessing, André	312, 375
Blumtritt, Jonathan	360, 404
Bock, Sina	118
Borek, Luise	19, 401
Borst, Janos	91
Brandes, Vanessa	408
Bürgermeister, Martina	234
Börner, Ingo	406
Brunner, Annelen	397
Bruns, Oleksandra	181
Burch, Thomas	203, 330
Burckhardt, Daniel	83
Burghardt, Manuel	13, 91, 138, 304
Busch, Anna	iv, 408
Busch, Hannah	401
Calvo Tello, José	160
Christ, Andreas	381
Cotgrove, Louis	414
Cugliana, Elisa	102, 404
Czmiel, Alexander	42
Dahnke, Michael	22
Dang, Sarah-Mai	77, 81
Decker, Franziska	364
Deicke, Aline	74, 203
Dias, Renata	357
Diebel, Richard	381
Diecke, Josephine	81
Dieckmann, Lisa	390
Diehr, Franziska	178
Dietrich, Elisabeth	326
Dinger, Patrick	22, 165
Dipper, Stefanie	399

Dönicke, Tillmann	45
Doppler, Carina	16
Dürfeld, Michael	357
Dörk, Uwe	329
Du, Keli	309
Duan, Tinghui	77
Dubova, Alona	170
Dumont, Stefan	121
Durdağı, A. Nursen	75
Dutschke, René	426
Ebel, Carla	51
Egger, Nils	257
Eggersglüß, Christoph	69
Ehlers, Lena	395
Eichhorst, Dana	283
Eigner, Johanna	234
Eiser, Isabel	252, 388
Elwert, Frederik	67, 399
Ewerth, Ralph	81
Fadeeva, Yuliya	22
Fath, Laura	203
Feichtinger, Moritz	74
Feidicker, Charlotte	34
Fischeidl, Kathrin	366
Fischer, Andreas	98
Fischer, Frank	114, 177, 406
Fischer, Tim	252, 388
Flüh, Marie	88, 355
Fliegl, Heike	181
Frank, Ingo	378
Frank, Laura	283
Franken, Lina	257
Franzini, Greta	174
Gaertner, Markus	422
Galka, Selina	217, 428
Gebhard, Henning	399
Geißel, Pia	346
Geiger, Jonathan D.	13
Geißler, Nils	404
Gengnagel, Tessa	72, 77, 360, 404
Gerber, Anja	34, 77, 95, 109
Gerstorfer, Dominik	30, 189, 410
Ghosh, Sharanya	348
Giovannini, Luca	406
Gius, Evelyn	189, 227, 410
Glas, Julia	286
Glawion, Anastasia	134
Günther, Johanna	430
Grallert, Till	77, 83
Grießer, Martina	419
Große, Peggy	384
Grote, Brigitte	207
Gärtner, Markus	58
Grundig de Vazquez, Katja	142
Guhr, Svenja	410
Hadassah Wendl, Katharina	283
Hadjakos, Aristotelis	146
Hagen, Thora	321
Hagener, Malte	69
Haider, Thomas Nikolaus	358
Hall, Mark	75, 417
Hanzer, Helene	419
Harvey, Francis	75
Hatzel, Hans Ole	227

Heßbrüggen-Walter, Stefan	231	Krüger, Katharina	373
Hegel, Philipp	404	Kristen, Maximilian	273
Hein, Pascal	312, 375	Kröncke, Merten	156
Helling, Patrick	62	Kudella, Christoph	404
Henning, Pia	368	Kuhn, Jonas	58
Henny-Krahmer, Ulrike	42	Kunze, Petra	370
Hensen, Kilian	404	Kurz, Stephan	369
Henzel, Katrin	381	Kurzmeier, Michael	356
Herbst, Yannik	296	Kushnarenko, Volodymyr	375
Hess, Jan	312, 375	Kusnick, Jakob	16
High-Steskal, Nicole	67	Lambertz, Michael	245
Hilger, Agnes	386	Lameris, Bregt	81
Hiltmann, Torsten	185	Lamers, Teresa	419
Hinzmann, Maria	54	Lamminger, Florian	364
Hodel, Tobias	98	Lang, Christian	415
Holz, Alex	424	Lang, Sabine	212
Homburg, Timo	27	Lang, Sarah	19, 72, 77
Hopp, Meike	265	Lasch, Alexander	426
Horstmann, Jan	13, 22, 165, 301	Lück, Christian	301, 421
Howanitz, Gernot	80, 393	Lecroq, Axelle	412
Hörschemeyer, Jörg	83, 404	Leinen, Peter	194
Häußler, Julian	391	Lemaire, Marina	37, 75
Hunziker, Manuel	25	Lemke, Karoline	404
Illmer, Viktor Jonathan	177	Leonhardt, Susann	373
Jacke, Janina	317, 422	Liebl, Bernhard	138
Janka, Anna	98	Liem, Johannes	16, 51
Jannidis, Fotis	156, 198	List, Ferdinand	357
Jettka, Daniel	42	Lordick, Harald	404
Jha, Vandana	399	Lu, Christopher	406
Jänicke, Steffan	16	Ludwig, Bernd	286
Jung, Kerstin	62, 312, 375	Luger, Daniel	364
Junginger, Pauline	341	Lyding, Verena	174
Kababgi, Daniel	62	Marie, Flüh	125
Kailus, Angela	372	Marquart, Aron	170
Kaltseis, Magdalena	393	Marschner, Michèle	357
Kampkaspar, Dario	72	Martini, Annett	283
Karcher, Stefan	72	Mayer, Desiree	372
Katrin, Hein	397	Mayer, Manuela	369
Keck, Jana	78, 84	Mayr, Eva	16, 51
Kepper, Johannes	249	Möbus, Dennis	257
Ketschik, Nora	58, 401, 422	Meier-Vieracker, Simon	426
Kipke, Marta	129	Meister, Malte	30
Kirchweiger, Franz	419	Milling, Carsten	406
Klaes, Jan Sebastian	373	Müller-Laackman, Jonas	72
Klammt, Anne	28, 105	Müller, Michael	278
Klee, Anne	54	Müller-Tamm, Jutta	177
Kleymann, Rabea	13	Münzmay, Andreas	249
Kloser, Peter	419	Moeller, Katrin	37
Klug, Helmut	377	Murphy, Órla	356
Klug, Helmut W.	217, 428	Nantke, Julia	88, 355
König, Mareike	40, 83	Neitzke, Thorben	45
Koch, Gertraud	252, 388	Neuber, Frederike	404, 412
Koch, Julia	58	Neudecker, Clemens	67
Koch, Steffen	52	Neuefeind, Claes	404
Kolbe, Ines	326	Nicolaou, Angelos	364
Konle, Leonard	156	Niekler, Andreas	91, 305
Konstanciak, Johanna	54	Nieländer, Maret	373
Kou-Herrema, Tianyi	336	Niemann, Klara	105
Kovacs, Tamas	364	Noichl, Maximilian	261
Kraft, Tobias	121	Normann, Immanuel	301
Krause, Celia	69	Nowakowski, Matthias	146
Krautter, Benjamin	291	Nyhan, Julianne	78
Kreffft, Annett	142	O'Sullivan, James	356
Kröger, Bärbel	48	Offert, Fabian	28

Panzer, Lukas	261	Seltmann, Melanie Elisabeth-H.	22, 269, 404
Park, Yohan	151	Sievers, Martin	404
Passecker, Markus	16	Skorinkin, Daniil	406
Pechlivanos, Miltos	238	Sluyter-Gäthje, Henny	406, 410
Pektor, Katharina	234	Soethaert, Bart	237
Peter, Leinen	125	Sommer, Kai	374
Petersen, Britta	381	Sonnberger, Jakob	369
Petersen Frey, Fynn	252, 339, 388	Sporleder, Caroline	45
Petkov, Radoslav	245, 330	Söring, Sibylle	37
Pfeifer, Ulrike	426	Stahn, Lena-Luise	350
Pichler, Axel	423	Stanicka-Brzezicka, Ksenia	69
Pidd, Michael	356	Steffes, Moritz	54
Piel, Patrick	430	Stein, Christian	357
Pielström, Steffen	62, 118	Stein, Regine	402
Pittroff, Sarah	95	Steindl, Christoph	234
Pollin, Christopher	377, 419	Steiner, Christian	377
Pons, Jessie	67	Steiner, Elisabeth	419
Posthumus, Etienne	181	Steller, Jonatan	95
Probst, Nora	19, 77	Steyer, Timo	22, 390
Probst, Stefan	52	Stiemer, Haimo	227
Radisch, Erik	362	Ströbel, Phillip	98
Rahman, Zead	357	Störiko, Johanna	48
Rastinger, Nina C.	343	Strobel, Jochen	203
Rüdiger, Jan Oliver	413	Strobl, Maren	207
Regeler, Lukas Nils	177	Strutz, Sabrina	377
Rensinghoff, Berenike	25	Suárez Cronauer, Elena	203
Reul, Christian	88, 296, 355	Sutor, Nadine	337
Richter, Matthias	221	Tögel, Philipp	399
Rißler-Pipka, Nanette	160	Thelen, Julius	359
Ripoll, Elodie	410	Thiery, Florian	28
Roeder, Torsten	34, 296	Thoden, Klaus	142
Rosenthal, Maximilian	221	Thomas, Christian	121
Röttgermann, Julia	54	Thran, Niklas	357
Ruth, Nicolas	305	Tietz, Tabea	181
Sack, Harald	182	Tolksdorf, Julia	402
Schauffler, Nadja	58	Tonne, Danah	283, 399
Schöch, Christof	54, 410	Towara, Nadine	374
Schelbert, Georg	278	Töpfer, Marlene	430
Schellhammer, Stefan	165	Trautmann, Tatjana	329
Schenk, Nicolas	312	Trilcke, Peer	iv, 407, 408, 410
Schennach, Stephanie	358	Trippel, Thorsten	194
Scherbaum, Stefan	426	Troglauser, Patrick	165
Scherer, Thomas	81	Tseng, Chiao-I	138
Scheurer, Patricia	98	Tu, Ngoc Duyen Tanja	415
Schirmer, Miriam	345	Ullrich, Rebecca	283
Schlesinger, Claus-Michael	312	Varachkina, Hanna	45
Schlögl, Matthias	51	Vasold, Gunter	419
Schmid, Florian	257	Vater, Christian	402
Schmitz, Jascha	13	Vetter, Angila	381
Schmunk, Stefan	75	Vogeler, Georg	364
Schneider, Florian	252, 388	Volk, Martin	98
Schneider, Philipp	185	Vollmer, Ricarda	273
Schneider, Roman	416	Volodina, Anna	416
Schneider, Stefanie	273	Vukovic, Teodora	81
Scholger, Martina	419	Wachter, Christian	74
Schrade, Torsten	72, 181, 402	Wagner, Sarah	170
Schulz, Daniela	404	Walkowski, Niels-Oliver	390
Schulz, Julian	37, 83	Walsh, David	417
Schumacher, Mareike	30, 125	Wehrheim, Lino	90
Schwab, Michel	114	Weigelt, Lucie	426
Schwandt, Silke	13	Weimer, Lukas	194, 403
Scius, Anna	98	Weis, Joëlle	390
Seemann, Sophia	426	Weitin, Thomas	134
Seifert, Sabine	121	Wesche, Jörg	359

Wessels, Bridgette	356
Wettlaufer, Jörg	37, 48
Weyand, Sandra	203
Widmer, Jonas	98
Wierzoch, Jan	121
Wierzock, Alexander	329
Windhager, Florian	16, 51
Winko, Simone	156
Wintjes, Jorit	118
Witt, Andreas	194
Wolf, Beat	98
Wolfer, Sascha	413
Wolff, Christian	286
Wuttke, Ulrike	72, 390
Yilmaz, Yasir	369
Zimmermann, Ronny	408
Çakir, Dilan Canan	424
von Berenberg-Gossler, Luise	177
von Pippich, Waltraud	25
von dem Bussche, Ruth	265

DHd2023

OPEN HUMANITIES

OPEN CULTURE



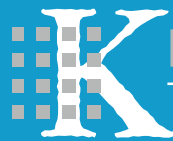
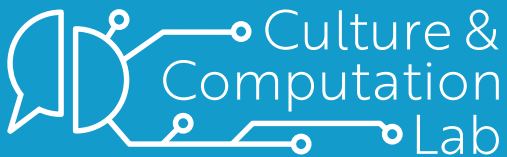
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY



UNIVERSITÉ DU
LUXEMBOURG



digital humanities im
deutschsprachigen raum



Kompetenzzentrum

Trier Center for Digital Humanities

