UNIVERSITÄT PASSAU

Book of Abstracts DHd2024

26.02. – 01.03.2024

digital humanities im deutschsprachigen raum

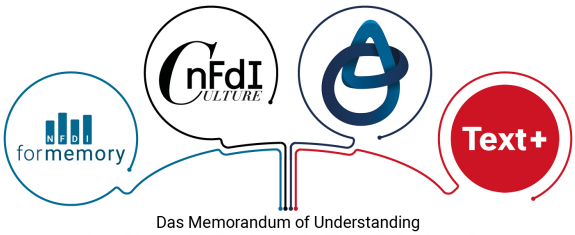
UNIVERSITÄT PASSAU

DHd2024

Quo Vadis DH

Partner und Sponsoren

Transkribus



Das Memorandum of Understanding
der geistes- und kulturwissenschaftlichen NFDI-Konsortien

Gefördert durch

DFG Deutsche Forschungsgemeinschaft

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

DHd2024: Quo Vadis DH?

Universität Passau

26.02. – 01.03.2024

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:
Joëlle Weis, Estelle Bunout, Thomas Haider
Redaktion und Korrektur der Auszeichnungen:
Elisabeth Huber, Markus Gerstmeier, Tobias Perschl, Anke Debbeler, Nicole Majka, Patrick Helling, Thomas Haider
Data Steward des Verbandes:
Patrick Helling

Konvertierung TEI nach PDF: Thomas Haider
https://github.com/tnhaider/DHd2024-BoA
Historie der Autorinnen und Autoren sowie Versionen der Konversionsskripte:
Nina Seemann (2020)
https://github.com/NinaSeemann/DHd2020-BoA
Attila Klett (2019)
https://github.com/texttechnologylab/DHd2019BoA
Claes 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

Lokales Konferenzteam:
Thomas Haider, Heidi Riederer, Jessica Nieder, Johann-Mattis List, Malte Rehbein
Konferenz-Logo, Umschlagdesign, Gestaltung der Webseite:
Thomas Haider

Online verfügbar: https://doi.org/10.5281/zenodo.10686565

Passau, 2024

Vorwort

Vor einem Jahrzehnt hat sich die DH-Community aus dem deutschsprachigen Raum erstmals zu ihrer jährlichen Konferenz in Passau getroffen. Seitdem wird die Veranstaltung jedes Jahr an wechselnden Orten in Deutschland, im deutschsprachigen Ausland sowie einmalig online abgehalten.

Anlässlich des Jubiläums thematisiert die diesjährige Tagung, erneut in Passau stattfindend unter dem Motto "Quo Vadis DH?", die Zukunft der Digital Humanities. Sie wirft zugleich einen Blick zurück auf die positiven wie negativen Erfahrungen der Vergangenheit, mit dem Ziel, Entwicklungspotenziale zu diskutieren und mögliche Wege für die nächsten zehn Jahre aufzuzeigen.

Auch dieses Jahr stand das Programmkomitee vor der großen Herausforderung, eine Auswahl unter 190 Einreichungen für Vorträge, Poster, Panels, Workshops und das Doctoral Consortium zu treffen.
Wir danken den Autor:innen für die facettenreichen Einblicke in ihre Forschung und den engagierten 170 Kolleg:innen, die sich bereit erklärt haben, bei der Qualitätssicherung für die DHd2024 mitzuwirken. Gemeinsam haben wir beeindruckende 518 Gutachten erstellt. Die Gutachtenden sind ein wesentlicher Bestandteil der DHd-Konferenzen, ohne die die Gestaltung der Tagung unmöglich wäre. Wir sind überzeugt, dass ein vielfältiges Programm von hoher Qualität entstanden ist, das die Transdisziplinarität und Offenheit unserer Community widerspiegelt.

Zur wissenschaftlichen Qualitätssicherung wurde der zweistufige Begutachtungsprozess beibehalten. Wie bereits in den beiden Vorjahren wurde das sogenannte Open-Peer-Review-Verfahren angewandt, bei dem die Namen der Gutachter:innen den Autor:innen offengelegt werden. Für die Zuordnung der Gutachter:innen wurden fachliche Interessen und Expertise berücksichtigt. Auch dieses Jahr gab es eine Erwiderungsphase, die von den Autor:innen breit genutzt wurde. Insgesamt wurden 210 Antworten auf Gutachten verfasst, was bei 71 Beiträgen zu einer Anpassung der vergebenen Punktezahl seitens der Gutachter:innen führte. Obwohl das Verfahren mit zusätzlichem Aufwand verbunden ist, sehen wir darin eine große Chance, das Peer-Review für die DHd so fair und transparent wie möglich zu gestalten - zum Vorteil aller Beteiligten. Darüber hinaus kann die Community bereits Monate vor dem Treffen in einen Austausch treten, was die wissenschaftliche Diskussion auch während der Konferenz bereichert. So wird das Review-Verfahren zu einem Beitrag zu offenen (Geistes-)Wissenschaften, denen wir uns als DH-Community verpflichtet fühlen, wie auch das Konferenzmotto "Open Humanities, Open Science" von 2023 verdeutlichte.

Der Dialog zwischen Einreichenden und Gutachter:innen trägt nicht nur zur Steigerung der inhaltlichen Qualität der Beiträge und zur Transparenz des Begutachtungsprozesses bei, sondern unterstützt auch das Programmkomitee bei der endgültigen Entscheidungsfindung. Um die Leistung der Gutachter:innen zu würdigen, hat das Programmkomitee - nachdem die ursprünglich verantwortliche Task Force "Optimizing Peer Review" nicht mehr aktiv ist - dieses Jahr zudem die Ehre, den Review Award zu verleihen. Wir freuen uns sehr, dass wir mit Lisa Eggert eine junge und engagierte Preisträgerin gefunden haben.

Wir danken allen Mitgliedern des Programmkomitees für ihr Engagement: Anne Baillot, Noah Bubenhofer, Anna Busch, Alexander Czmiel, Lisa Dieckmann, Evelyn Gius, Katrin Glinka, Andreas Henrich, Andreas Münzmay, Patrick Sahle, Martina Scholger und Silke Schwandt. Unser Dank gilt auch allen lokalen Mitorganisator:innen in Passau, die die Arbeit des Programmkomitees entscheidend unterstützt haben. Patrick Helling und seinem Team danken wir für die Unterstützung beim Erstellen des Book of Abstracts.

Wir freuen uns auf eine bereichernde und zukunftsweisende Konferenz!

Trier, Belval und Passau, im Februar 2024
Joëlle Weis, Estelle Bunout und Thomas Haider
für das Programmkomitee und das lokale Organisationsteam der DHd2024

Reviewer:innen der DHd2024

Review Award 2024

Eggert, Lisa

Nominierte für den Review Award 2024

Guhr, Svenja
Hinzmann, Maria
Jannidis, Fotis
Lang, Sarah
Neudecker, Clemens
Trippel, Thorsten
Veit, Joachim

Reviewer:innen

Achmann, Michael
Akkermann, Miriam
Andresen, Melanie
Andrews, Tara
Arnold, Eckhart
Baillot, Anne
Bamberg, Claudia
Barabucci, Gioele
Barzen, Johanna
Bell, Peter
Bernhart, Toni
Blaschitz, Edith
Bläsi, Christoph
Blumtritt, Jonathan
Börner, Ingo
Brunner, Annelen
Bubenhofer, Noah
Bunout, Estelle
Burch, Thomas
Burckhardt, Daniel
Burr, Elisabeth
Busch, Anna
Busch, Hannah
Capelle, Irmlind
Casties, Robert
Cremer, Fabian
Czmiel, Alexander
Dang, Sarah-Mai
Deicke, Aline
Dieckmann, Lisa
Dinger, Patrick
Draxler, Christoph
Du, Keli
Duan, Tinghui
Düring, Marten
Eggert, Lisa
Eide, Øyvind
Elwert, Frederik
Fechner, Martin
Flüh, Marie
Frank, Markus
Freyberg, Linda
Fritze, Christiane
Geiger, Jonathan
Gengnagel, Tessa
Gerber, Anja
Gerstenberg, Annette
Gerstmeier, Markus
Gerstorfer, Dominik
Giovannini, Luca
Gius, Evelyn
Glinka, Katrin
Gradl, Tobias
Grallert, Till
Grote, Brigitte
Grund, Vera
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
Hinzmann, Maria
Hodel, Tobias
Hohmann, Georg
Homburg, Timo
Horstmann, Jan
Howanitz, Gernot
Illmayer, Klaus
Jannidis, Fotis
Jäschke, Robert

Jung, Kerstin
Jünger, Jakob
Kampkaspar, Dario
Keck, Jana
Klammt, Anne
Klemstein, Franziska
Kleymann, Rabea
Klinke, Harald
Koch, Walter
Kocher, Ursula
König, Sandra
Konle, Leonard
Krautter, Benjamin
Kremer, Dominik
Kröger, Bärbel
Kuczera, Andreas
Kurz, Stephan
Lang, Sarah
Langner, Martin
Leinen, Peter
Lemaire, Marina
List, Johann-Mattis
Lucke, Alexa
Lüschow, Andreas
Mandl, Thomas
Mayr, Eva
Meier-Vieracker, Simon
Meyer, Holger J
Molitor, Paul
Moulin, Claudine
Münzmay, Andreas
Nantke, Julia
Neudecker, Clemens
Neuefeind, Claes
Niekler, Andreas
Noichl, Maximilian
Nunn, Christopher
Oberbichler, Sarah
Offert, Fabian
Pfeffer, Magnus
Pichler, Axel
Pielström, Steffen
Probst, Nora
Proisl, Thomas
Puppe, Frank
Rehm, Georg
Reiners-Selbach, Stefan
Reiter, Nils
Rißler-Pipka, Nanette
Roller, Ramona
Röttgermann, Julia
Rüdiger, Jan Oliver
Sack, Harald
Sahle, Patrick
Schaßan, Torsten
Schmidt, Thomas
Schmunk, Stefan
Schneider, Stefanie
Schöch, Christof
Scholger, Martina
Scholger, Walter
Schommer, Christoph
Schumacher, Mareike
Schwandt, Silke
Seltmann, Melanie Elisabeth
Söring, Sibylle
Stadler, Peter
Staecker, Thomas
Stede, Manfred
Steyer, Timo
Ströbel, Phillip Benjamin
Teich, Maximilian C.
Thomas, Christian
Tietz, Tabea
Trippel, Thorsten
Tu, Ngoc Duyen Tanja
Veit, Joachim
Venken, Machteld
Viehhauser, Gabriel
Vogeler, Georg
Voges, Ramon
Wagner, Andreas
Weis, Joëlle
Wettlaufer, Jörg
Wieneke, Lars
Windhager, Florian
Wolff, Christian
Wübbena, Thorsten
Wuttke, Ulrike
Zaagsma, Gerben
Zehe, Albin

Inhaltsverzeichnis

Keynotes

Workshops

Panels

Vorträge

Doctoral Consortium

Posterpräsentationen

Anhang

Keynotes

Comparative interdisciplinary research in History and Sociology

van Leeuwen, Marco
Utrecht University

I will talk about the role of theory, methods, data and collaboration in History - and by extension the Human Sciences - and in the Social Sciences - in particular Sociology. These rather abstract notions are illustrated with examples from my past ERC-project Towards Open Societies, using big register datasets in several European countries to study social mobility and social homogamy over the past two centuries. I will also talk about the limitations of that project and how other more qualitative, textual sources may help to overcome these.

Fiction - the key to bringing digital humanities and corpus linguistics closer together

Mahlberg, Michaela
University of Birmingham

In today's digital age, many areas of the humanities have seen a digital turn and there is plenty of innovation in methods and tools. Still, there are also parallel developments across different fields, and even the danger of reinventing the wheel. The study of fiction is a case in point. In digital humanities, literary texts have received much attention. With a specific focus on language, literary texts have also been studied in corpus linguistics. While there is productive overlap between digital humanities and corpus linguistics research, there is still huge potential for more collaboration and more exchange of ideas. At times, it seems divisions between literary studies and linguistics have been carried over to separate approaches in digital humanities and corpus linguistics. I want to suggest that the way we approach fiction is the key to more collaborative thinking across different fields and disciplines. I focus on the example of corpus linguistics, but some of the points I will make do extend to other areas of humanities. Crucially, fiction is never just fiction. Fiction and non-fiction can be described along a continuum. We use fundamentally the same language to create fictional worlds and to talk about what we perceive as the `real world'. In this talk, I will look at how corpus approaches, as linguistic approaches, aim to account for properties of literary texts. On the other hand, I will consider how the study of literature and literary history can be linked up with questions in corpus-assisted discourse studies. Ultimately, collaborative approaches will not only lead to better methods and tools, but also to a better understanding of the fuzzy boundaries between fiction and the real world.

Workshops

BEACONe unde venis et quo vadis?

Al-Eryani, Susanne
al-eryani@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-1741-1645

Gerstner, Eva-Maria
gerstner@maxweberstiftung.de
Max Weber Stiftung, Bonn, Deutschland
ORCID: 0000-0001-9920-4238

Hegel, Philipp
philipp.hegel@tu-darmstadt.de
Technische Universität Darmstadt, Deutschland
ORCID: 0000-0001-6867-1511

Lordick, Harald
lor@steinheim-institut.org
Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte, Essen, Deutschland
ORCID: 0000-0002-5070-4263

Schnöpf, Markus
schnoepf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-2529-8248

Schulz, Daniela
schulz@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0003-3167-5089

Sikora, Uwe
sikora@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

BEACON

Der Workshop zielt im Sinne der FAIR-Prinzipien und im Hinblick auf konkrete Standardisierungsbestrebungen auf die Vernetzung geistes- und kulturwissenschaftlicher Daten durch Erstellung sogenannter BEACON-Dateien[1] auf Basis der Gemeinsamen Normdatei (GND). Maschinenlesbare BEACON-Dateien stellen über normdatenbasierte Web-URLs oder eine Schnittstelle Normdaten-IDs für Ressourcen zur Verfügung und ermöglichen die gegenseitige Verlinkung von Webseiten, die ihrerseits Inhalte mittels Normdaten verknüpft haben. Das können digitale Editionen, Enzyklopädien, Bibliographien, aber auch andere Ressourcen wie wissenschaftliche Blogs sein.

Neben der Motivation, die auch für die generelle Annotation mit Normdaten gilt, nämlich die eindeutige Identifizierung von Entitäten und inhaltliche Anreicherung der eigenen Daten, geht es bei der Verwendung von BEACON-Dateien insbesondere um die gegenseitige Verlinkung und erhöhte Sichtbarkeit der eigenen Website, und das, wie die Erläuterungen zur Struktur und Anwendung dieser Dateien zeigen werden, auf niederschwelligem Weg.

In der Praxis werden dazu meist Identifier der Gemeinsamen Normdatei verwendet. Eine typische Anwendung ist eine direkt in das jeweilige digitale Angebot integrierte „See-Also“-Funktionalität, die zu passenden ‚Treffern‘ anderer Online-Angebote führt.

BEACON-Dateien werden zudem auch von der Deutschen Nationalbibliothek genutzt, um GND-bezogene Daten Dritter in den Datendienst *Entity Facts*[2] zu übernehmen. WikiData verwendet und exportiert ebenfalls BEACON-Dateien.

Das BEACON-Findbuch[3][4], STI Linked Data Service[5] und CorrespSearch[6] bieten auf GND-BEACON basierende übergreifende (domänenspezifische) See-Also-Funktionen an. Außerdem werden BEACON-Dateien von biographischen und bibliographischen Portalen wie beispielsweise dem *Bayerischen Musiker-Lexikon Online*, der *Deutschen Biographie* oder der *Landesbibliographie Baden-Württemberg* verwendet.[7]

Beispiele und praktische Übungen

In dem Workshop werden zunächst zwei Beispiele vorgeführt, an denen das grundlegende Vorgehen erläutert wird. Im ersten Beispiel werden die strukturierten Metadaten von Online-Publikationen und auch Blogbeiträgen verwendet und aus diesen eine BEACON-Datei erzeugt. Beim zweiten Beispiel wird auf die Volltextdaten der Sitzungsprotokolle der Preußischen Akademie (bis 1806) als XML-Dateien ohne Normdaten zurückgegriffen. Anhand dieser Beispiele werden verschiedene Schritte prototypisch vorgestellt und praktisch durchgespielt, die notwendig sind, um eine digitale Ressource mit einer online zugänglichen BEACON-Datei zu versehen und so ihre Vernetzung zu ermöglichen:

1. In den Beispieldaten werden die zu berücksichtigenden Entitäten festgelegt.
2. Für die definierten Entitäten werden die GND-IDs unter Verwendung von OpenRefine[8] und in Kombination mit der Lobid-API[9] recherchiert und eingetragen.
3. Aus den Einträgen wird eine BEACON-Datei (Plain Text, UTF-8) erstellt.

4. Die Publikation der BEACON-Datei wird anhand eines konkreten fachspezifischen Dienstes für Judaica und durch den Eintrag der neuen BEACON-Datei in die ‚Registry' für BEACON-Dateien (Wikipedia:BEACON) vorgeführt.

Der Schwerpunkt des Workshops liegt auf der praktischen Anwendung und Hands-On-Erprobung des BEACON-Verfahrens. Teilnehmer:innen können dazu eigene Daten mitbringen. Hierzu wird es im Vorfeld einen Austausch geben. Sie können für die praktische Übung aber, je nach Interesse, auch eine der vorbereiteten Online-Publikationen oder eines der genannten Sitzungsprotokolle nutzen. Die Ergebnisse, die Erfolge und Schwierigkeiten werden anschließend gemeinsam diskutiert.

Ausblicke

In einem ersten Ausblick wird gezeigt, was hinter den im Workshop genutzten See-Also-Services (u.a. STI Linked Data Service) steckt. In einem zweiten Ausblick sollen die Möglichkeiten von ‚Künstlicher Intelligenz', namentlich ChatGPT, in diesem Bereich erkundet werden. Lassen sich derartige Techniken sinnvoll und effizient einsetzen, um Entitäten in Metadaten und Volltexten zu identifizieren und mit Normdatensätzen zu verknüpfen? In einem dritten Ausblick werden mit der GND-Agentur und entityXML neuere Entwicklungen aufgezeigt, die ein agileres Zusammenspiel von Projekten und Normdaten ermöglichen.

Weder prinzipiell noch technisch sind BEACON-Dateien auf die GND als Normdatei und auf Personen als Entitäten festgelegt. Sie sind auch nicht auf bestimmte Arten von digitalen Ressourcen eingeschränkt. Diskutiert werden kann daher auch, welche Rolle im Sinne einer Internationalisierung des Formats und der Services VIAF und Wikidata heute schon und in zukünftigen Anwendungen spielen können. Gefragt werden soll darüber hinaus, welche technischen und organisatorischen Hindernisse überwunden werden müssen, damit ein breiterer Einsatz des BEACON-Verfahrens möglich wird.

Format, Methodik und Ziele des Workshops

Der Workshop ist als Präsenzveranstaltung geplant. Das didaktische Konzept sieht eine einführende Vermittlung von Wissen sowie Praxisphasen in Gruppenarbeit und Plenumsdiskussionen vor. Es sind vier Stunden angesetzt, um möglichst viel Zeit für die eigene Arbeit und den Austausch zu gewährleisten. Der Workshop ist folgendermaßen strukturiert:

- Mehrere kurze Impulsvorträge zu unterschiedlichen Aspekten münden in einem ersten Austausch zum Thema im Plenum (vgl. 1., Dauer: 45 Minuten).
- Im Anschluss daran beginnt der praktische Teil des Workshops, in dem die Teilnehmenden nach einer kurzen Einführung in mehreren Gruppen mit unterschiedlichen thematischen Schwerpunkten BEACON-Dateien erstellen und idealtypische Workflows bis hin zur Publikation einer BEACON-Datei erproben. Auf Wunsch geschieht dies an eigenem Material. Der Block schließt mit einem Austausch zu den gemachten Erfahrungen. (vgl. 2., Dauer: 120 Minuten, exkl. Pausen).
- Abschließend erfolgt ein Ausblick, der die bestehende BEACON-Landschaft mit aktuellen Entwicklungen kontextualisiert und nach Möglichkeiten, aber auch nach Grenzen der weiteren Entwicklung fragt (vgl. 3., Dauer: 45 bis 60 Minuten).

Am Ende des Workshops sollten die Teilnehmer:innen wissen, was GND-BEACON-Dateien sind und welche Möglichkeiten, aber auch welche Grenzen der Vernetzung durch ihren Einsatz bestehen. Sie haben verschiedene Anwendungsfelder kennengelernt sowie Hilfsmittel, die der Erstellung von BEACON-Dateien dienen, und haben diese auch exemplarisch ausprobiert. Ebenso wissen sie, welche BEACON-bezogenen Services existieren und wo deren Schwerpunkte liegen. Sie haben darüber hinaus einen Einblick in die Funktionsweise dieser Dienste erhalten. Anhand von unterschiedlichen Beispieldaten haben sie erfahren, wie mögliche Workflows aussehen und eventuell selbsttätig eine BEACON-Datei erstellt, publiziert und dadurch ihre Daten mit anderen Daten GND-basiert vernetzt.

Eine Publikation der im Workshop gesammelten Erfahrungen und Ergebnisse im Rahmen von Blogbeiträgen ist geplant. Ebenso werden diese Ergebnisse auch in Handreichungen des NFDI-Konsortiums Text+ zum Thema eingehen.

Zielpublikum und Teilnehmer*innenzahl

Der Workshop richtet sich an Forschende aller Disziplinen der (Digital) Humanities mit Interesse an Vernetzung von Daten und Linked Data, insbesondere auch an Nachwuchswissenschaftler*innen. Vorwissen wird nicht vorausgesetzt. Um einen guten Betreuungsschlüssel zu gewährleisten, können bis zu 25 Personen am Workshop teilnehmen.

Organisationsteam bzw. Beitragende

- Susanne Al-Eryani ist Fachreferentin für Orientalistik und Kulturanthropologie/Europäische Ethnologie an der Staats- und Universitätsbibliothek Göttingen und wirkt in der Arbeitsgruppe *Metadaten und Datenkonversion* der SUB Göttingen am Aufbau einer Text+ GND-Agentur mit.

- Eva-Maria Gerstner ist Referentin für bibliothekarisch-informationswissenschaftliche Services mit Schwerpunkt Forschungsdatenmanagement.
- Philipp Hegel ist Mitarbeiter im Sonderforschungsbereich 980: *Episteme in Bewegung* und im Konsortium Text+ der Nationalen Forschungsdateninfrastruktur.
- Harald Lordick forscht zur deutsch-jüdischen Geschichte, Digital Humanities und geisteswissenschaftlichen Infrastrukturen.
- Markus Schnöpf koordiniert das Forschungsdatenmanagement der Berlin-Brandenburgischen Akademie der Wissenschaften, darunter zahlreiche Editionsprojekte aus verschiedenen Epochen.
- Daniela Schulz betreut an der Herzog August Bibliothek die *Hybridedition der deutschsprachigen Werke des Martin Opitz* sowie das NFDI-Konsortium Text+.
- Uwe Sikora arbeitet im Bereich *Metadaten und Datenkonversion* an der Niedersächsischen Staats- und Universitätsbibliothek in Göttingen.

Die Einreichenden verbindet ihr Interesse an der Verknüpfung digitaler geistes- und kulturwissenschaftlicher Ressourcen. Sie sind im NFDI-Konsortium Text+ engagiert.

Benötigte Ausstattung

Benötigt wird ein größerer Seminarraum mit WLAN und Beamer, der sowohl Diskussion im Plenum als auch das verteilte Arbeiten in mehreren Gruppen ermöglicht. Für die praktische Arbeit an den Daten genügt seitens der Teilnehmenden das Mitbringen eines Laptops.

Fußnoten

1. https://de.wikipedia.org/wiki/Wikipedia:BEACON (zugegriffen: 18. Juli 2023).
2. https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/Entity-Facts/entityFacts_node.html (zugegriffen 18. Juli 2023).
3. http://beacon.findbuch.de/ (zugegriffen: 18. Juli 2023).
4. https://github.com/hbeyer/beacon (zugegriffen: 18. Juli 2023).
5. http://www.steinheim-institut.de/see-also/query.html (zugegriffen: 18. Juli 2023).
6. https://correspsearch.net/de/start.html (zugegriffen: 18. Juli 2023).
7. https://www.bmlo.lmu.de, https://data.deutsche-biographie.de/beta/beacon-open und https://www.statistik-bw.de/LABI (zugegriffen jeweils: 18. Juli 2023).
8. https://openrefine.org/ (zugegriffen: 18. Juli 2023).
9. https://lobid.org/gnd/api (zugegriffen: 18. Juli 2023).

Bibliographie

Lordick, Harald. 2022. „Digitale Editionen und die vernetzte GND-BEACON-Landschaft.“ In Text + Blog, 10.11.2022. https://textplus.hypotheses.org/752 (zugegriffen am 19. Juli.2023).

Lordick, Harald und Mache, Beata. 2018. „Annotationen anhand der Gemeinsamen Normdatei aus einer anwendungsorientierten Perspektive historischer Forschung“. In Kritik der digitalen Vernunft, hg. von Georg Vogeler. https://doi.org/10.5281/zenodo.1188229 (zugegriffen: 19. Juli 2023).

Sikora, Uwe. 2022. entityXML: Handbuch. https://entities.pages.gwdg.de/entityxml (zugegriffen: 19. Juli 2023).

Stadler, Peter. 2012. „Normdateien in der Edition“. In Editio 26: 174-183.

Voß, Jakob und Mathias Schindler. 2017. BEACON Link Dump Format. https://gbv.github.io/beaconspec/beacon.html (zugegriffen: 19. Juli 2023).

Das richtige Tool für die Volltextdigitalisierung

Baierer, Konstantin
konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin, Deutschland
ORCID: 0000-0003-2397-242X

Hinrichsen, Lena
hinrichsen@hab.de
Herzog August Bibliothek, Deutschland
ORCID: 0000-0002-9286-2390

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-4615-4753

Reul, Christian
christian.reul@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland
ORCID: 0000-0002-1776-1469

Sautter, Lilja
sautter@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Mustafa, Mehmed
mehmed.mustafa@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung
Göttingen, Deutschland

Will, Larissa
larissa.will@uni-mannheim.de
Universität Mannheim, Deutschland
ORCID: 0009-0004-6220-8939

Sowohl Einrichtungen als auch Forschende, die Volltexte generieren möchten, stehen vor einer Vielzahl an Tools, die entweder Open Source oder kostenpflichtig sind. Jedes der Tools bringt spezifische Vor- und Nachteile mit sich. Eine Auswahl des Tools kann von verschiedenen Faktoren abhängig sein, beispielsweise Art und Menge des Ausgangsmaterials, verfügbare Hardware, verfügbare Softwarekenntnisse sowie die gewünschte Qualität. Auch aufgrund der vielen verschiedenen Anforderungen werden im OCR-D-Projekt[1] unterschiedliche Lösungen entwickelt. „Dazu wurde ein Koordinationsprojekt gebildet, das in der ersten Projektphase Entwicklungsbedarfe identifizierte. Diese wurden in der zweiten Projektphase von insgesamt acht Modulprojekten bearbeitet. In der derzeitigen dritten Projektphase steht die konzeptionelle Vorbereitung für die automatische Generierung von Volltexten für die Verzeichnisse der im deutschen Sprachraum erschienenen Drucke des 16., 17. und 18. Jahrhunderts im Fokus. Außerdem arbeiten vier Implementierungsprojekte daran, OCR-D in bestehende Anwendungen und Infrastrukturen zu integrieren, während drei Modulprojekte OCR-D-Werkzeuge weiter optimieren.“ (https://ocr-d.de/de/about)

OCR4all und eScriptorium

OCR4all[2] ist eine benutzerfreundliche Anwendung für die OCR mit einer grafischen Benutzeroberfläche, die unter anderem ein Training der Layout- und Texterkennung niedrigschwellig ermöglicht. In einem der OCR-D-Implementierungsprojekte wird das „Open-Source-Werkzeug OCR4all so erweitert und angepasst werden, dass Bibliotheken und Archive bei ihrer Massendigitalisierung die im Rahmen des OCR-D-Projekts erarbeiteten Lösungen niederschwellig, flexibel und eigenständig einsetzen können. Eine zusätzliche visuelle Erklärungskomponente soll darüber hinaus Unterstützung bei der Erstellung und Konfiguration optimaler OCR-Workflows bieten.

Als Use Case fungiert die Forschungsbibliothek des GEI Braunschweig mit ihren digitalisierten Schulbüchern des 17. und 18. Jahrhunderts. Um zunehmende Komplexitäten der so entstehenden OCR-Lösung nutzerorientiert aufzufangen, wird die bestehende grafische Benutzerschnittstelle in enger Kooperation und unter Anleitung des HCI Lehrstuhl der Universität Würzburg angepasst und weiterentwickelt.“ (ebenda)

Auch eScriptorium,[3] entwickelt von der Université PSL, ist als Open-Source-Lösung geeignet für DH-Anwender*innen, um manuell und automatisch Transkriptionen zu erstellen. Während der Fokus bei OCR4all auf einer großen Anzahl an verschiedenen OCR-Tools liegt, die hier in einer Anwendung integriert sind, hat eScriptorium Vorteile in den Bereichen des Datenaustauschs und der Ergonomie. eScriptorium und OCR4all haben beide grafische Benutzeroberflächen, die es Forschenden ermöglichen, Volltexte zu erstellen und dabei einen vergleichsweise niedrigschwelligen Einstieg zu haben. Die Tools zu kennen und auszuprobieren, versetzt Forschende in die Lage, für ihre eigenen Projekte geeignete Methoden auszuwählen. eScriptorium und OCR4all sind eine Auswahl unter vielen verschiedenen OCR-Tools, die verfügbar sind. Forschende können nach ihrem Einstieg in die OCR ggf. Weitere, spezialisertere Tools ausprobieren, in vielen Fällen werden sie allerdings mit eScriptorium und/oder OCR4all ihre Bedarfe für die Forschung decken können und haben damit eine gute Grundlage für weitere Arbeiten. Daher können diese beiden Anwendungen im Workshop ausprobiert werden. Dazu können gern eigene Daten mitgebracht werden.

Individuelle OCR-Workflows für Forschende und bestandshaltende Einrichtungen

Je nach Tool ist es ggf. notwendig, geeignete Workflows und Modelle auszuwählen. Schritte einer OCR (Optical Character Recognition) oder einer HTR (Handwritten Text Recognition) können unter anderem umfassen:

- Binarisierung: Das Umwandeln von Bildern in Farbe oder Graustufen in Schwarz-Weiß-Bilder
- Cropping: Ausschneiden des Bildes auf den Bereich, der erkannt werden soll (ohne gegenüberliegende Seiten oder störende Elemente)
- Denoising: Entfernen von störenden Artefakten
- Deskewing: Geraderücken eines schiefen Scans
- Dewarping: Entfernen von Verzerrungen
- Layoutanalyse auf Regionen-, Zeilen-, Wort- und Glyphenebene.
- Texterkennung
- Nachkorrektur

Dabei können diese Schritte auf verschiedene Arten kombiniert werden und einzelne Schritte mehrmals ausgeführt werden. Für jeden Prozessierungsschritt gibt es bei OCR-D bzw. OCR4all in der Regel mehr als einen Prozessor, der zur Verfügung steht. Zudem arbeiten manche Prozessoren modellbasiert, sodass ein solches Modell ausgewählt (und ggf. weiter trainiert) werden muss. Weitere mögliche Parameter bei der Erstellung eines Workflows kommen ergänzend hinzu. Um hier die richtige Wahl für Workflows und Modelle bei OCR4all, eScriptorium oder anderen Tools zu

treffen, sind grundlegende Kenntnisse der OCR und Deep Learning notwendig, die im Workshop vermittelt werden.

Nutzungsszenarien und Ressourcen von OCR-D

Innerhalb des Projekts OCR-D mit dem Fokus auf Masssenvolltextdigitalisierung werden Ressourcen bereitgestellt, die auch im Rahmen individueller OCR-Arbeiten von Forschenden nützlich sein können.

Aktuell können in OCR-D die Workflows über die Kommandozeile ausgeführt werden. Mit der neuen Version von OCR4all, das die OCR-D-Prozessoren integriert hat, ist es außerdem möglich, die Konfigurationen über eine grafische Benutzeroberfläche vorzunehmen. Dies macht individuelle Workflows zugänglicher. OCR4all als freie und auf Nutzerfreundlichkeit konzentrierte Software bietet alle notwendigen Workflowschritte an und integriert das Tool LAREX (Layout Analysis and Region Extraction) für Layouterkennung und Korrektur von Zwischenergebnissen der Layout- sowie Texterkennung.

Für bestimmte Materialien kann ein einfacher Workflow mit wenigen Schritten ausreichend sein, um eine gute Qualität zu erzeugen, während Materialien mit komplexen Layouts aufwändigere Workflows notwendig machen. Für eine Vorauswahl stellen wir Standard-Workflows sowie das im OCR-D-Projekt entwickelte Benchmarking-Tool QUIVER bereit.[4] Mit dieser Entwicklung erstellt OCR-D Werte für Durchsatz und Qualität bestimmter Workflows auf verschiedenen Materialien.

Werden Modelle (nach-)trainiert und dafür Ground Truth erstellt, erreicht das Vorhaben eine weitere Komplexitätsstufe. Bei der Transkription von Ground Truth helfen die in OCR-D entwickelten und gepflegten Ground-Truth-Guidelines.[5] „Mit den OCR-D-Ground-Truth-Guidelines wurden Richtlinien geschaffen, die eine Format-Dokumentation des vorhandenen OCR-D-Ground-Truth darstellt und als Handlungsanweisung für die Ground-Truth-Erstellung genutzt werden kann. Mit dieser Normierung kann der Ground-Truth technisch validiert werden. Darüber hinaus können vorhandene Transkriptionen auf Grundlage dieses Regelwerkes überprüft und gegebenenfalls in Ground-Truth-Daten umgewandelt werden. Das Datenformat des OCR-D-Ground-Truth ist PAGE-XML. Dieses Format wurde initial durch das PRImA Research Lab an der Universität Salford Greater Manchester entwickelt und innerhalb des EU-Projektes IMPACT grundlegend erweitert. Zurzeit wird es vom PRImA Research Lab betreut. Um eine Weiterentwicklung und Pflege dieses Formates zu gewährleisten, wurde auf Initiative von OCR-D ein PAGE-XML-Board geschaffen.“ (https://ocr-d.de/de/gt-guidelines/trans/) Zusätzlich finden regelmäßige Onlinemeetings (GT-Call)[6] statt, um Fragen zu erörtern.

Workshop

Im ganztägigen Workshop erlangen die Teilnehmenden erforderliche Kenntnisse, um Tools und Workflows für die Volltexterschließung unter der Vielzahl von Angeboten auszuwählen. Dabei legen wir einen besonderen Fokus auf die genannten Open-Source-Produkte wie die OCR-D, OCR4all und eScriptorium.

Geplante Inhalte des Workshops sind:

- Einführung in Deep Learning, OCR, Layoutanalyse, Evaluation, OCR-D und passende Workflows für bestimmte Vorlagen
- Praktisches Arbeiten mit OCR4all und LAREX
- Praktisches Arbeiten mit eScriptorium
- Möglichkeiten zum Hochskalieren mit OCR-D Processing Server/Workern
- Unterstützung beim individuellen Einrichten der Werkzeuge

Bei der Anwendung von OCR4all mit LAREX und eScriptorium ist der Workshop interaktiv gestaltet und die Teilnehmenden können die Tools selbst ausprobieren. Sie erhalten grundlegende Kenntnisse für die Anwendung. Dabei kann anhand selbst mitgebrachter Vorlagen bereits festgestellt werden, wo Potenziale, aber auch Grenzen der OCR liegen können.

Nach Abschluss des Workshops kennen die Teilnehmenden frei verfügbare OCR-Tools und Anlaufstellen für Ressourcen, die sie bei der Einrichtung eines OCR-Workflows benötigen.

Vortragende

Konstantin Baierer (SBB Berlin), Lena Hinrichsen (HAB Wolfenbüttel) und Matthias Boenig (BBAW) arbeiten im OCR-D-Koordinierungsprojekt.

Christian Reul leitet die Digitalisierungseinheit des Zentrums für Philologie und Digitalität „Kallimachos“ (ZPD) and er Universität Würzburg. In OCR-D ist er unter anderem mitverantwortlich für ein Implementierungspaket für OCR4all.

Lilja Sautter ist an der Niedersächsischen Staats- und Universitätsbibliothek Göttingen in der Einheit Software- und Service-Entwicklung tätig. In OCR-D ist sie beteiligt an den Projekten OPERANDI sowie OLA-HD.

Mehmed Mustafa ist Entwickler in der Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen. In OCR-D arbeitet er im Koordinierungsprojekt sowie den Projekten OPERANDI und OLA-HD.

Larissa Will ist Referentin für Forschungsdatenmanagement und Digitalisierung (Digital Humanities) an der Universität Mannheim.

Zielpublikum, technische Ausstattung

Teilnehmende sollten ein spezifisches Interesse daran haben, selbst Volltexte zu erstellen. Dies können Forschende sein, die Volltexte für ihre Projekte benötigen, sowie Mitarbeitende von Einrichtungen, die Volltexte und/oder OCR-Services in der Zukunft anbieten möchten. Um die Tools auszuprobieren, wird ein Laptop benötigt, möglichst mit Windows oder Linux.

Fußnoten

1. https://ocr-d.de/
2. https://www.ocr4all.org/
3. https://gitlab.com/scripta/escriptorium
4. https://ocr-d.de/quiver-frontend/#/workflows?view=list
5. https://ocr-d.de/en/gt-guidelines/trans/
6. https://ocr-d.de/en/community

Bibliographie

eScriptorium. https://gitlab.com/scripta/escriptorium [19.07.2023].

Website OCR-D. https://ocr-d.de [19.07.23].

Website OCR.4all. https://www.ocr4all.org/ [19.07.2023].

Der Weg zum grünen Forschungsdaten-managementplan

Gerber, Anja
anja.gerber@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-2576-1511

Rosendahl, Lisa
rosendahl@beethovens-werkstatt.de
Beethovens Werkstatt/Beethoven-Haus Bonn, Deutschland
ORCID: 0000-0002-4826-4553

Forschungsdatenmanagement berührt inzwischen alle Bereiche in der Wissenschaft, da während jedes Forschungsprozesses – insbesondere auch in den Digital Humanities – Daten jeder Art anfallen, die als Forschungsdaten bezeichnet werden. Mittelgeber fordern, dass bei der Antragstellung bereits ein Datenmanagementplan (DMP) vorgelegt werden muss, um den Umgang mit den Daten während, aber auch nach, der Projektlaufzeit genau zu erfassen. Aspekte des Klimaschutzes werden hier noch nicht berücksichtigt, jedoch schreitet die Klimakrise zeitgleich zur steigenden Anzahl an digitalen Daten und Methoden voran. Dies berührt Aspekte der Digital Humanities, da hier neben steigenden Mengen digitaler Daten auch Methoden, technische Prozesse und Speicherroutinen zu berücksichtigen sind. Daher sollen in diesem Workshop gemeinsam Fragen zu einem und der Prozess eines grünen Forschungsdatenmanagements diskutiert sowie ein Musterdatenmanagementplan entwickelt werden.

Theoretische Grundlagen Forschungsdatenmanagement

Während jedes Forschungsprozesses, insbesondere auch im Bereich der Digital Humanities, fallen Daten jeder Art an, z. B. Messdaten, Texte, Bilddateien, audiovisuelle Daten oder 3D-Modelle. Daten, die während des Forschungsprozesses entstehen oder das Ergebnis dessen sind, werden auch als Forschungsdaten bezeichnet (Kindling/Schirmbacher 2013). Das Forschungsdatenmanagement (FDM) begleitet alle Aktivitäten in Zusammenhang mit diesen Forschungsdaten, hierunter sind Prozesse ihrer Aufbereitung, Speicherung, Archivierung und Veröffentlichung zu verstehen. Es umfasst den gesamten Forschungsdatenzyklus – den Prozess von der Planung über die Erhebung und Verarbeitung bis hin zur Archivierung, Nachnutzung oder auch Löschung der Daten (Forschungsdaten.info 2023, Kindling et al. 2013).

Drittmittelgeber und Forschungsförderer geben in ihren Richtlinien vor, dass ein angemessener Umgang mit den in Projekten erstellten Forschungsdaten essenziell für qualitätsorientierte und anschlussfähige Forschung ist. Bereits während der Antragstellung ist daher der Umgang mit den Daten und Objekten, die diesen zugrunde liegen, mitzubedenken. Das betrifft die Planung, Dokumentation und Beschreibung, aber auch die Nachnutzung dieser. Es wird ebenfalls angeraten, fachspezifische einschlägige Empfehlungen zu Standards, Methoden und Infrastrukturen zu berücksichtigen (DFG 2023).

Neben einer genauen Beschreibung der Daten, die in den Forschungsprojekten entstehen, ist eine Dokumentation dieser sowie der Maßnahmen für eine hohe Datenqualität erforderlich. Hier werden die Datentypen und der zu erwartende Umfang erfasst, ebenso Methoden, wie diese weiterverarbeitet werden können. Methoden zur nachvollziehbaren Beschreibung der Daten sind ebenfalls zu erfassen, z. B. ob bereits vorhandene Standards und Ontologien nachgenutzt werden, sowie die für die Datenerfassung notwendigen Methoden und Software. Die Aspekte der Speicherung und technischen Sicherung der Daten sind ein weiterer Aspekt, der schon im Vorfeld bedacht werden muss, hierunter fallen z. B. die Art der Speicherung der Daten und Zugriffsrechte. Zusätzlich sind rechtliche Rahmenbedingungen zu

bedenken, wie zu erwartende Einschränkungen bei einer späteren Veröffentlichung bzw. Onlinestellung der Daten, was den Aspekt der Nutzungs- und Urheberrechtsfragen und Lizenzierungen der Daten betrifft. Ein weiterer wichtiger Punkt des Forschungsdatenmanagements ist der Austausch und die dauerhafte Zugänglichkeit von Daten. Kriterien für die Auswahl von nachnutzbaren Forschungsdaten müssen transparent dokumentiert werden, ebenso die Archivierung in einer geeigneten Infrastruktur und mögliche Sperrfristen. Rollen und Verantwortlichkeiten im Umgang mit Forschungsdaten sind, zusammen mit zeitlichen und materiellen Ressourcen, zu benennen, auch ist die weitere Datenpflege nach Projektende zu benennen (DFG 2021).

Für die übersichtliche und nachvollziehbare Dokumentation der Forschungsdaten wird ein Datenmanagementplan (DMP) genutzt. Dieser ist ein lebendiges Dokument, das während der Projektlaufzeit weiter angepasst und aktualisiert werden muss und beschreibt, wie mit den Daten während und nach der Laufzeit verfahren wird. Er hat mehrere Ebenen, wie „Überblick“ (Ziel, Angaben zu Projektverantwortlichen und -laufzeit), „Datenbestand“, „Datengenese“, „Workflow“, „Dateningest“, „Konsolidierung“, „Verbreitung“, „Rollen des Datenmanagements“ und „Ressourcen“ (Forschungsdaten.info 2023).

Forschungsdaten und Klimawandel

Im Kontext des Klimawandels gewinnt auch das Forschungsdatenmanagement eine zunehmende Bedeutung, da es einen erheblichen Einfluss auf die Umweltauswirkungen von wissenschaftlichen Aktivitäten hat. Das Thema der Nachhaltigkeit wird im Rahmen der FAIR-Prinzipien zwar häufig thematisiert, allerdings geht es hier meist nicht um ökologische Aspekte, sondern um die langfristige Sicherung der Daten (Wilkinson et al. 2016).

Server und Rechenzentren sind wesentliche Bestandteile des Forschungsdatenmanagements, jedoch gehen mit ihrer Nutzung erhebliche CO2-Emissionen einher. Die kontinuierliche Energieversorgung, die Kühlung der Hardware und die Speicherung großer Datenmengen erfordern beträchtliche Mengen an Energie, die oft aus fossilen Brennstoffen gewonnen wird. Dieser Energieverbrauch trägt zur Freisetzung von Treibhausgasen und somit zum Klimawandel bei. Hier gilt es, nachhaltige Lösungen für die Infrastruktur der Datenspeicherung zu entwickeln.

Bereits durch den Einsatz energieeffizienter Serverhardware und Optimierung der Kühlungsmechanismen können erhebliche Energieeinsparungen erzielt werden. Die Verwendung von Servern mit geringerem Stromverbrauch und die Implementierung effektiver Kühlungsstrategien, beispielsweise durch die Nutzung erneuerbarer Energiequellen oder Abwärmenutzung, tragen zur Reduzierung der CO2-Emissionen bei. Sollen Cloud Computing-Dienste und virtuelle Infrastrukturen genutzt werden, so ist es bei der Auswahl von Cloud-Anbietern wichtig, deren Einsatz erneuerbarer Energiequellen und deren Umweltbilanz zu berücksichtigen (Shao et al. 2022).

Eine weitere Möglichkeit, den ökologischen Einfluss des Forschungsdatenmanagements zu reduzieren, besteht darin, eine genaue Selektion der Daten vorzunehmen, die gespeichert werden sollen. Durch eine Priorisierung der zu speichernden Daten können Einsparungen erzielt werden. Zusätzlich ist es in vielen Fällen möglich, Dateigrößen zu komprimieren, indem z. B. bereits im Vorhinein entschieden wird, Bilddateien nicht in der größtmöglichen, sondern in einer dem Zweck angepassten Auflösung bereitzustellen. Ein klimagerecht orientiertes FDM sollte sicherstellen, dass die Ressourcen (z. B. Hardware, Rechenzeit, Datenspeicherung) in angemessenem Verhältnis zu den erzielten Ergebnissen (z. B. Outputs, erwarteten Erkenntnissen) stehen (vgl. für das Beispiel des Natural Language Processing: Bender et al. 2021). Hierbei kann es sinnvoll sein, Ansätze des Minimal Computing zu nutzen, bei denen der Energieverbrauch bei der Datenverarbeitung minimiert wird (vgl. Pereira et al. 2017; Abbing 2021).

Zusätzlich zu diesen Maßnahmen können auch andere Aktivitäten zur Förderung eines grünen Forschungsdatenmanagements beitragen. Eine Verlängerung der Lebensdauer von Hardware kann die Nachhaltigkeit verbessern, indem die Notwendigkeit für häufige Neuanschaffungen reduziert wird. Des Weiteren können Tools zur Berechnung des Energieverbrauchs bei der Datenverarbeitung eingesetzt werden, um die Auswirkungen auf die Umwelt zu quantifizieren und geeignete Maßnahmen zur Energieeinsparung zu identifizieren.

Ein umweltfreundliches Forschungsdatenmanagement ist von großer Bedeutung, um den ökologischen Fußabdruck der Digital Humanities-Forschung zu verringern (Juckes et al. 2022). Indem ökologische Aspekte in den Fokus gerückt und nachhaltige Praktiken implementiert werden, können Wissenschaftler*innen dazu beitragen, den Klimawandel zu bekämpfen und eine nachhaltige Zukunft zu gestalten (vgl. für das Beispiel der Langzeitspeicherung in Archiven: Pendergrass et al. 2019). Von der „Information, Measurement and Practice Action Group“ der Digital Humanities Climate Coalition gibt es bereits eine Leitlinie für die Erstellung klimabewusster Forschungsdatenmanagementpläne (Baker et al. 2022). Dieser kann auf die Kontexte an deutschen Universitäten und je nach Forschungsfeldern und -kontexten angepasst werden. Der Workshop soll dafür erste Grundlagen bilden.

Zielstellung des Workshops

Der ganztägige Workshop soll dazu dienen, gemeinsam in Kleingruppen Ideen für ein grünes Forschungsdatenmanagement zu entwickeln und zu diskutieren. Hier können und sollten auch Szenarien aus dem eigenen Umfeld eingebracht werden, um die Realisierbarkeit der diskutierten Pläne sicherzustellen. Der Workshop dient nicht der Präsentation fertiger Lösungen sondern der Diskussion von Fragestellungen aus dem Bereich des klimabewussten FDM. Anhand der Problemstellungen werden dann gemeinsam Lösungsansätze skizziert sowie Best Prac-

tice-Beispiele und ein Musterdatenmanagementplan entwickelt. Eine Vorlage wird im Rahmen des Workshops zur Verfügung gestellt, Vorschläge für ein grünes Template können ebenfalls seitens der Teilnehmenden eingebracht und diskutiert werden.
Der grüne DMP soll gemeinsam mit einem Aufsatz über die Ergebnisse des Workshops im Nachgang zur DHd publiziert und der Community zur Verfügung gestellt werden. Zudem gibt es in der AG GreeningDH konkrete Überlegungen, darauf aufbauende Workshops zu grünem Forschungsdatenmanagement zu organisieren.

Der Workshop ist für eine Teilnehmendenanzahl von maximal 20 Personen konzipiert, Vorkenntnisse sind nicht erforderlich. Um sicherzustellen, dass die Inhalte des Workshops an die Teilnehmer angepasst sind, werden ihre jeweiligen Vorkenntnisse im Vorfeld durch eine Umfrage abgefragt. Um die Szenarien zu diskutieren und gemeinsam an einem DMP zu arbeiten, wird ein Computerpool mit Beamer und Whiteboard oder Flipchart inklusive Moderationskoffer benötigt.

Ablauf des Workshops

Für den Workshop ist folgender Ablauf geplant:

1. Begrüßung und Einführung (30 Minuten): Vorstellung der AG und der Teilnehmenden sowie der Ziele des Workshops
2. Theorie (30 Minuten): Überblick über Forschungsdatenmanagement und FAIR-Prinzipien, ausgerichtet auf Vorkenntnisse der Teilnehmenden und Bereitstellung von Materialien

[Pause]

3. Themensammlung (30 Minuten): Sammlung von Themen oder Fragen rund um klimabewusstes Forschungsdatenmanagement
4. Sessionplanung (10 Minuten): Themenwahl und Planung der Sessions sowie Zusammenstellung von Kleingruppen
5. Session 1, Diskussion und Ideenaustausch in Kleingruppen (30 Minuten): Diskussion auf Grundlage bereitgestellter Materialien und Fallstudien aus der eigenen Arbeitspraxis der Teilnehmenden

[Pause]

6. Session 2, Entwicklung von Handlungsansätzen in Kleingruppen (60 Minuten): Erarbeitung konkreter Maßnahmen, Best Practices oder Richtlinien

[Pause]

7. Kurzpräsentationen (60 Minuten): Präsentation und Diskussion der entwickelten Handlungsansätze

[Pause]

8.1 Erstellung eines Forschungsdatenmanagementplans (60 Minuten): Gemeinsame Erstellung eines Forschungsdatenmanagementplans anhand eines Templates

[Pause]

8.2 Fortsetzung: Erstellung eines Forschungsdatenmanagementplans (60 Minuten)

[Pause]

9. Zusammenfassung und Ausblick (60 Minuten): Zusammenfassung der wichtigsten Ergebnisse und Erkenntnisse des Workshops mit Ausblick auf mögliche Folgeaktivitäten und weiteren Austausch

Der Zeitplan kann flexibel auf die Bedingungen vor Ort und die Bedürfnisse der Teilnehmenden angepasst werden.

Organisationsteam

Anja Gerber (ORCID 0000-0003-2576-1511, anja.gerber@klassik-stiftung.de), ist seit 1.8.2023 an der Klassik Stiftung Weimar für die Task Area 6 „Qualification, Harmonisation and Integration“ der NFDI4Objects u. a. für die Erstellung von Objektbiografien und die Entwicklung einer N4O CoreOntology sowie Fragen des Forschungsdatenmaangements zuständig. Sie hat fundierte Kenntnisse in Datenmodellierung und Metadatenstandards sowie Erfahrungen im Umgang mit heterogenen Forschungsdaten. In ihrer bisherigen Tätigkeit an der Berlin-Brandenburgischen Akademie der Wissenschaften war sie beim Corpus Vitrearum Medii Aevi unter anderem konzeptionell an der Entwicklung des „CVMA Digitaler Ressourcen Managers“ beteiligt, einer Erfassungsumgebung für Bilddaten, die nach einer projekteigenen Spezifikation mit Metadaten annotiert und als XMP in die Header der Bilddateien geschrieben werden. Im Sommersemester 2023 lehrte sie an der Fachhochschule Potsdam „Metadatenvertiefung“ im Bachelorstudiengang „Bibliothekswissenschaften“. Sie hat Informationswissenschaften und Digitales Datenmanagement studiert.

Lisa Rosendahl (ORCID 0000-0002-4826-4553 , rosendahl@beethovens-werkstatt.de), Wissenschaftliche Mitarbeiterin am Beethoven-Haus Bonn im Projekt *Beethovens Werkstatt*, studierte Musikwissenschaft, Geschichte und Digital Humanities in Düsseldorf und Münster. Ab Dezember 2022 arbeitete sie als wissenschaftliche Hilfskraft in der Beethoven-Gesamtausgabe sowie ab April 2021 am Musikwissenschaftlichen Seminar Detmold/Paderborn im DFG/AHRC-Projekt *Beethoven in the House: Digitale Studien zu Bearbeitungen für Hausmusik* (Kooperation mit der Universität Oxford, RISM Digital und dem Beethoven-Haus Bonn).

Beide Autorinnen engagieren sich in der AG „Greening DH“, die 2021 mit dem Ziel gegründet wurde, 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, grundlegende Veränderungen, die sich daraus für das Fach ergeben, epistemologisch zu begleiten. Einige ihrer wichtigsten Arbeitsergebnisse sind die aktuelle Arbeit an „GreeningDH Guidelines“ für den DHd-Verband sowie das bereits on-

line publizierte „The Digital Humanities Climate Coalition Toolkit“ zusammen mit der Digital Humanities Climate Coalition (DHCC). Neben technischen Prozessen, Fragen an Speichermanagement oder Infrastrukturen betrifft das auch den Bereich des Datenmanagements.

Bibliographie

Abbing, Roel Roscam. 2021. „This Is a Solar-Powered Website, Which Means It Sometimes Goes Offline: A Design Inquiry into Degrowth and ICT”. In: LIMITS’21: Workshop on Computing within Limits (14./15. Juni 2021).

Baker, James, Christopher Ohge, Lisa Otty, Jo Lindsay Walton. 2022. „A Researcher Guide to Writing a Climate Justice Oriented Data Management Plan (v0.6)”. Zenodo. DOI: https://doi.org/10.5281/zenodo.6451499.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell. 2021. „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”. In: *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Hrsg. von der Association for Computing Machinery* , New York 2021, S. 610–623. DOI: https://doi.org/10.1145/3442188.3445922 .

Deutsche Forschungsgemeinschaft (DFG). 2021. „Checkliste zum Umgang mit Forschungsdaten.“ https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf (zugegriffen: 15.7.2023).

Deutsche Forschungsgemeinschaft (DFG). 2023. „Umgang mit Forschungsdaten.“ https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/index.html (zugegriffen: 9.7.2023).

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: 9.7.2023).

Forschungsdaten.info. 2023. „Forschungsdaten und Forschungsdatenmanagement. Der Datenlebenszyklus.“ (https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/ (zugegriffen: 9.7.2023)

Forschungsdaten.info. 2023. „Forschungsdaten und Forschungsdatenmanagement. Der Datenmanagementplan“. https://forschungsdaten.info/themen/informieren-und-planen/datenmanagementplan/ (zugegriffen 15.7.2023).

Juckes, Martin, Charlotte Pascoe, Lucy Woodward u. a. 2022.„ *Interim Report: Complexity, Challenges and Opportunities for Carbon Neutral Digital Research”*. Zenodo. DOI: https://doi.org/10.5281/ZENODO.7016952 .

Kindling, Maxi and Schirmbacher, Peter. 2013. „ *Die digitale Forschungswelt* als Gegenstand der Forschung. Lehrstuhl Informationsmanagement“. In: Information - Wissenschaft & Praxis 64, no. 2-3 (2013): 127-136. https://doi.org/10.1515/iwp-2013-0017 .

Kindling, Maxi, Schirmbacher, Peter und Simukovic, Elena. 2013. „Forschungsdatenmanagement an Hochschulen: das Beispiel der Humboldt-Universität zu Berlin.“ LIBREAS. Library Ideas, 23 (2013). Online verfügbar unter: http://libreas.eu/ausgabe23/07kindling/, DOI: 10.18452/9041.

Pendergrass, Keith, Walker Sampson, Tim Walsh, Laura Alagna. 2019 „ *Toward Environmentally Sustainable Digital Preservation”*. In: The American Archivist 82/1, S. 165–206. DOI: https://doi.org/10.17723/0360-9081-82.1.165.

Pereira, Rui, Marco Couto, Francisco Ribeiro. u. a. 2017. „Energy Efficiency across Programming Languages: How Do Energy, Time, and Memory Relate?”. In: Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering, Vancouver 2017, S. 256–267. DOI: https://doi.org/10.1145/3136014.3136031

Shao, Xiaotong, Zhongbin Zhang, Ping Song, Yanzhen Feng, Xiaolin Wang. 2022. „ *A Review of Energy Efficiency Evaluation Metrics for Data Centers”*. In: Energy and Buildings 271. DOI: https://doi.org/10.1016/j.enbuild.2022.112308.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg u. a. 2016. „ *The FAIR Guiding Principles for scientific data management and stewardship”*. In: Scientific Data 3, 160018 (2016). DOI: https://doi.org/10.1038/sdata.2016.18.

Edierst Du noch oder trainierst Du schon? Forschungsdaten als Grundlage von Trainingsdaten für die automatische Texterkennung

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-4615-4753

Baierer, Konstantin
konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Deutschland
ORCID: 0000-0003-2397-242X

Hinrichsen, Lena
hinrichsen@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0002-9286-2390

Würzner, Kay-Michael
kay-michael.wuerzner@slub-dresden.de
Sächsische Landesbibliothek — Staats- und Universitätsbibliothek Dresden (SLUB), Deutschland
ORCID: 0000-0002-9039-4124

Reul, Christian
christian.reul@uni-wuerzburg.de
Zentrum für Philologie und Digitalität (ZPD) der Universität Würzburg, Deutschland
ORCID: 0000-0002-1776-1469

Einführung

Wichtigste Grundlage der textorientierten Forschung in den Digital Humanities ist eine ausreichende Verfügbarkeit von hochwertigem maschinenlesbarem Text. Diese Anforderung kann bei grundständig digitalen Texten häufig einfacher erfüllt werden als bei historischen Texten, wo zunächst die Transformation vom gedruckten oder geschriebenen Wort auf Papier in eine geeignete digitale Repräsentation zu realisieren ist.

Mit der Anwendung von Verfahren des maschinellen Lernens in der automatischen Texterkennung ist in den letzten zehn Jahren ein enormer Fortschritt vollzogen worden. Dies betrifft vor allem die Zeichenerkennung und deren Genauigkeit. Hierbei kommen Methoden zum Einsatz, die dem Paradigma *Lernen aus Beispielen* folgen. Die dazu nötigen Trainingsdaten werden als Ground Truth (GT) bezeichnet.

> „Der Ursprung des Wortes Ground Truth ist das deutsche Wort Grundwahrheit. Im OCR-Zusammenhang bedeutet das, dass alles auf der gedruckten Seite in gleicher Art und Weise nach definierten Regeln unter anderem in digitaler Form wiedergegeben wird.“[1]

Aber GT dient nicht nur dem Training der Zeichenerkennung (sowohl dem Training eines neuen Modells „from scratch“, als auch dem „Finetuning“ eines bestehenden Modells auf einen spezifischen Anwendungsfall hin), sondern wird auch zur Datenvalidierung, -evaluation und -referenzierung eingesetzt. Neben der Zeichenerkennung können aber weitere Teilprozesse der automatischen Texterkennung vom Einsatz maschinellen Lernens profitieren. Dies gilt insbesondere für die Erkennung und Auszeichnung der Seitenstruktur bzw. des Seitenlayouts. Diese unterschiedlichen Anwendungen setzen differenzierte GT-Typen voraus. Allgemein kann zwischen Struktur-GT und Text-GT unterschieden werden.

Die Erstellung von GT erfolgt zu einem Großteil manuell, was einen hohen zeitlichen und finanziellen Aufwand erfordert. Um brauchbaren GT zu erstellen, sind abgestimmte Konventionen und Richtlinien notwendig. Aus diesem Grund entwickelt, pflegt, vermittelt und diskutiert das Projekt OCR-D[2] neben technischen Lösungen für die Massenvolltexterschließung historischer Drucke vom 16. bis 19. Jahrhundert eigene GT-Richtlinien[3]. Diese Richtlinien werden in einer offenen, zur kollaborativen Datenkultur verpflichtenden Umgebung erstellt und sollen sicherstellen, dass nachnutzbare Forschungsdaten entstehen sowie der Aufwand in der GT-Erstellung minimiert werden kann.[4]

Forschungsdaten im Kontext des Deutschen Textarchives

Im Rahmen des vorgeschlagenen Workshops soll eine solche offene Datenkultur am Beispiel von Forschungsdaten des Deutschen Textarchivs (DTA)[5] gemeinsam gelebt und so mittelbar ein wertvoller Beitrag zur Qualität historischer Textkorpora geleistet werden. Die Analyse des DTA vor dem Hintergrund der GT-Erstellung soll den Teilnehmenden zeigen, welche Möglichkeiten (und Grenzen) diese Daten bieten.

Betrachtung des DTA-Datenbestandes

Das DTA wurde im Rahmen eines sprachwissenschaftlich orientierten DFG-Projektes erstellt. Der Kernbestand besteht aus 1500 Druckpublikationen mit einem Gesamtumfang von 540.000 Seiten. Die Text- und Textsortenauswahl, die zeitliche Spanne des Publikationszeitraumes vom frühen 17. bis frühen 20. Jahrhundert, die Verwendung von Erstausgaben und die vorlagengetreue Transkription kennzeichnen diesen Bestand als Grundlage eines Referenzkorpus der frühneuhochdeutschen Sprache. Die Bereitstellung der digitalen Texte erfolgt sowohl in einem XML-basierten Format als auch als unannotierter Rohtext.

Für die Einschätzung der Nutzbarkeit des DTA als Quelle für GT sind nicht nur die Ergebnisdaten relevant. Ein genauerer Blick auf die einzelnen Etappen des ursprünglichen Datenerfassungsworkflows im DTA zeigt bisher ungenutzte Potenziale der einzelnen Datenstände als Trainingsmaterialien für Text- und Strukturerkennung. Die folgende Abbildung illustriert die beiden grundsätzlichen Wege der Volltexterstellung, die im DTA zur Anwendung kamen: Automatische Texterkennung mit anschließender Nachkor-

rektur („OCR way“) und manuelle Transkription im Vier-Augen-Prinzip („Double Keying way“).

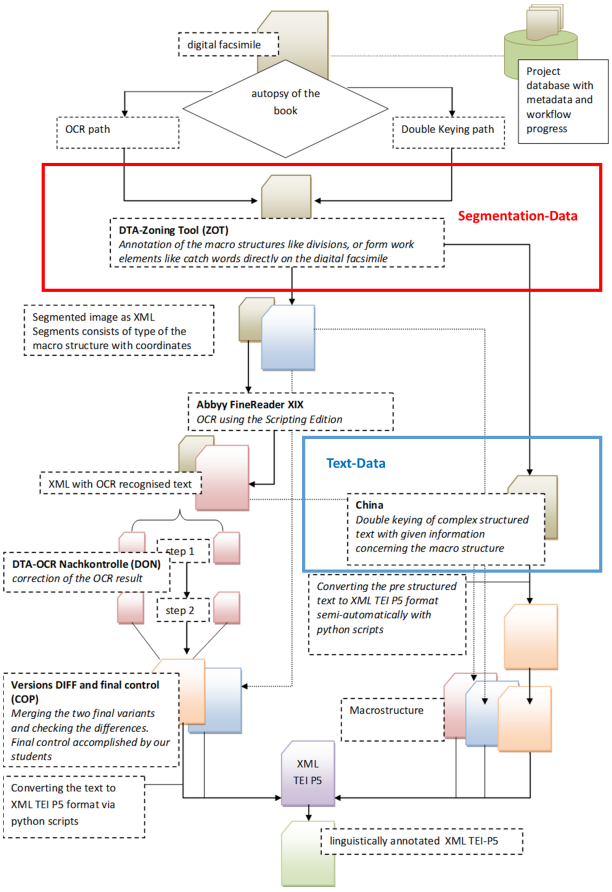


Abbildung 1: Schematische Darstellung des DTA-Datenerfassungsworkflows

Letzterer kam für den Großteil des Bestands zur Anwendung. Das Double-Keying-Verfahren wurde von Nicht-Muttersprachlern vorgenommen und ist sehr genau. Die Zeichengenauigkeit kann mit 99,99 % angesetzt werden (Haaf, 2013; Geyken, 2012). Mit OCR wurden hauptsächlich Titel des 19. und Mitte des 18. Jahrhunderts erfasst.[6] Für dieses Schrifttum existieren hoch performante und zuverlässige OCR-Modelle.[7]

Beiden Wegen gemein ist ein manueller Segmentierungsschritt. In diesem wurden Textzonen und Abbildungen lokalisiert und klassifiziert. Diese Segmentierung diente zwar „nur“ der nachträglichen Auszeichnung der Volltexte im XML (und nicht etwa der Unterstützung der automatischen Texterfassung). Sie bilden aber dennoch eine der größten bekannten Sammlungen an Strukturdaten für historische deutschsprachige Drucke. Aus der Untersuchung des Datenerfassungsworkflows können somit Segmentierungsdaten und Textdaten identifiziert werden, die für die Verwendung als GT in Frage kämen. Größtes Manko der Datensammlung ist jedoch die fehlende Verknüpfung zwischen Text und Bild, die die Einsatzszenarien als Trainingsdaten massiv einschränkt. An dieser Stelle setzt der vorgeschlagene Workshop an.

Ziel

Die Teilnehmenden des Workshops werden mit Verfahren und Methoden der Erstellung, Erschließung und Speicherung von GT für die automatische Texterkennung vertraut gemacht. Der Workshop ist in zwei Teile geteilt: einen theoretischen und einen praktischen. Ziel des theoretischen Teils ist, dass die Teilnehmenden in die Lage versetzt werden, anhand einer Liste von Kriterien sowie einer Validierung der Daten, Forschungsdaten für die Erstellung von GT einzuschätzen. Mit den OCR-D-GT-Richtlinien bekommen die Teilnehmenden eine in der Praxis erprobte Anleitung für die Erstellung von GT zur Verfügung gestellt. Inhalt und Aufbau, aber auch die Möglichkeiten der praktischen Anwendung dieser Richtlinien im jeweiligen Projekt bilden in diesem ersten Workshopteil den Schwerpunkt. Im praktischen Teil sollen nun die Teilnehmenden in verschiedenen Szenarien GT-Daten erstellen. Dabei werden Forschungsdaten des DTA und vorhandener GT geprüft und eingeschätzt. Dazu werden die im theoretischen Teil vorgestellten Metriken und Validierungsmethoden angewendet. Mit Transformations- und Konvertierungsprogrammen kann in der Folge nun der GT automatisiert erstellt werden. Ebenfalls können spezielle Softwareprogramme für die manuelle Erstellung von GT verwendet werden. Um sich sowohl mit dem Funktionsumfang als auch mit der Leistungsfähigkeit der Tools vertraut zu machen, ist es notwendig, diese im theoretischen Teil kennenzulernen. Der unmittelbare Umgang und die Handhabung des Tools für die GT-Erstellung stehen nicht im Mittelpunkt, sondern die Entscheidung, welches Tool für das jeweilige Vorhaben am geeignetsten scheint.

Zum Abschluss steht die Speicherung des GT in einem Repositorium. So können die Daten entsprechend der FAIR-Prinzipien zugänglich gemacht werden. Erklärungen zum Aufbau des Repositoriums sowie die Erschließung mit Metadaten, die Nutzung des OCR-D-GT-Repo-Template[8] schließen diesen Teil und den Workshop ab.

Inhalte

Den Teilnehmenden des Workshops sollen verschiedene Methoden und Verfahren der GT-Erstellung vorgestellt werden.

Theoretischer Teil

1. Prüfung und Bewertung von Forschungsdaten
2. Vorstellung der OCR-D-GT-Richtlinien
3. Vorstellung von Verfahren zur Alignierung von existierenden Transkriptionen und generierter Segmentierung
4. Vorstellung von Softwaretools

Praktischer Teil

1. Erstellung von GT aus Forschungsdaten
   1. Bewertung der Forschungsdaten
   2. Transformation, Konvertierung der Forschungsdaten
   3. Erstellung und Validierung des GT
2. Erstellung des GT durch Transkription
   1. Vorstellung und Anwendung von Transkription-GT-Tools
3. Speicherung und Veröffentlichung des GT
   1. Erstellung eines GT-Repositoriums auf GitHub

Benötigte technische Ausstattung:

die jeweiligen Teilnehmenden verfügen über:

- einen GitHub-Account
- Laptop mit installierten Werkzeugen (Liste wird vorab per E-Mail geschickt)
- Ggf. eigene Daten

Der Raum verfügt über:

- Beamer, Whiteboard (wenn möglich)
- Internet via W-Lan

Umfang:

- vier Stunden (90 Minuten Theorie 30 Minuten Pause 120 Minuten Praxis)

Forschungsfeld der Beitragenden

Matthias Boenig ist Informationswissenschaftler sowie Kunsthistoriker. Er hat Bibliotheks- und Informationswissenschaftler und Kunstgeschichte an der Humboldt-Universität zu Berlin studiert. Seit seinem Studium beschäftigt er sich mit der digitalen Transformation von Textdaten in digitale, strukturierte und XML-basierte Forschungsdaten. Dazu war er in verschiedenen Projektkontexten von 1997 an, zu Beginn am Computer- und Medienservice der Humboldt-Universität, dem Institut für Bibliotheks- und Informationswissenschaft und heute an der Berlin-Brandenburgischen Akademie der Wissenschaften tätig. Zurzeit ist Matthias Boenig wissenschaftlicher Mitarbeiter im Projekt OCR-D. Im Rahmen dieses Projekts hat er die OCR-D-GT-Richtlinien entwickelt und betreut diese. Sein derzeitiges praktisches und forschungsorientiertes Interesse besteht in der Erstellung, der Bereitstellung und Standardisierung von GT für die OCR. Matthias Boenig war Mitarbeiter am „Deutschen Textarchiv“.

Kay-Michael Würzner ist Sprachwissenschaftler und hat Computerlinguistik und Germanistik studiert. Nach dem Studium arbeitete er als wissenschaftlicher Mitarbeiter an der Universität Potsdam und der Berlin-Brandenburgischen Akademie der Wissenschaften im Bereich korpuslinguistischer Forschungsdateninfrastrukturen. Seit April 2019 ist Kay-Michael Würzner an der SLUB tätig und bearbeitet Themen des maschinellen Lernens und der automatischen Sprachverarbeitung. Er koordiniert außerdem die Angebote der SLUB rund um einen offenen Forschungskreislauf.

Konstantin Baierer arbeitet seit 2018 als wissenschaftlicher Mitarbeiter für die Staatsbibliothek zu Berlin am Projekt OCR-D, insbesondere an der technischen Interoperabilität der entwickelten Lösungen, der OCR-D/core Softwarebibliothek und dem Release Management. Besonders wichtig sind ihm transparente, inklusive und robuste Methoden für verteilte Softwareentwicklung und gute Schnittstellen zwischen Kulturerbeeinrichtungen und Digital Humanities.

Lena Hinrichsen ist wissenschaftliche Mitarbeiterin an der Herzog August Bibliothek Wolfenbüttel und dort seit 2021 als Koordinatorin im Projekt OCR-D tätig. Ihr Studium der Buchwissenschaft absolvierte sie an der Johannes Gutenberg-Universität Mainz.

Dr. Christian Reul leitet die Digitalisierungseinheit des Zentrums für Philologie und Digitalität (ZPD) der Universität Würzburg. Seine Forschungsschwerpunkte sind die OCR/HTR auf historischem Material sowie die Neu- und Weiterentwicklung der entsprechenden Software.

Fußnoten

1. vgl. OCR-D-GT-Richtlinie < https://ocr-d.de/de/gt-guidelines/trans/trLevels.html >
2. DFG-Projekt OCR-D : Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR), Koordinierungsprojekt < https://ocr-d.de/de/about >
3. OCR-D-GT-Richtlinien <https://ocr-d.de/de/gt-guidelines/trans/>
4. siehe dazu Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke (Engl 2020)
5. Deutsches Textarchiv <https://deutschestextarchiv.de/>
6. siehe dazu Deutsches Textarchiv – Der Digitalisierungsworkflow im DTA < https://deutschestextarchiv.de/doku/workflow>
7. Siehe dazu GT4Hist-GT-Datensatz mit Korrekturen: https://code.bib.uni-mannheim.de/ocr-d/GT4HistOCR, Training Fraktur from GT4HistOCR: https://github.com/tesseract-ocr/tesstrain/wiki/GT4HistOCR, Modelle: GT4HistOCR: https://code.bib.uni-mannheim.de/ocr-d/GT4HistOCR/src/branch/master/models, frak2021: https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/
8. OCR-D/gt-repo-template: A template for creating a ground truth repo with the various functions and features: such as metadata creation, data analysis and presentation. (github.com) < https://github.com/OCR-D/gt-repo-template >

Bibliographie

Engl, Elisabeth; Boenig, Matthias; Baierer, Konstantin; Neudecker, Clemens; Hartmann, Volker. 2020: „Volltexte für die Frühe Neuzeit : Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke“. Zeitschrift für Historische Forschung, 47(2), 223-250. doi:10.3790/zhf.47.2.223.

Haaf, Susanne; Wiegand, Frank; Geyken, Alexander: „Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text“. Journal of the Text Encoding Initiative (jTEI) 4, 2013. doi:10.4000/jtei.739.

Geyken, Alexander; Haaf, Susanne; Jurish, Bryan; Schulz, Matthias; Thomas, Christian; Wiegand, Frank: „TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv“. Jahrbuch für Computerphilologie – online, 2012, http://computerphilologie.digital-humanities.de/jg09/geykenetal.html.

Erstellung von DH Workflows im SSH Open Marketplace

Barbot, Laure
laure.barbot@dariah.eu
DARIAH
ORCID: 0000-0002-6008-7959

Illmayer, Klaus
klaus.illmayer@oeaw.ac.at
Österreichische Akademie der Wissenschaften - ACDH-CH, Österreich
ORCID: 0000-0001-7253-996X

König, Alexander
alex@clarin.eu
CLARIN
ORCID: 0000-0002-8540-2396

Kontaktdaten der Workshopleiter:innen und Forschungsinteressen

Klaus Illmayer, klaus.illmayer@oeaw.ac.at, Austrian Centre for Digital Humanities and Cultural Heritage an der Österreichischen Akademie der Wissenschaften. Forschungsinteressen: Digitale Forschungsplattformen und -infrastrukturen, Datenmodellierung, Erstellung digitaler Forschungsworkflows besonders im Kontext kultureller Ausdrucksformen

Alexander König, alex@clarin.eu, CLARIN ERIC. Forschungsinteressen: Digitale Forschungsinfrastrukturen, Metadatenstandards, Forschungsdatenmanagement (FAIR)

Michael Kurzmeier, michael.kurzmeier@oeaw.ac.at, Austrian Centre for Digital Humanities and Cultural Heritage an der Österreichischen Akademie der Wissenschaften

Detailbeschreibung

Der SSH (Social Sciences and Humanities) Open Marketplace (in weiterer Folge kurz SSHOMP) – https://marketplace.sshopencloud.eu – wurde im Zuge des EU-finanzierten Horizon 2020-Projektes SSHOC (Social Sciences and Humanities Open Cloud, https://sshopencloud.eu/ , Grant Agreement #823782) entwickelt. Nach dem Laufzeitende von SSHOC 2022 wird der SSHOMP durch ein Konsortium von DARIAH (https://www.dariah.eu/), CLARIN (https://www.clarin.eu/), CESSDA (https://www.cessda.eu/) und Partner:innen weiterbetrieben. Ein Editorial Board – in dem die Workshopleiter:innen Mitglieder sind – sorgt für eine beständige Kuration, Dissemination und Erweiterung des SSHOMP. Ein besonderes Augenmerk wird darauf gelegt, neue Inhalte zu generieren und mit den bereits vorhandenen Einträgen zu verbinden. Damit sollen neben der beständigen Qualitätssteigerung der Metadaten auch kontextuelle Zusammenhänge verstärkt aufgezeigt werden. Zu diesem Zwecke wird ein besonderer Fokus auf die Erstellung von Workflows durch Forscher:innen möglichst vieler Fachbereiche gerichtet, die digitale Methoden in den Geistes- und Kulturwissenschaften sowie den Sozialwissenschaften (SSH) anwenden.

Der SSHOMP ist ein „discovery portal“, um digitale Ressourcen auffindbar zu machen, die in der Forschung mit digitalen Mitteln angewandt werden. Im Zentrum stehen dabei Tools und Services (oder anders formuliert: Softwareprodukte und Webanwendungen), die in Forschungsabläufen eingesetzt werden. Daneben werden auch Publikationen, Trainingsmaterialien und Datensätze verzeichnet, die mit diesen Tools und Services in Verbindung stehen, wenn sie bspw. erläutern, wie ein Tool verwendet werden kann.

Erfasst werden im SSHOMP die Metadaten und Relationen dieser digitalen Ressourcen. Dafür wird zum einen ein Harvesting-Service betrieben, der die Daten von existierenden Plattformen wie TAPoR (https://tapor.ca/), dem Programming Historian (https://programminghistorian.org/) oder dblp (https://dblp.org/ - für Abstracts von verschiedenen DH-Konferenzen) einpflegt. Zum anderen wird eine Kuration dieser importierten Daten vorgenommen (manuell und automatisiert) und es werden neue Einträge vorgenommen.

Eine spezielle Kategorie von Ressourcen sind die Workflows (https://marketplace.sshopencloud.eu/search?categories=workflow), die einen Anwendungsfall beschrei-

ben, der aus mehreren Schritten („steps“) besteht. Ein solcher Workflow kann die Anwendung eines Tools zu Forschungszwecken sein (z.B. wie mit dem Tool „Celluloid“ Videoannotationen vorgenommen werden können: https://marketplace.sshopencloud.eu/workflow/tYY6xe), die Vorgehensweise um eine digitale Methode umzusetzen (z.B. die digitale Langzeit-Archivierung von 3D-Objekten: https://marketplace.sshopencloud.eu/workflow/ctyv0s) oder der Ablauf eines Forschungsprojekts (z.B. die Erstellung einer digitalen Edition eines Musikkorpus: https://marketplace.sshopencloud.eu/workflow/bKefY4). Workflows haben einen beschreibenden Charakter und abstrahieren eine digitale Vorgehensweise auf der Ebene kleinerer Einheiten, nämlich der methodischen Schritte die nötig sind, um eine Aufgabe zu beginnen, auszuführen und zu beenden. Der Detailgrad der Ausführungen unterscheidet sich und ist pragmatisch anzulegen. Entscheidend sind eine verständliche Darlegung und die Relation zu Tools und Services, die für die einzelnen Schritte oder auch für den gesamten Workflow Verwendung finden.

Workflows erfüllen dadurch mehrere Zwecke: Sie machen andere Einträge des SSHOMP besser auffindbar, sie kontextualisieren diese Einträge und sie regen zur Auseinandersetzung mit “best practices” eines Forschungsbereichs an. Insbesondere zielen sie darauf ab, das Zusammenspiel von Tools und Services in der Umsetzung eines Forschungsablaufs transparent aufzuzeigen, womit Einsteiger:innen eine Handhabung zur Umsetzung ihrer Forschungsidee aufgezeigt wird und fortgeschrittenen Benutzer:innen Anregungen zur Überprüfung und Erweiterung ihres Forschungsablaufs gegeben werden.

Die Erstellung von Workflows wiederum bietet Forscher:innen die Möglichkeit, ihre digitale Vorgehensweise zu dokumentieren und im Forschungsfeld – aber auch darüber hinaus – zu teilen. Damit kann ein Beitrag geleistet werden zur Überprüfbarkeit von Forschungsresultaten, aber insbesondere sollen davon Forscher:innen profitieren, die noch wenig erfahren sind mit digitale Methoden, indem sie Workflows aufgreifen, anwenden und weiterentwickeln. Durch die Relation der Workflowschritte zu Tools und Services kann eine Forschungsumgebung rekonstruierbar vermittelt werden. Zugleich kann implizites Wissen weitergegeben und die Verwendung spezifischer Tools und Services gefördert werden, die bekannt dafür sind, Forschungsstandards zu unterstützen.

Im Workshop werden Forscher:innen, Research Software Engineers, Data Stewards und alle weiteren Interessierten dazu angeleitet, einen Workflow auszuarbeiten, diesen im SSHOMP einzutragen und anschließend zu publizieren. Damit einher geht eine grundsätzliche Auseinandersetzung mit den Möglichkeiten des SSHOMP um neue Einträge zu erstellen, bestehende zu ergänzen und miteinander zu verlinken. Besonderes Augenmerk wollen wir auf die Anreicherung des SSHOMP durch diese neuen Workflows legen, die eine verbesserte Kontextualisierung ermöglichen und dadurch das Sucherlebnis steigern, besonders wenn Suchkriterien zu Beginn noch unklar sind (z.B. mittels „serendipity“).

Diskutiert werden im Workshop auch die Vorteile der Veröffentlichung von Workflows im SSHOMP. Hervorzuheben sind dabei die Möglichkeiten der Reflexion der eigenen methodischen Vorgehensweise, der Erfahrungsaustausch durch eine Peer-Gruppe vor Ort – z.B. der Austausch über die Verwendung von unterschiedlichen Tools für vergleichbare digitale Verfahren – oder durch die Benutzer:innengruppe des SSHOMP nach Veröffentlichung des Workflows sowie die Weiterverwendung von Workflows in anderen Kontexten bspw. durch Referenzierung in einem Blog-Beitrag oder einem Forschungsartikel.

Interessiert sind wir an einem ausführlichen Feedback durch die Workshopteilnehmer:innen, um die Weiterentwicklung des SSHOMP besser steuern und gestalten zu können. Neben dem Auffinden allfälliger Bugs und Hinweisen für eine besseres User Experience (UX-Design), sind wir an Vorschlägen zur Erweiterung des Metadatenschemas sowie für neue Ideen zur Kontextualisierung und Erweiterung der Suchmöglichkeiten des SSHOMP interessiert.

Zum Ende des Workshops sollen Teilnehmer:innen neben der Erstellung eines Workflows auch ein besseres Verständnis für die Herausforderungen einer digitalen „discovery platform“ hinsichtlich dem kuratorischen und institutionellen Aufwand gewonnen haben. Zugleich bietet sich am Beispiel von SSHOMP die Möglichkeit, über Herangehensweisen zur Dokumentation, Auffindbarkeit und Weiternutzung („FAIR principles“) digitaler Methoden zu reflektieren.

Format

Der Workshop ist auf einen halben Tag (4 Stunden, inkl. Pause) ausgelegt. Zunächst wird eine generelle Einführung in den SSHOMP gegeben, um danach beispielhaft die Erstellung eines Workflows im Detail zu zeigen (1 Stunde). Anschließend werden offene Fragen geklärt und die Benutzer:innen-Accounts können über die Federated AAI der EOSC erstellt werden (Forscher:innen mit einem Eduroam-Account können sich über diesen institutionellen Zugriff registrieren, daneben ist ein Login mit ORCID, Google und weiteren Authentifizierungsservices möglich). Zugleich kann bereits in Kleingruppen oder auch alleine eine Idee für einen Workflow formuliert werden. Dieser erste Block vor der Pause wird mit einer Rundum-Präsentation dieser Ideen abgeschlossen (1 Stunde). Nach der Pause (20 Minuten), die auch zur Vernetzung und zum Austausch genutzt werden kann, gibt es eine kurze Feedback-Runde durch die Workshopleiter:innen, bevor die Workflow-Ideen als Hands-On-Session individuell oder in Gruppen im SSHOMP erstellt werden. Dabei stehen die Workshopleiter:innen unterstützend zur Seite, um Probleme zu besprechen und Tipps zu geben (1 Stunde und 10 Minuten). Die verbleibenden 30 Minuten werden für eine Kurzpräsentation der erstellten Workflows sowie eines

Feedbacks durch die Teilnehmer:innen und Workshopleiter:innen genutzt.

Das Format wurde bereits mehrfach erprobt, u.a. auf der DH2023 in Graz (Barbot et al., 2023), und hat sich als produktiv erwiesen. Da Workflows auch noch nach Ende des Workshops weiterbearbeitet werden können – bzw. die Publikation der Workflows nicht sofort nötig ist –, ist eine individuelle Weiterführung der begonnenen Arbeit jederzeit möglich. Die Workshopleiter:innen sind auch nach Ende des Workshops per E-Mail erreichbar, um die nötige Unterstützung zur Finalisierung eines Workflows zu geben.

Zielpublikum

Eingeladen sind insbesondere Forscher:innen, die eine Forschungsvorgehensweise digital abbilden möchten. Dabei sind eigene Herangehensweisen aber auch abstrakte Abläufe denkbar, die im Workshop als Workflows modelliert und schrittweise beschrieben werden. Eine gute Möglichkeit ist auch, eine Posterpräsentation oder einen Vortrag auf der DHd als Workflow abzubilden.

Es ist kein besonderes Vorwissen nötig, außer das Interesse an der Erstellung von Inhalten auf einer digitalen Plattform und der Bereitschaft, diese Inhalte zu veröffentlichen. Falls keine eigenen Workflow-Ideen gefunden werden, stellen die Workshopleiter:innen Material zur Verfügung, auf dessen Basis ein Workflow erstellt werden kann. Auch kann statt der Erstellung eines Workflows bereits bestehende Workflows ergänzt werden.

Anzahl Teilnehmer:innen

20 Teilnehmer:innen sind eine gute Zahl, um eine individuelle Betreuung zu gewährleisten. Erfahrungsgemäß findet sich ein Großteil der Teilnehmer:innen schnell mit dem SSHOMP und den Editierfunktionen zurecht. Da wir zumindest zwei Workshopleiter:innen sein werden, ist eine Ansprechperson dauerhaft für generelle Fragen aus allen Gruppen verfügbar. während die andere Person individuelle Problemlösungen vornehmen kann.

Im Prinzip sind auch mehr als 20 Teilnehmer:innen möglich. Wir würden in so einem Fall versuchen, eine dritte Person als Workshopleiter:in hinzuzuziehen, damit dem vermehrten Betreuungsaufwand ausreichend begegnet wird.

Benötigte technische Ausstattung

Für die einleitende Präsentation benötigen wir einen Beamer mit HDMI-Anschluss, damit die Folien gezeigt werden können. Des Weiteren ist ein gut funktionierender Internetzugang eine Voraussetzung, da die Workflows direkt online in der Plattform erstellt werden.

Teilnehmer:innen sollten idealerweise einen Laptop mitbringen. Es ist aber auch möglich, sich ohne Laptop in einer Kleingruppe an der konzeptuellen Entwicklung eines Workflows zu beteiligen bzw. gemeinsam den Workflow auf einem Laptop zu erstellen.

Workshop-spezfischer Call for Papers

Ist nicht vorhanden und nicht vorgesehen. Es wird überlegt, in Kooperation mit DARIAH und CLARIN aus verfügbaren Projektgeldern eine oder zwei Anmeldegebühr/en für den DHd-Kongress zu finanzieren, um eine Workshopteilnahme zu ermöglichen. Dafür würde es dann einen gesonderten Call geben, dies ist aber noch in Verhandlung.

Bibliographie

Barbot, Laure, Edward J. Gray, und Klaus Illmayer. 2022. “The SSH Open Marketplace: Workflow Curation Sprint.” https://doi.org/10.5281/zenodo.6375078 (zugegriffen: 19. Juli 2023).

Barbot, Laure, Elena Battaner Moro, Stefan Buddenbohm, Cesare Concordia, Maja Dolinar, Matej Ďurčo, Edward Gray, Cristina Grisot, Klaus Illmayer,Martin Kirnbauer, Mari Kleemola, Alexander König, Michael Kurzmeier, Barbara McGillivray, Clara Parente Boavida, Christian Schuster, Irena Vipavc Brvar und Magdalena Wnuk. 2023. “Creating a DH workflow in the SSH Open Marketplace.” In *Digital Humanities 2023: Book of Abstracts*, hg. von Baillot, Anne, Walter Scholger, Toma Tasovac und Georg Vogeler, 31-33. https://doi.org/10.5281/zenodo.7961822(zugegriffen: 19. Juil 2023).

Gray, Edward, Nicolas Larrousse, Clara Petitfils, Laure Barbot, Frank Fischer, Matej Ďurčo, Klaus Illmayer, Cesare Condordia, Alexander König, Dieter Van Uytvanck und Stefan Buddenbohm. 2021. “D7.4 Marketplace – Data Population & Curation.” https://doi.org/10.5281/zenodo.5783358 (zugegriffen: 19. Juli 2023).

König, Alexander und Dieter Van Uytvanck. 2020. “D7.3 Marketplace - Interoperability”. https://doi.org/10.5281/zenodo.5871651 (zugegriffen: 19. Juli 2023).

Puren, Marie, Klaus Illmayer, Laurent Romary, Nicolas Larrousse, Laure Barbot, Matej Ďurčo, Edward Gray und Charles Riondet. 2023. “How to create a workflow in the SSH Open Marketplace?” https://marketplace.sshopencloud.eu/workflow/hmGpmv (zugegriffen: 19. Juil 2023).

SSHOC. 2022a. “SSHOC Legacy Booklet.” https://doi.org/10.5281/zenodo.6394462 (zugegriffen: 19. Juli 2023).

SSHOC. 2022b. “SSHOcingly Good and Sustainable Tools.” https://doi.org/10.5281/zenodo.6404957 (zugegriffen: 19. Juil 2023).

Evaluating Digital Humanities Methods and Tools for the OpenMethods Metablog: An OpenMethods Edit-a-thon

Nunn, Christopher
christopher.nunn@theologie.uni-heidelberg.de
Universität Heidelberg, Deutschland
ORCID: 0000-0001-7208-8636

Horváth, Alíz
aliz.horvath06@gmail.com
Eötvös Loránd University, Ungarn
ORCID: 0000-0002-9131-5504

Wuttke, Ulrike
ulrike.wuttke@fh-potsdam.de
Fachhochschule Potsdam, Deutschland
ORCID: 0000-0002-8217-4025

Introduction

In digitally oriented disciplines, such as the Digital Humanities (DH), the exchange of scholarly information is not confined solely to peer-reviewed journals and academic book series. Instead, it occurs naturally in various online platforms such as blogs, Git repositories, pre-prints, video tutorials, podcasts, social media, and other web 2.0 and 3.0 discourse spaces (AG Digitales Publizieren, 2021). This is particularly crucial when sharing critical reflections and expertise related to tools and methodologies (König, 2015).

The DARIAH-EU-sustained OpenMethods metablog aims to explore and deliver solutions for the rich and dynamically evolving DH scene that challenges the traditional formats of scholarly communication that is deeply rooted in print culture. It offers a comprehensive platform for consolidating diverse forms of open access content in multiple languages, specifically focusing on methods and tools within the field of DH. It provides a convenient way for DH experts from around the globe to select, propose, curate, highlight and evaluate open access content. It disseminates knowledge and enhances the recognition of DH methods and tools among peers and embraces a broad range of publication formats, aiming for inclusivity by acknowledging content types that often go unnoticed in formal research evaluation and conventional research papers and book chapters (Eve, 2020, Neuber und Sahle, 2022).

The OpenMethods metablog approach means that its Editorial Team is actively involved in the selection and curation of published content. The content is proposed by the OpenMethods Editorial Team members themselves as well as by community volunteers (Engelhardt et al., 2017). Additionally, the nomination of content is open to anyone, either through social media or through the nomination tool on the OpenMethods platform. External collaborators, including students studying DH, are especially encouraged to contribute and be recognized on the OpenMethods website. After the process of commenting on, filtering, and curating the nominations based on the established evaluation criteria, successful entries not only get republished on the platform, but also get categorized using the Taxonomy of Digital Research Activities in the Arts and Humanities (TaDiRAH) developed by Borek et al. (2016; 2021). Additionally, a brief introduction in English is provided by one of the Editorial Team members, explaining the significance and relevance of the contribution.

OpenMethods aims to ensure a wide range of coverage in various languages. The Editorial Team has deliberately chosen an interdisciplinary and multilingual approach with the goal of showcasing the diversity of DH discussions within different regional, national, and language communities (Del Rio Rande et al., 2018). Moreover, OpenMethods aims not only to enhance the representation of traditionally marginalized languages and actors, such as female tool-makers in the Digital Humanitiesand a particular focus on non-Anglophone, under-resourced languages (such as those with non-Latin scripts), but to also ensure that the content selected handles various topics, including method and tool descriptions, critical evaluations of tools and methods, and both practical and theoretical reflections on the digital conduct of humanities research. This is done to explore how the growing influence of digital methods and tools is reshaping scholarly attitudes and scientific practices within humanities research. This objective holds significance both for the OpenMethods platform specifically and for the broader discourse in the field of Digital Humanities, as highlighted by Horvath (2021) and related research.

Recognizing the vast and dynamic nature of born-digital scholarly discourse spaces, there is a need to establish mechanisms for filtering and building trust in the openly accessible content within them (Eve, 2020; Fitzpatrick, 2011; Nyhan, 2020). This requirement has been addressed by OpenMethods, a platform that leverages the expertise of the scholarly community to establish such filtering and trust-building mechanisms. During the OpenMethods Edit-a-Thon at DH Graz (Tóth-Czifra et al., 2023) a vivid discussion arose around the selection and evaluation process that showed the necessity to further develop the scope of the metablog's selection and the direction of the evaluation process to meet inclusivity and diversity criteria.

Goals of the Workshop

In this workshop, we invite scholars in the arts and humanities, regardless of their career stage or disciplinary and geographical backgrounds, to explore the OpenMethods metablog as an innovative platform for publication.

Nominated papers for the OpenMethods metablog are reviewed according to seven different evaluation criteria: scope, openness, relevance, clarity, diversity, language, as well as assessment and validation. During the workshop, each of these criteria will be considered and discussed in more detail. For example, should contributions only be considered if they are in formats without peer review like blogposts or podcasts? Should the preparation of an introduction by the editors follow certain guidelines in order to increase its quality and thus also gain value for their own publication list (Neuber und Sahle, 2022)? Under specific circumstances, could reports of research results without explicit method or tool criticism also be considered? Given the development of the computational humanities (Piotrowski, 2020), should it be possible to nominate papers that are aimed exclusively at a technically literate audience? How can or should a metablog do justice to diversity in the sense of uncovering data biases and hidden hierarchies in working processes (D'Ignazio und Klein, 2020)? Is multilingualism justified and timely, or is it even more inclusive to focus on a few dominant languages and use more and more improved translation software?

One of the main goals of the workshop is to revisit the OpenMethods metablog evaluation criteria through an interactive conceptual discussion with the workshop participants. The discussion will be followed by a hands-on phase: selecting and evaluating nominations and writing introductions. Ultimately the engagement of the participants with the metablog will serve to foster sustainability through reuse and community building by involving participants in the editorial process and enticing them to spread the word and join the team.

Workshop Outline

The workshop is designed as a half-day event (4 hours) consisting of an in-depth critical discussion of the review criteria and a hands-on writing phase.

The format will be an edit-a-thon, with a major focus on interactive discussion, as well as a hands-on writing phase after an introductorypresentation on OpenMethods as a platform, its goals and the evaluation criteria.

Structure of the workshop (total 240 min):

Introduction and Q&A (30 min)

Introduction to the evaluation criteria and discussion (30 min)

Nomination phase (suggestion key topical areas: non-English content) (30 min)

BREAK (15 min)

Self-Evaluation (Checkbox exercise) (30 min)

Deliberation phase (15 min)

Groups formation and choosing of content to collaboratively write introductions (15 min)

Break (15 min)

Hands-On Writing Phase (50 min.)

Wrap-up, takeaways, reflections (10 min)

Target Audience and Participant Number

The intended target group is people with an interest in the topic of the workshop. There are no prerequisites for participation. The main language of the workshop presentations will be English, German contributions to the discussions are welcomed. Nominations are encouraged in any language.

For the hands-on part, each participant should bring a computer or tablet.

Group size: 20 participants.

Required Technical Equipment

For technical equipment, a projector is required. Due to the interactive nature of the workshop, online participation will unfortunately not be feasible.

Furthermore, the following equipment will be required: whiteboard or flipchart.

Facilitators and Research Interests

Alíz Horváth has a PhD in East Asian Languages and Civilizations from the University of Chicago and currently works as assistant professor of East Asian History and Digital Humanities at Eötvös Loránd University (Hungary). She is a member of the core editorial team of OpenMethods since 2020 and has been a contributor to the pioneering project New Languages for NLP, organized by Princeton. She is also founder and co-chair of the DARIAH-EU Multilingual DH Working Group.

Christopher Nunn is Postdoc at Heidelberg University. He has a PhD in Church History (Ancient Christianity) and founded the Interdisciplinary Forum of Digital Textual Studies (InFoDiTex) as well as the TheoLab, a research network for computational theology at Heidelberg University. He is a member of the OpenMethods Editorial Team since 2019. His research interests are in community building and science communication among others.

Ulrike Wuttke has been the Deputy Chief Editor since the launch of OpenMethods in 2017. She has a PhD in Medieval Dutch Literature from Ghent University (Belgium) and currently is professor of Library Science at University of Applied Sciences Potsdam. Her research interests include scholarly communication and Open Science. Among others she is a member of the editorial board of the Zeitschrift für Digitale Geisteswissenschaften (ZfDG).

Acknowledgments: OpenMethods was initiated during the Horizon 2020 project Humanities at Scale (2020). It has since then been sustained by DARIAH-EU (Digital Research Infrastructure for the Arts and Humanities). DARIAH's approach to supporting and strengthening the Digital Humanities entails validation of Data Services and tools, building networks of people, providing training and education opportunities, enabling transnational and transdisciplinary approaches across Europe.

Bibliographie

AG Digitales Publizieren. 2021. Digitales Publizieren in den Geisteswissenschaften: Begriffe, Standards, Empfehlungen (= Zeitschrift für digitale Geisteswissenschaften / Working Papers, 1). *Zeitschrift für Digitale Geisteswissenschaften*. https://doi.org/10.17175/wp_2021_001 (zugegriffen: 18. Juli 2023).

Borek, Luise, Quinn Dombrowski, Jody Perkins und Christof Schöch. 2016. "TaDiRAH: A Case Study in Pragmatic Classification." DHQ 10, Nr. 1. http://www.digitalhumanities.org/dhq/vol/10/1/000235/000235.html (zugegriffen: 18. Juli 2023).

Borek, Luise, Canan Hastik, Vera Khramova, Klaus Illmayer und Jonathan D. Geiger. 2021. "TaDiRAH Revised, Formalized and FAIR", In *Information between Data and Knowledge*. Information Science and its Neighbors from Data Science to Digital Humanities - Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th - 10th March 2021 (=Schriften zur Informationswissenschaft 74), hg. von Thomas Schmidt und Christian Wolff, 321-332. Glückstadt: Hülsbusch. https://epub.uni-regensburg.de/44951/ (zugegriffen: 18. Juli 2023).

D'Ignazio, Catherine und Lauren F. Klein. 2020. *Data Feminism*. Cambridge (Mass.), London: The MIT Press.

Engelhardt, Claudia, Claudio Leone, Nicolas Larrousse, Delphine Montoliu, Yoann Moranville, Pierre Mounier, Jenny Oltersdorf, Paulin Ribbe und Ulrike Wuttke. 2017. *Open Humanities Methods Review Journal (Research Report)*. DARIAH; TGIR Huma-Num (UMS3598); Göttingen State and University Library. https://hal.archives-ouvertes.fr/hal-01685852 (zugegriffen: 18. Juli 2023).

Eve, Martin Paul. 2020. Violins in the Subway: Scarcity Correlations, Evaluative Cultures, and Disciplinary Authority in the Digital Humanities. In *Digital Technology and the Practices of Humanities Research*, hg. von Jennifer Edmond, 103–122. Cambridge, United Kingdom: Open Book Publishers. https://doi.org/10.11647/OBP.0192 (zugegriffen: 18. Juli 2023).

Fitzpatrick, Kathleen. 2011. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York: New York University Press.

Horváth, Aliz. 2021. "Enhancing Language Inclusivity in Digital Humanities: Towards Sensitivity and Multilingualism: Includes interviews with Erzsébet Tóth-Czifra and Cosima Wagner". In *Modern Languages Open* 1, http://doi.org/10.3828/mlo.v0i0.382 (zugegriffen: 18. Juli 2023).

König, Mareike. 2015. Herausforderung für unsere Wissenschaftskultur: Weblogs in den Geisteswissenschaften. In *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität*, hg. von Wolfgang Schmale, 91:57–74. Historische Mitteilungen – Beihefte. Franz Steiner Verlag. https://doi.org/10.25162/9783515111508 (zugegriffen: 18. Juli 2023).

Neuber, Frederike und Patrick Sahle. 2022. "Nach den Büchern: Rezensionen digitaler Forschungsressourcen". *H-Soz-Kult* (Forum: Rez.). www.hsozkult.de/debate/id/fddebate-132457 (zugegriffen: 18. Juli 2023).

Nyhan, Julianne. 2020. "7. The Evaluation and Peer Review of Digital Scholarship in the Humanities: Experiences, Discussions, and Histories". In *Digital Technology and the Practices of Humanities Research*, hg. von Jennifer Edmond, 163–82. Open Book Publishers. https://doi.org/10.11647/obp.0192.07 (zugegriffen: 18. Juli 2023).

Piotrowski, Michael. 2020. „Ain't no way around it: Why we need to be clear about what we mean by ‚Digital Humanities'". SocArXiv. https://doi.org/10.31235/osf.io/d2kb6 (zugegriffen: 18.07.2023).

Del Rio Riande, Gimena, Gabriel Calarco, Gabriela Striker und Romina De León. 2018. Humanidades Digitales: Construcciones locales en contextos globales. Buenos Aires: Universidad de Buenos Aires. http://eprints.rclis.org/32718/1/Actas_Humanidades%20Digitales_Construcciones%20locales%20en%20contextos%20Globales_all.pdf (zugegriffen: 18. Juli 2023).

Tóth-Czifra, Erzsébet, Ulrike Wuttke, Alíz Horváth und Christopher Nunn. 2023. Amplifying unheard voices in Digital Humanities: an OpenMethods edit-a-thon. In Digital Humanities 2023: Book of Abstracts. Digital Humanities 2023. Collaboration as Opportunity. (DH2023), Graz, Austria, hg. von Anne Baillot, Toma Tasovac, Walter Scholger, und Georg Vogeler, 26–27. Graz: Zenodo. https://doi.org/10.5281/zenodo.7961822 (zugegriffen: 18. Juli 2023).

Explorationen unbekannter Korpora mit Topic Modeling und manueller Annotation. Zusammenarbeiten von Mensch und Maschine revisited

Franken, Lina
lina.franken@uni-vechta.de
Universität Vechta, Deutschland
ORCID: 0000-0002-2587-4068

Dennis, Möbus
dennis.moebus@fernuni-hagen.de
FernUniversität Hagen, Deutschland
ORCID: 0009-0008-9064-7460

Fragestellung

Computationelle Verfahren können in den seltensten Fällen erkenntnisreich sein, wenn sie nicht mit einer manuellen, respektive qualitativen Untersuchung des Korpus einhergehen. In der Regel werden zu diesem Zweck Mensch-Maschine-Interaktionen umgesetzt. Doch wie verändern sich Erkenntnisse und Forschungsprozesse durch diese Erweiterungen? Noch immer ist zu wenig evaluiert, welche Konsequenzen die Anwendung von Verfahren der Digital Humanities für Forschungsergebnisse haben (Franken/Möbus 2023 in Review). Im Workshop setzen wir an dieser Stelle an und untersuchen gemeinsam, aufbauend auf bisherigen Studien der Workshopleitenden, wo und wie sich hermeneutische Erkenntnisproduktion verändert. Hierfür wird Topic Modeling exemplarisch herangezogen, weil dies als unüberwachtes maschinelles Lernen besonders gut geeignet ist, um komplexe und bisher kaum erschlossene Korpora zu untersuchen (Andorfer 2017; Philipps 2018). Es wird kombiniert mit manuellen Annotationen in Tradition der Grounded Theory (Charmaz 2014; Franken/Koch/Zinsmeister 2020), die im Workshop mit dem Annotationstool Catma umgesetzt werden (Gius et al. 2023). Workshopteilnehmende lernen nicht nur das Zusammenspiel von computationellen und manuellen Arbeitsschritten kennen, sondern gestalten dies selbst und diskutieren die epistemologischen Konsequenzen. Die Erkenntnisse ebenso wie das konkrete Vorgehen sind im Anschluss an den Workshop auf eigene Korpora übertragbar.

Hintergrund der Workshopreihe

Ausgehend von den auf der DHd 2023 diskutierten Ergebnissen eines Workshops zur Gegenüberstellung manueller Annotation und durch maschinelles Lernen erzeugter Annotation (Egger et al. 2023; Franken/Möbus 2023 in Review) möchten wir einen wesentlich überarbeiteten Workshop anbieten.

Der erste Workshop, der im Juni 2022 digital als Kooperation zwischen der FernUniversität in Hagen (Möbus) und der LMU München (Franken) durchgeführt wurde, orientierte sich grob am Aufbau eines Turing Tests und diente der Beantwortung der Frage, wie angesichts der rapide steigenden digitalen Verfügbarkeit lebensgeschichtlicher Interviews (etwa im Rahmen von Oral-History.Digital[1]) computationelle und traditionelle Auswertungsverfahren ineinandergreifen können. Die Teilnehmenden verschiedener Disziplinen und Professionalisierungsgrade wurden in zwei Gruppen unterteilt, von denen eine zunächst über Topic Modeling, die andere über manuelle Annotation in ein digital aufbereitetes Korpus lebensgeschichtlicher Interviews eingestiegen ist (Lebensgeschichte und Sozialkultur im Ruhrgebiet[2])
. Zur besseren Betreuung und ethnografischen Dokumentation wurden diese beiden Gruppen nochmal in jeweils zwei Teilgruppen geteilt. Mit diesem Aufbau konnten wir erheben, in welcher Reihenfolge den Teilnehmenden ein besserer Einstieg in die völlig unbekannten Daten gelang. Durch teilnehmende Beobachtung und eine nachträgliche Befragung haben wir wichtige Erkenntnisse gesammelt, um sowohl eine optimierte Topic-Modeling-Pipeline als auch einen neu konzipierten explorativen Workshop zu entwickeln.

Ziele des Workshops

Im Rahmen des Workshops möchten wir unsere Ergebnisse kurz vorstellen und diese weiterentwickeln. Dabei erproben die Workshopteilnehmenden gemeinsam, wie Zusammenarbeiten von Mensch und Maschine in der Forschungspraxis aussehen. Ziel ist es mittelfristig, Empfehlungen für den Einsatz computationeller Verfahren in der qualitativen Forschung aussprechen zu können.

Die aufgeworfenen Fragen wurden bisher mit einem heterogenen Teilnehmendenkreis diskutiert. Der Workshop im Rahmen der DHd 2024 richtet sich hingegen explizit an die DH-Community und damit an Forschende, deren Blick auf die sich verändernden Epistemologien bereits geschult ist. Hierfür muss kein explizites Vorwissen vorhanden sein, aber entsprechende Fragen tauchen doch häufig auf, wenn DH-Forschungen realisiert werden. Im Workshop überprüfen wir gemeinsam, wer in diesen Settings Sinn konstruiert und welche Unterschiede sich ergeben.

Aufgrund der Erfahrungen im ersten Workshop wird der Workshop so aufgebaut, dass ein offenes Explorieren sowohl in der Gruppe als auch in Einzelarbeit möglich ist.

Wir konnten feststellen, dass Gruppenkonstellationen großen Einfluss auf die Wahrnehmung der Methoden haben. Um die Wahrnehmungen einzelner Disziplinen und unterschiedliche Wissensbestände besser kondensieren zu können, überlassen wir es den Teilnehmenden, sich bei der Zusammensetzung der Gruppen selbst zu organisieren und an gemeinsamen Interessen zu orientieren.

Entsprechend werden, anders als beim ersten Workshop, weder Textstellen noch eine thematische Fragestellung vorgegeben, um das offene Explorieren nicht einzuschränken. Stattdessen stehen die Forschungsinteressen der Teilnehmenden im Mittelpunkt, um eine realitätsnahe Arbeitsweise und keine künstliche Laborsituation untersuchen zu können.

Als Ergebnis unseres ersten Workshops hat sich ergeben, dass das Topic Modeling einen optimalen Einstiegspunkt in die Exploration unbekannter Daten bietet und der manuellen Annotation vorgeschaltet werden sollte. Allerdings kann ein iteratives Verfahren qualitative Analysen sogleich miteinbeziehen. Durch entsprechende Funktionalitäten in der im Workshop verwendeten Topic-Modeling-Pipeline kann zwischen dem Distant-Reading des Korpus und der Ansicht konkreter Textstellen zur inhaltlichen Begutachtung on-the-fly umgeschaltet werden. Das ermöglicht die Suche nach interessanten Phänomenen im Rahmen einer Annäherung von Topic Modeling und theoretischem Sampling nach Grounded Theory.

Schließlich werden, ausgehend von der konstruktiven Kritik am ersten Workshop, für das manuelle Annotieren keine Textlängen oder Formen der Verschlagwortung vorgegeben, um dem in der Grounded Theory üblichen flexiblen Vorgehen zu entsprechen. Wir entwickeln ein Verfahren, um die statisch vergebenen Topics mit den dynamisch vergebenen Annotationen in der nachträglichen Auswertung der Workshopergebnisse ins Verhältnis zu setzen.

Format und Zielpublikum

Der Workshop ist als Hands-On-Workshop konzipiert. Der klassische Bestand Lebensgeschichte und Sozialkultur im Ruhrgebiet des Archivs “Deutsches Gedächtnis” wird in einem CoLab mit einem vortrainierten Topic Model exploriert. Nach einer kurzen Einführung durch die Workshopleitenden und der Vorstellung der Workshopteilnehmenden beschäftigen wir uns am ersten Tag mit der Korpuserkundung durch Topic Modeling. Am zweiten Tag steigen wir davon ausgehend in die manuelle Annotation von interessanten Textpassagen ein, wobei ein iteratives Hin- und Herspringen bei Bedarf bereits am ersten Tag umgesetzt wird. Dabei gilt es, fortlaufend unsere Erkenntnissprozesse zu reflektieren. Hierfür werden Methoden kollaborativen Forschens (Bieler et al. 2021; Fortun et al. 2021) und Methoden des Walkthroughs (Light et al. 2013, Amelang 2023) genutzt.

Das explorative Setting des Workshops dient der weiteren forschenden Erkundung des Zusammenwirkens von Mensch und Maschine, so dass während des Workshops teilnehmende Beobachtungen durch die Workshopleitenden stattfinden, die für künftige Untersuchungen verwendet werden. Im Anschluss an den Workshop sollen die Ergebnisse durch vertiefende Interviews als Walk-Throughs durch die getätigten Annotationen und durch das CoLab reflektiert werden. Die Bereitschaft zur Teilnahme an einem solchen Interview ist wünschenswert, aber keine Voraussetzung.

Es ist kein technisches Vorwissen für die Teilnahme notwendig.

Beitragende zum Workshop

Dennis Möbus ist Wissenschaftlicher Mitarbeiter am Institut für Geschichte und Biographie der FernUniversität in Hagen. Er koordiniert die Forschungsgruppe digital humanities - Forschen im digitalen Raum und ist im Rahmen von Oral-History.Digital für den Aufbau des Pilotarchivs Archiv “Deutsches Gedächtnis” online sowie die automatische Vorverarbeitung und Erschließung lebensgeschichtlicher Interviews zuständig. Seine Forschungsinteressen liegen in den Bereichen Neuere/Neueste und Zeitgeschichte, Oral History und Erfahrungsgeschichte sowie Text Mining für historische Quellen.

Lina Franken ist Universitätsprofessorin für Digital Humanities an der Universität Vechta. Ihre Forschungsinteressen liegen in der Weiterentwicklung computationeller Verfahren für qualitative Forschung und in der Beforschung epistemologischer Veränderungen in den und mit den DH anhand (diskurs)ethnografischer Methoden. In den Critical Code und Data Studies im Kontext der Science and Technology Studies forscht sie zu Algorithmen im Alltag und deren Einbezug in ethnografische Forschungen. Infrastrukturentwicklung für qualitativ-ethnografische Forschungen setzt u.a. sie als Mitglied der Design Group der Plattform for Experimental Collaborative Ethnography (PECE) um.

Benötigte technische Ausstattung

Workshopteilnehmende müssen ein eigenes Gerät mitbringen, wir benötigen Internetzugang. Im Vorfeld des Workshops werden den Teilnehmenden sowohl ein CoLab als auch der Zugang zu einer Catma-Gruppe zur Verfügung gestellt, es wird ein Google- sowie ein Catma-Account benötigt. Vor Ort werden ein Beamer sowie eine flexibel anpassbare Kombination aus Tischen und Stühlen benötigt, hilfreich wäre zudem ein Whiteboard o.ä.

Fußnoten

1. https://www.oral-history.digital/
2. https://www.fernuni-hagen.de/geschichteundbiographie/forschung/projekte/lusir-online.shtml

Bibliographie

Amelang, Katrin. 2023. „Wie Apps erforschen? Zum Zusammentreffen neuer Forschungsgegenstände und alter Methoden“. In: *Hamburger Journal für Kulturanthropologie 16*. https://journals.sub.uni-hamburg.de/hjk/article/view/2073.

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.“ In: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2017_002.

Bieler, Patrick/Bister, Milena D./Hauer, Janine/ Klausner, Martina/Niewöhner, Jörg/Schmid, Christine/ von Peter, Sebastian. 2021. „Distributing Reflexivity through Co-laborative Ethnography“. In: *Journal of Contemporary Ethnography 50*. 77–98.

Charmaz, Kathy. 2014. *Constructing Grounded Theory. Introducing Qualitative Methods*. 2. Aufl. Los Angeles u.a.

Egger, Nils; Franken, Lina; Möbus, Dennis; Schmid, Florian. 2023. „Oral History auf dem Weg zu Big Data: menschliche und maschinelle Annotation lebensgeschichtlicher Interviews im Vergleich“. In: *Digital Humanities im deutschsprachigen Raum (DHd) 2023. Book of Abstracts*. https://zenodo.org/record/7715317.

Fortun, Mike/Poirier, Lindsay/Morgan, Alli/ Callahan, Brian/Fortun, Kim. 2021. “What's So Funny about PECE, TAF and Data Sharing?” *Collaborative Anthropology Today. A Collection of Exceptions*, hg. Von Dominic Boyer und George E. Marcus. Ithaca.

Franken, Lina; Möbus, Dennis. 2023 (in Review). „Mensch und Maschine als Team. Exploratives Topic Modeling und manuelle Annotation in der qualitativen Sozialforschung“. In: *Zeitschrift für digitale Geisteswissenschaften*.

Franken, Lina; Koch, Gertraud; Zinsmeister, Heike. 2020. „Annotationen als Instrument der Strukturierung“. *Annotations in Scholarly Editions and Research. Functions, Differentiation, Systematization*, hg. von Julia Nantke und Frederik Schlupkothen, Berlin/München/Boston. 89–108. https://doi.org/10.1515/9783110689112-005.

Light, Ben/Burgess, Jean/Duguay, Stefanie. 2017. “The Walkthrough Method. An Approach to the Study of Apps”. In: *new media & society 20/3*. 881–900. DOI: 10.1177/1461444816675438.

Philipps, Axel. 2018. „Text Mining-Verfahren als Herausforderung für die rekonstruktive Sozialforschung“. In: *Sozialer Sinn 19/2*. 367–387.

Generative KI, LLMs und GPT bei digitalen Editionen

Czmiel, Alexander
czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Dumont, Stefan
dumont@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0002-6923-0950

Fischer, Franz
franz.fischer@unive.it
Venice Centre for Digital and Public Humanities, Ca‘ Foscari Università Venezia, Italien

Pollin, Christopher
christopher.pollin@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz, Österreich
ORCID: 0000-0002-4879-129X

Sahle, Patrick
sahle@uni-wuppertal.de
Lehrstuhl für Digital Humanities, Bergische Universität Wuppertal, Deutschland
ORCID: 0000-0002-8648-2033

Schaßan, Torsten
schassan@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0002-8902-4775

Scholger, Martina
martina.scholger@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz, Österreich
ORCID: 0000-0003-1438-3236

Vogeler, Georg
georg.vogeler@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz, Österreich
ORCID: 0000-0002-1726-1712

Roeder, Torsten
dh@torstenroeder.de
Zentrum für Philologie und Digitalität, Universität Würzburg, Deutschland
ORCID: 0000-0001-7043-7820

Fritze, Christiane
christiane.fritze@wienbibliothek.at
Wienbibliothek, Österreich
ORCID: 0000-0001-5099-8970

Henny-Krahmer, Ulrike
ulrike.henny-krahmer@uni-rostock.de
Institut für Germanistik, Universität Rostock, Deutschland
ORCID: 0000-0003-2852-065X

Kurzzusammenfassung

Dieser Workshop konzentriert sich auf die Erforschung der Anwendungsmöglichkeiten und Herausforderungen von KI-basierten Anwendungen wie GPT und Large Language Models (LLMs) im Kontext digitaler Editionen. GPT-4, mindestens bis Juli 2023 das führende Modell, bietet erhebliche Potenziale, z.B. für die Umwandlung von unstrukturiertem Text in strukturierte Daten und die Erkennung von benannten Entitäten. Dennoch liefert es bislang noch unbefriedigende Ergebnisse, weshalb sorgfältige Überwachung und Feedbacksysteme unerlässlich sind. Die Integration von LLMs in Arbeitsabläufe und Webentwicklungsprojekte ist vielversprechend, erfordert jedoch noch konzeptionelle und dann auch technische Vorstudien. In Anbetracht der rasanten KI- und LLM-Entwicklungen lädt der Workshop dazu ein, zu experimentieren und Strategien für den effektiven Einsatz dieser Modelle in digitalen Editionsprojekten zu diskutieren.

Generative KI, LLMs und GPT bei digitalen Editionen

KI-basierte Anwendungen wie Generative Pre-trained Transformer (GPT) sind revolutionär im Umgang mit Text. Ihre Stärke liegt in der Dialogfähigkeit, der natürlichsprachigen Textgenerierung, der Einbeziehung von Kontext, der Nutzung als semantischer “Wissensbasis”, der Mehrsprachigkeit und der Möglichkeit der Feinabstimmung auf spezifische Aufgaben *(Chen et al. 2023, 1–2)* . Außerdem können sie für die Entwicklung von Algorithmen zur Verarbeitung von Information eingesetzt werden. Zum Zeitpunkt der Erstellung dieses Abstracts (Stand Juli 2023) ist GPT-4 *(OpenAI 2023)* das leistungsfähigste Werkzeug. Zum Zeitpunkt des Workshops, im März 2024, wird GPT-4 seit einem Jahr eingeführt und seitdem für die verschiedensten Einsatzgebiete getestet worden sein. Vielleicht werden bis dahin auch alternative Modelle und Szenarien im KI-Kontext verwendet.

Bei der Erschließung von historischen Dokumenten im digitalen Paradigma richten sich Editionen auf Texte, deren Inhalte und Strukturen etwa durch Markup-Sprachen aufbereitet werden. Aber auch Planung, Modellierung und Umsetzung digitaler Editionen manifestieren sich üblicherweise in textueller Form, z.B. in User Stories, Code für Daten und Algorithmen oder Dokumentationen. Sind generative Sprachmodelle dann das ultimative Werkzeug zur Umsetzung digitaler Editionen?

Diese Frage lässt sich aufgrund der rasanten Entwicklung in diesem Bereich noch nicht eindeutig beantworten. Wir befinden uns in einer Übergangsphase, in der die Möglichkeiten und Potenziale, aber auch die Herausforderungen, Grenzen und Risiken dieser neuen Technologie noch ausgelotet werden müssen. Dazu sollten Experimente durchgeführt werden, bei denen typische Aufgaben in Editionsprojekten gepromptet, getestet, angewandt, überprüft, eingeordnet und reflektiert werden. Mit diesem Ziel vor Augen möchten wir einen Aufruf zu Experimenten (Call for Experiments) veröffentlichen, um in einem Workshop zusammen mit der Fachgemeinschaft Versuche im Zusammenhang mit Large Language Models (LLMs) *(Zhao et al. 2023)* , GPT oder alternativen Modellen und digitalen Editionen zu erörtern. Folgende Themenbereiche scheinen aktuell von Interesse zu sein. AI-unterstütze Experimente müssen sich aber nicht auf diese beschränken, sondern können sich auf jegliche Aspekte im Zusammenhang mit digitalen Editionen beziehen.

- **Überführung von unstrukturiertem Text (Transkription) in strukturierten Text (Markup)**
  Unstrukturierte oder semistrukturierte Texte lassen sich derzeit besonders effizient durch LLMs in strukturierte Daten umwandeln. Beispielsweise kann GPT-4 zur automatischen Annotation, Klassifikation, Erzeugung von Textstrukturen, sowie zur Erstellung von Zusammenfassungen und Abstracts eingesetzt werden. Dies ermöglicht die Implementierung von Workflows, in denen Texte halbautomatisch in Interaktion eines Menschen mit einem LLM oder sogar vollautomatisch in einer Skript-Pipeline nach TEI-XML oder anderen Metadatenstandards und Auszeichnungssprachen konvertiert werden können.
- **Überführung von strukturiertem Text in explizite Datenstrukturen**
  Die überwiegende Mehrheit wissenschaftlicher Editionen liegt in gedruckten Ausgaben vor. Das alte Ziel, diese Wissensschätze in standardkonform strukturierte Daten zu überführen, die z.B. nach dem TEI-Standard kodiert sind, besteht weiterhin. Viele Editionen sind inzwischen bilddigitalisiert und stehen oft mit OCR-Daten (z.B. PAGE XML) oder in PDFs zur Verfügung. Der weitere Prozess der Digitalisierung müsste aus einer Mischung regelgeleiteter “Lesung” der typografischen Strukturen und der Berücksichtigung der Un-

schärfe in der Einhaltung der typografischen Regeln (insbesondere zur Trennung von Text und Paratext) sowie der Bereinigung von OCR-Fehlern bestehen.

- **Named Entity Recognition, Normalisierung und Anreicherung**
  Eine weitere potentielle Anwendung ist die Erkennung, Extraktion und Annotatierung von Entitäten wie Personen, Organisationen, Orten und Werken, deren Identifizierung und Normalisierung ebenso wie die historischer Wert- und Datumsangaben. In diesem Zusammenhang erscheint auch die Verknüpfung mit Normdaten (Named Entity Linking, NEL) interessant, wie die Gegenüberstellung zu anderen Tools wie OpenRefine, oder die Einbindung von Reconciliation.
- **Kontextspezifische Annotationen**
  LLMs und insbesondere GPT-4 zeichnen sich durch ihre Fähigkeit aus, Texte zu analysieren und auf Kontextwissen zuzugreifen, wodurch semantische Strukturen in Texten (semi-)automatisch annotiert werden können. Diese annotierten, semantischen Schichten, auch als "assertive Schichten" *(Vogeler 2019)* bezeichnet, können verschiedene Aspekte wie Transaktionen, Kommunikationsprozesse oder Rechtsmittel in historischen Quellen, sowie geografische Phänomene wie Reise- und Transportinformationen umfassen. GPT-4 ist in der Lage, z.B. geografischen Kontext in die Generierung von Antworten einzubeziehen, einschließlich von Informationen über Ländergrenzen, Flussverläufe und Gebirgsketten *(Roberts et al. 2023)* . Könnte GPT-4 basierend auf einer historischen Quelle einen Kommentar generieren, der geografischen Kontext nutzt, um die Machbarkeit einer in der Quelle erwähnten Reiseroute zu bewerten?
- **Fehlermanagement und Datenkontrolle**
  Wie jedes generative KI-Modell erzeugt auch GPT-4 häufig unbefriedigende Ergebnisse. Diese können durch fehlende oder tendenziöse Trainingsdaten, schlechtes Prompting, "Halluzinationen", fehlerhaftes Reasoning, übermäßig komplexe Aufgaben, eine begrenzte Anzahl von Ein- und Ausgabetokens oder eingeschränkte "Aufmerksamkeitsspannen" entstehen. Es ist von entscheidender Bedeutung, die erzeugten Daten gründlich zu prüfen und Feedback sowie redaktionelle Entscheidungen miteinzubeziehen, wie beispielsweise Studien aus dem medizinischen Bereich nahelegen *(Nori et al. 2023)* . Ggf. ließe sich auch GPT-4 selbst zur Validierung der eigenen Resultate einsetzen.
- **Integration in Arbeitsabläufe**
  Workflows *(Møller et al. 2023)* können iterativ gestaltet sein, wobei mehrere Prompts nacheinander eingesetzt werden, um verschiedene Aspekte der Annotation, Normalisierung oder des Fehlermanagements zu übernehmen. Ein erster Prompt erstellt z.B. die Textstruktur, ein zweiter verwendet die Ergebnisse des ersten Prompts für die Named Entity Annotation, ein dritter annotiert semantische Phänomene basierend auf den Ergebnissen des zweiten Prompts und der vierte führt eine automatische Datenqualitätsprüfung durch, basierend auf "Ground Truth" TEI-XML-Dateien. Zwischen den Iterationen könnte ein:e Editor:in involviert sein, der bzw. die Fehler korrigiert und Feedback gibt. Allerdings sind hierfür geeignete Infrastrukturen und Softwarelösungen erforderlich, die derzeit noch nicht existieren.
- **Planung, Konzeption und Evaluierung digitaler Editionen**
  Vor der Entwicklung eines digitalen Editionsprojekts ist es wichtig, die Anforderungen sowie die Forschungsfragen genau zu verstehen. Dies beinhaltet die Sammlung und Analyse von Anforderungen, User Stories von Stakeholdern sowie Kontextwissen von Expert:innen.
- **Webentwicklung und Benutzeroberflächen**
  LLMs können auch in Webanwendungen integriert werden, um dynamische und interaktive Inhalte zu erstellen. Dies kann z.B. durch APIs geschehen, die es Entwicklern ermöglichen, auf die Fähigkeiten von z.B. GPT-4 zuzugreifen und diese in ihre Webanwendungen zu integrieren. Darüber hinaus unterstützen generative Sprachmodelle die Implementierung von digitalen Editionen, da sie sehr effizient im Schreiben von Quellcode sind.
- Fine Tuning **,** Prompt Tuning **,** Prompt Engineering **und Vektordatenbanken**
  Fine Tuning ist ein Prozess, bei dem ein vortrainiertes Modell weiter trainiert wird, um es für eine spezifische Aufgabe zu optimieren. Prompt Tuning *(Wu et al. 2023)* hingegen steuert ein Sprachmodell durch gezielte Auswahl und Gestaltung von Eingabeaufforderungen (Prompts), ohne das Modell selbst zu ändern. Prompt Engineering *(White et al. 2023)* bezieht sich auf die Entwicklung effektiver Eingabeaufforderungen für Sprachmodelle. Vektordatenbanken *(Windsor and Choi 2023)* können zur Speicherung von Embeddings genutzt werden, die bei der Feinabstimmung, dem Prompt Tuning und dem Prompt Engineering hilfreich sind. Sie erlauben es, eine Sammlung von effektiven Prompts zu speichern und zu verwalten, wobei die Speicherung als Vektoren einen schnellen Zugriff auf ähnliche oder verwandte Prompts ermöglicht. Experimente könnten zeigen, wie und für welche Aufgaben diese Methoden im Bereich digitaler Editionen angewendet werden können.

Im Workshop sollen konkrete Anwendungsfälle vorgestellt werden, in denen GPT-4 oder andere LL-Modelle für die oben genannten Szenarien auf der Basis unterschiedlichen Quellenmaterials experimentell erprobt und im Hinblick auf Potenziale, Grenzen und Probleme sowie ethische *(Baktash and Dawodi 2023)* und theoretische Implikationen diskutiert werden. In einem weiteren Schritt soll auf Basis der Erkenntnisse aus den Experimenten reflektiert werden, welchen Einfluss diese Technologien auf die digitalen Editionen der Zukunft nehmen können. Dabei sollte GPT-4 nicht allein betrachtet werden, sondern insbesondere andere zukünftige Open Source-Alternativen wie bei-

spielsweise Orca *(Mukherjee et al. 2023)* berücksichtigt werden. Diskutiert werden sollen aber auch die immer noch bestehenden Grenzen und zu überwindenden Hürden, wie z.B. die Differenz zwischen einem rein auf Zeichen und Tokens aufbauenden Textverständnis und der Visualität handschriftlicher oder typografischer Dokumente, die eigentlich ein "bildliches" Lesen erfordern würden.

Das Format des Workshops, der acht Stunden (zwei Tageshälften) dauert, besteht aus der Vorstellung von Experimenten und einer kritischen Diskussion. Die Experimente werden in vier Blöcken zusammengefasst, in denen nach der Vorstellung der Experimente mindestens 30 Minuten für die Diskussion und Ergebnissicherung reserviert sind. Die Ergebnisse der Blöcke des ersten halben Tages werden in einer Zusammenfassung nach der Mittagspause sowie in einer abschließenden Runde gesammelt. Das Ergebnis des Workshops soll ein Experimentbericht sein, der zur Veröffentlichung zeitnah als Blogbeitrag und anschließend in einer einschlägigen Fachzeitschrift (ZfdG, DHQ, magazén) vorgesehen ist. Die Teilnehmer:innen setzen sich aus Beitragenden (Einzelpersonen oder Gruppen, die Experimente präsentieren), aktiv mitdiskutierenden Teilnehmer:innen und den Workshopleiter:innen aus dem Kreis der Einreichenden zusammen.

Zeitplan

- 09:00 - 09:30 Uhr: Einführung
- 09:30 - 10:45 Uhr: Block 1 - Experimentvorstellung und Diskussion
- 10:45 - 11:00 Uhr: Pause
  11:00 - 12:15 Uhr: Block 2 - Experimentvorstellung und Diskussion
- 12:15 - 12:45 Uhr: Mittagspause (30 Minuten)
- 12:45 - 13:15 Uhr: Wrap-Up
- 13:15 - 14:30 Uhr: Block 3 - Experimentvorstellung und Diskussion
- 14:30 - 14:45 Uhr: Pause (15 Minuten)
- 14:45 - 16:00 Uhr: Block 4 - Experimentvorstellung und Diskussion
- 16:00 - 17:00 Uhr: Abschlussdiskussion und Festhalten von Ergebnisse

Workshop

- Format: Diskussionen anhand von präsentierten Experimenten
- Dauer: zwei halbe Tage
- Zielgruppe: Forscher:innen, Entwickler:innen und Fachleute im Bereich digitaler Editionen
- Maximale Teilnehmer:innenzahl: 20
- Erforderliche technische Ausstattung: Beamer, Computer, Internetzugang

Call for Experiments: Generative KI, LLMs und GPT bei digitalen Editionen

Im Zuge unseres bevorstehenden Workshops "Generative KI, LLMs und GPT bei digitalen Editionen" auf der Dhd24, lädt das Institut für Dokumentologie und Editorik (IDE) zur Einreichung experimenteller Beiträge ein, die sich mit den Anwendungsmöglichkeiten und Herausforderungen KI-basierter Tools wie GPT und Large Language Models (LLMs) in digitalen Editionen auseinandersetzen. In diesem Workshop sollen konkrete Anwendungsfälle präsentiert und diskutiert werden, in denen GPT-4 oder alternative LL-Modelle auf Basis unterschiedlichen Quellenmaterials für die vorgegebenen Szenarien experimentell erprobt und hinsichtlich ihrer Potenziale, Grenzen, Probleme sowie ethischen und theoretischen Implikationen diskutiert werden, um darauf aufbauend anhand der gewonnenen Erkenntnisse zu erörtern, welchen Einfluss diese Technologien auf die digitalen Editionen der Zukunft haben könnten.

Wir starten daher einen Call for AI-Experiments zu jeglichen Themenbereichen die im Zusammenhang mit digitalen Editionen stehen, einschließlich (aber nicht beschränkt auf):

- Überführung von unstrukturiertem Text (Transkription) in strukturierten Text (Markup)
- Überführung von strukturiertem Text in explizite Datenstrukturen
- Named Entity Recognition, Normalisierung und Anreicherung
- Kontextspezifische Annotationen
- Fehlermanagement und Datenkontrolle
- Integration in Arbeitsabläufe
- Planung, Konzeption und Evaluierung digitaler Editionen
- Webentwicklung und Benutzeroberflächen
- Fine Tuning, Prompt Tuning, Prompt Engineering und Vektordatenbanken

Um einen Beitrag einzureichen, senden Sie bitte eine halbseitige Skizze und Link zum Chat-Protokoll (wenn möglich) an ki@i-d-e.de bis zum 14. Januar 2024. Die Vorschläge werden von den Organisator:innen (Mitgliedern des IDE) des Workshops begutachtet, die Benachrichtigung über die Annahme erfolgt bis zum 26. Januar 2024.

Bibliographie

Baktash, Jawid Ahmad, and Mursal Dawodi. 2023. 'Gpt-4: A Review on Advancements and Opportunities in Natural Language Processing'. arXiv. https://doi.org/10.48550/arXiv.2305.03195.

Chen, Zhutian, Chenyang Zhang, Qianwen Wang, Jakob Troidl, Simon Warchol, Johanna Beyer, Nils Gehlenborg, and Hanspeter Pfister. 2023. 'Beyond Generating Code: Evaluating GPT on a Data Visualization Course'. arXiv. https://doi.org/10.48550/arXiv.2306.02914.

Møller, Anders Giovanni, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. 'Is a

Prompt and a Few Samples All You Need? Using GPT-4 for Data Augmentation in Low-Resource Classification Tasks'. arXiv. https://doi.org/10.48550/arXiv.2304.13861.

Mukherjee, Subhabrata, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. 'Orca: Progressive Learning from Complex Explanation Traces of GPT-4'. arXiv. https://doi.org/10.48550/arXiv.2306.02707.

Nori, Harsha, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. 'Capabilities of GPT-4 on Medical Challenge Problems'. arXiv. https://doi.org/10.48550/arXiv.2303.13375.

OpenAI. 2023. 'GPT-4 Technical Report'. arXiv https://doi.org/10.48550/ARXIV.2303.08774.

Roberts, Jonathan, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. 'GPT4GEO: How a Language Model Sees the World's Geography'. arXiv. http://arxiv.org/abs/2306.00020.

Vogeler, Georg. 2019. 'The "Assertive Edition"'. In International Journal of Digital Humanities 1 (2), 309–22 https://doi.org/10.1007/s42803-019-00025-5.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. 'A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT'. arXiv. https://doi.org/10.48550/arXiv.2302.11382.

Windsor, Brad, and Kevin Choi. 2023. 'Thistle: A Vector Database in Rust'. arXiv. https://doi.org/10.48550/arXiv.2303.16780.

Wu, Junda, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. 2023. 'InfoPrompt: Information-Theoretic Soft Prompt Tuning for Natural Language Understanding'. arXiv. https://doi.org/10.48550/arXiv.2306.04933.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2023. 'A Survey of Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2303.18223.

Hands-on-Workshop DNBLab

Taube, Anke

a.taube@dnb.de
Deutsche Nationalbibliothek, Deutschland
ORCID: 0000-0003-4178-086X

Palek, Stephanie

s.palek@dnb.de
Deutsche Nationalbibliothek, Deutschland
ORCID: 0000-0003-1624-896X

Einführung:

Der Workshop bietet einen praktischen Einstieg in den Bezug und die Analyse der Daten und freien digitalen Objekte der Deutschen Nationalbibliothek (DNB). Neben einem Überblick über das vielseitige Datenangebot der DNB ist das Ziel des Hands-on-Workshops, den Teilnehmer*innen eine niedrigschwellige praktische Einführung in automatisierte Abfragen über die verschiedenen offenen Bezugswege sowie geeignete Datenformate zu geben. Die Teilnehmer*innen des Workshops erhalten einen Einblick in das DNBLab als Service für Wissenschaft und Forschung sowie das Arbeiten mit dem Bestand und den Normdaten der DNB. Durch das gemeinsame Bearbeiten einer exemplarischen Fragestellung werden Grundlagen für die Entwicklung weiterer Forschungsideen geschaffen.

Beitragende:

Bei den Veranstaltenden handelt es sich um zwei Mitarbeiterinnen der Deutschen Nationalbibliothek, die als Teil des DNBLab-Teams seit dem experimentellen Start im Sommer 2020 das DNBLab als zentralen Anlaufpunkt für den Zugang zu Metadaten und digitalen Ressourcen (Taube 2021) betreuen und weiterentwickeln. Anke Taube ist Bibliothekarin und Informationswirtin und arbeitet seit 2016 im Bereich Digitale Bereitstellung der Abteilung Digitale Dienste. Stephanie Palek ist Historikerin mit informationstechnischem Hintergrund und arbeitet seit 2020 als Referentin für Digitale Dienste an der DNB. Ihre Schwerpunkte liegen dabei in den Bereichen Persistent Identifier, Visualisierung sowie Angebote und Services für Wissenschaft und Forschung.

Zielsetzung:

Die Deutsche Nationalbibliothek (DNB) ist die zentrale Archivbibliothek Deutschlands. Sie sammelt, dokumentiert und archiviert alle Medienwerke in Schrift, Bild und Ton, die seit 1913 in und über Deutschland oder in deutscher Sprache veröffentlicht werden. Ob Bücher, Zeitschriften, CDs, Schallplatten, Karten oder Online-Publikationen – sie sammelt ohne Wertung, im Original und lückenlos. Die DNB bietet ihre über 50 Mio. Metadatensätze sowie viele digitale Objektsammlungen frei zugänglich an. Die umfassende Sammlung auf Basis verlegerischer Pflichtabgaben lädt zum Forschen und Analysieren ein. Neben der Sammlung physischer Medienwerke mit mehr als 31 Millionen Objekten und Metadaten stellt die DNB für Wissenschaft und Forschung einen schnell wachsenden digitalen Bestand mit mittlerweile über 12 Millionen E-Books, E-Journals, E-Paper und anderen digitalen und digitalisierten Werken für Text und Data Mining zur Verfügung. So umfasst der seit 2010 gesammelte E-Paper-Bestand alle in Deutschland digital erschienenen Ta-

ges- und Wochenzeitungen mit mehr als 3,6 Millionen Ausgaben. Forschende rücken als Bibliotheksnutzende immer mehr in den Fokus und möchten ihre Anforderungen im Bibliotheksservice der DNB berücksichtigt und gefördert wissen (Döhl 2019). Als zentraler Anlaufpunkt für die Präsentation, den Zugriff und die Nachnutzung der digitalen Ressourcen bietet das DNBLab verschiedene Zugänge zu Daten, frei verfügbaren Objektdateien und Volltexten für alle an Text und Data Mining Interessierte[1]. Für automatisierte Analysen von den zum großen Teil urheberrechtlich geschützten Ressourcen (Volltexten, Bildern oder AV-Medien) gibt es Kooperationsmöglichkeiten durch DH-Call und DH-Stipendien[2]. Nach den Regelungen des § 60d UrhG unterstützt die DNB aktiv Forschungsvorhaben durch die Bereitstellung von Metadaten, digitalen Objekten und Infrastrukturen sowie bei Bedarf auch Digitalisierung analoger Objekte (Döhl et al. 2020). Diese Angebote für Wissenschaft und Forschung werden kontinuierlich und so weitgehend wie möglich ausgebaut. Auch in Form von Workshops, Netzwerktreffen und Lehrkooperationen verstärkt die Deutsche Nationalbibliothek die Zusammenarbeit mit den Digital Humanities mit dem Ziel, Erfahrungen mit Expert*innen im Themenfeld auszutauschen, neue Wege der Zusammenarbeit einzuschlagen und das Potenzial der eigenen digitalen Datenbestände auszuloten. Verschiedene Veröffentlichungen und Anwendungen zeigen bereits die vielseitigen Nutzungsmöglichkeiten der digitalen Sammlungen der Deutschen Nationalbibliothek und inspirieren zu neuen Ideen und Analysen. So wurden neben der Untersuchung von Heftromanen im Vergleich zur Hochliteratur (Jannidis et al. 2019) die Nutzung von Pseudonymen in der Literatur oder auch die Wahrnehmung und Darstellung von Faschismus und Nationalsozialismus in der Presse- und Publikationslandschaft von 1933-1945 mittels Bestandsanalysen in der DNB erforscht[3]. Durch die Aufbereitung und Dokumentation im DNBLab werden die Daten und frei verfügbaren Objekte der DNB auch immer mehr von externen Nutzer*innen analysiert und visualisiert. So werden als eines der gesammelten Praxisbeispiele für das Arbeiten mit den DNB-Daten im wissenschaftlichen Stammbaum von BibSonomy auf Grundlage der DNB-Katalogdaten Promovierte den Betreuer*innen ihrer Doktorarbeit zugeordnet[4]. In der Arbeit „Liebe und Tod in der Nationalbibliothek“ wurden interessante Fakten über Romane im DNB-Katalog herausgearbeitet (Fischer et al. 2018), beispielsweise zu welchem Autor es die meisten Romane im Bestand gibt oder was die durchschnittliche Seitenzahl der Romane ist. Bei einer Fallstudie zur Geokartierung bibliografischer Übersetzungsdaten aus dem DNB-Katalog wurde im Rahmen einer Doktorarbeit aus den exportierten Übersetzungen ausgewählter Autoren ein Prototyp erstellt (Teichmann 2022).

Workshop-Format und Zielpublikum:

Das DNBLab bietet verschiedene Zugriffsmöglichkeiten für Wissenschaft und Forschung auf die Daten der Deutschen Nationalbibliothek. In dem halbtägigen Hands-on-Workshop werden die verschiedenen Zugänge zu den Metadaten und freien digitalen Objekten gezeigt und exemplarisch für eine gemeinsame Analyse aufbereitet. Nach einer kurzen Vorstellung des DNBLab-Angebots für Digital Humanities wird gemeinsam eine exemplarische Fragestellung mit Hilfe eines Online Coding Tutorials (Jupyter Notebook) in der Webumgebung Binder[5] oder JupyterLite[6] bearbeitet. Die Workshop-Teilnehmer*innen erhalten über das gemeinsame Durchführen des Praxisbeispiels eine Einführung in das maschinelle Abfragen sowie mögliche Analysen des vielseitigen Bestands der DNB.

Die im Rahmen des Workshops eingesetzten Python-Bibliotheken und Codeteile können ohne Vorkenntnisse ausgeführt und angepasst werden. Grundlegende Kenntnisse in der Programmiersprache Python und Jupyter Notebooks werden nicht vorausgesetzt, sind aber hilfreich und können bei Bedarf vorab über ein Online-Tutorial (Althage et a. 2022) oder die Teilnahme an einer der regelmäßigen virtuellen DNBLab-Einführungsveranstaltungen[7] erworben werden. Die Teilnehmer*innen erhalten ohne vorherigen Installationsaufwand einen Überblick über den digitalen Bestand und die Zugriffsmöglichkeiten auf die Metadaten, freien Bilder, Audiodateien und Volltexte der Deutschen Nationalbibliothek. Es wird eine neue exemplarische Fragestellung mit einem vorher noch nicht veröffentlichten Jupyter Notebook durchgespielt, die mit den frei verfügbaren Daten und digitalen Objekten aus dem Onlineangebot des DNBLabs beantwortet werden kann. Basis wird eine CQL-basierte Abfrage der SRU-Schnittstelle nach dem Vorkommen eines thematischen Stichwortes in den Titeln der DNB-Katalogdaten sowie im zweiten Schritt in den Volltexten der digitalisierten Inhaltsverzeichnisse sein. Die selektierten Daten werden für die weitere Bearbeitung normalisiert. Neben einer Worthäufigkeitsanalyse in den Titeldaten im zeitlichen Verlauf werden die Inhaltsverzeichnisse mittels Topic Modeling exemplarisch exploriert und visualisiert. Mit der Abfrage der SRU-Schnittstelle werden Grundlagen zur Bildung eigener Datensets vermittelt und ein beispielhaftes Datenset aufbereitet und analysiert. Durch den gemeinsamen Perspektivwechsel beim Heranziehen bibliothekarischer Formatdokumentationen aus Sicht der Forschenden werden dabei der Umgang mit Erschließungsfallstricken sowie bibliothekarische Unterstützungsangebote thematisiert. Am Ende werden die Ergebnisse für die maschinelle Weiterverarbeitung aufbereitet und visualisiert.

Vorkenntnisse und Kompetenzen:

Es sind keine spezifischen Vorkenntnisse erforderlich. Grundlagenkenntnisse in einer Programmiersprache sind allerdings hilfreich. Wesentlich ist das Interesse an Schnittstellenabfragen und bibliothekarischen Datenformaten sowie die Motivation zum gemeinsamen Arbeiten und Austausch.

Technisches Setup:

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

Teilnehmer*innenzahl:

Max. 15

Workshop-Programm:

Bei dem Workshop können die Teilnehmer*innen auf dem eigenen Laptop gemeinsam ein Online-Tutorial (Jupyter Notebook) ausführen und sich dazu untereinander und mit den Betreuerinnen austauschen. Neben einem WLAN-Zugang und einem gängigen Webbrowser ist vorab keine Software-Installation erforderlich. Alternativ können die Teilnehmer*innen auch ihre gewohnte lokale Installation zum Ausführen des Jupyter Notebooks nutzen. Nach der Einführung in das DNBLab-Angebot arbeiten die Teilnehmer*innen parallel in einem vorbereiteten Jupyter Notebook, während eine Betreuerin die einzelnen Schritte zeigt und eine zweite Betreuerin ein Jupyter Notebook ebenfalls live bearbeitet und für Fragen und Support zur Verfügung steht (Strecker et al. 2021 und Hall 2021).

Programmablauf:

Programmpunkt	Inhalte	Zeit in Minuten
1. Einführung (Vortrag)	Was bietet das DNBLab?	20
2. Fragestellung und Bezugswege (Vortrag)	Vorstellung eines Anwendungsfalls und Auswahlkriterien für Bezugsweg und Format	20
3. Aufruf Jupyter Notebooks	Kurzüberblick zur Funktionsweise	20
Pause		15
4. Datenset erstellen (hands-on)	Schnittstellenabfrage und Ausgabe des Datensets	40
5. Datenbereinigung (hands-on)	Datenextraktion in ein Dataframe unter Berücksichtigung der Formatverzeichnung	50
Pause		15
6. Exemplarische Analyse (hands-on)	Überprüfung und Häufigkeitsanalyse der bereinigten Daten	20
7. Visualisierung (hands-on)	Grafische Darstellung der Daten	30
8. Abschluss (Diskussion)	Austausch zum Ergebnis und weiteren Forschungsideen	10

Fußnoten

1. DNBLab: Zugang zu Datensets und digitalen Objekten unter https://www.dnb.de/dnblab, zugegriffen 14. Juli 2023
2. DH-Call und DH-Stipendien unter https://www.dnb.de/dhd, zugegriffen 14. Juli 2023
3. Aktivitäten und Ergebnisse unter https://www.dnb.de/dhd, zugegriffen 14. Juli 2023
4. Unsere Daten in der Praxis unter https://www.dnb.de/dnblabpraxis, zugegriffen 14. Juli 2023
5. Binder unter https://mybinder.org/, zugegriffen 14. Juli 2023
6. JupyterLite unter https://jupyterlite.readthedocs.io/ zugegriffen 2. November 2023
7. Veranstaltungen unter https://www.dnb.de/dnblab, zugegriffen 14. Juli 2023

Bibliographie

Althage, Melanie; Dröge, Martin; Hiltmann, Torsten; Schneider, Torsten: Python für Historiker:innen. Ein anwendungsorientierter und interaktiver Einstieg, 20.07.2022, Jupyter Book, (v1.0), https://digital-history-berlin.github.io/Python-fuer-Historiker-innen/home.html , https://doi.org/10.5281/zenodo.6868043

Döhl, Frédéric; Zechmann, Dorothea: Digital Humanities und Recht: Zu den neuen Regeln für das Text und Data Mining (TDM) und ihrem strategischen Potential für die Bibliotheken, in b.i.t. online 23 (2020) Nr. 4, S. 397-404. https://www.b-i-t-online.de/heft/2020-04-fachbeitrag-doehl.pdf

Döhl, Frédéric: Deutsche Nationalbibliothek verstärkt Engagement in den Digital Humanities, in: Dialog mit Bibliotheken 2, 2019, S. 9-11. urn:nbn:de:101-2019081635

Dombrowski, Quinn: Jupyter notebooks for digital humanities. 2023. GitHub repository. https://github.com/quinnanya/dh-jupyter

Fischer, Frank; Jäschke, Robert: Liebe und Tod in der Deutschen Nationalbibliothek: Der DNB-Katalog als Forschungsobjekt der digitalen Literaturwissenschaft. In: DHd2018: ”Kritik der digitalenVernunft”, Digital Humanities im deutschsprachigen Raum, Cologne, Germany. 261-266. https://hal.archives-ouvertes.fr/hal-01787558/document

Hall, Marc: DHd-AG-ZZ Metadata Tutorial. 2021. GitHub repository. https://github.com/mmh352/metadata-tutorial

Jannidis, Fotis; Konle, Leonard; Leinen, Peter: Makroanalytische Untersuchung von Heftromanen. In: DHd 2019. Digital Humanities: multimedial & multimodal. Konferenz-abstracts. 2019, p. 167-172. https://zenodo.org/record/2596095#.Xde4jm5Fzct

Strecker, Dorothea, Lukas C. Bossert und Évariste Demandt. 2021. Das Versprechen der Vernetzung der NFDI.Bausteine Forschungsdatenmanagement.

Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und –managern Nr.3/2021: S. 39-55. https://doi.org/10.17192/bfdm.2021.3.8336

Taube, Anke. „DNBLab – Zugang zu freien Datensets und digitalen Objekten“. Dezember 2021. https://doi.org/10.5281/zenodo.5850794

Teichmann, Lisa. (2022). Geomapping bibliographic translation data from the German National Library catalogue: A casestudy. Spatial Humanities 2022, Ghent. Zenodo. https://doi.org/10.5281/zenodo.7058385

How to do Theory: Reflexive Praktiken in der DH Lehre

Geiger, Jonathan D.
Jonathan.Geiger@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0000-0002-0452-7075

Horstmann, Jan
jan.horstmann@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0001-8047-2232

Kleymann, Rabea
rabea.kleymann@phil.tu-chemnitz.de
Technische Universität Chemnitz, Deutschland
ORCID: 0000-0003-3856-2685

Schröter, Julian
j.schroeter@lmu.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-0168-2608

Einleitung: Ziele des Workshops

Die Anzahl akkreditierter Digital-Humanities-Studiengänge ist in den letzten zehn Jahren rasant gestiegen, sodass von einer Integration der Digital Humanities (DH) in der universitären Lehre gesprochen werden kann.[1] Vor diesem Hintergrund rückt die Frage in den Fokus, welche Kompetenzen, Inhalte und Praktiken in der DH-Lehre eigentlich vermittelt werden sollen. Dabei spielen nicht nur Schlagwörter wie *data* bzw. *code literacy* eine Rolle (vgl. Schmidt 2016). Vielmehr können Lehrende und Studierende auf ein Repertoire von Tools, Plattformen, Tutorials und weiteren infrastrukturellen Settings zurückgreifen (vgl. *Programming Historian*[2], *dariahTeach*[3] , *Ranke.2*[4]). Im Zuge dessen können digitale und computationale Methoden niedrigschwelliger, schneller und früher in DH-Lehrformate eingebunden werden. Zugleich wird deutlich, dass die Forderung nach kritischen Praktiken des Theoretisierens und Reflektierens unter anderem auch mit Blick auf die jüngsten Entwicklungen von generativen Sprachmodellen sowie den Einfluss der Datenwissenschaften immer stärker werden.

Thema des von der AG *Digital Humanities Theorie* organisierten halbtägigen Workshops ist eine Auseinandersetzung und Weiterentwicklung von theorieorientierten Inhalten, Formaten und Ressourcen für die DH-Lehre. Dabei geht es nicht nur um die Frage, wie Praktiken des Theoretisierens in bestehende DH-Curricula integriert und eingebettet sind.[5] Vielmehr widmet sich der Workshop auch der Entwicklung zukünftiger Formate und fragt, welche fachwissenschaftlichen Kompetenzen, technischen Fähigkeiten und infrastrukturellen Ressourcen benötigt werden und welche Rolle Überlegungen hinsichtlich der Repräsentativität von Grundlagenliteratur spielen.

Ausgangspunkt des Workshops ist ein von uns erstellter multimodaler Reader „DH-Theorie“, der im Rahmen der Veranstaltung erstmals vorgestellt und diskutiert werden soll. Ziel ist eine disziplinäre Selbstverständigung über die bereits vorhandenen Profile von DH-Studiengängen im deutschsprachigen Raum, indem folgende Fragen adressiert werden: Welche Rolle spielt die Theoriebildung bei der Entwicklung und Implementierung von DH-Studiengängen? Und welche Vorstellung von geisteswissenschaftlicher Theoriebildung wird dabei implizit vorausgesetzt? Gibt es erste Erfahrungswerte, wie mit der DH-Theorie in der Lehre umgegangen wird? Das gemeinsame Lernziel des Workshops ist ein vertieftes Verständnis unterschiedlicher Möglichkeiten des Theoretisierens und der Theoriebildung in den DH sowie die Kompetenz, Theorieaspekte in Studiengängen der Digital Humanities im deutschsprachigen Bereich vergleichend einschätzen zu können.

Problemlagen: Von geisteswissenschaftlichen Theorien zu Praktiken des Theoretisierens in der DH-Lehre

Für die Begriffe „Theorie“ und „Theoretisieren“ liegt in den DH bislang kein einheitliches Verständnis vor. Beide Begriffe zeichnen sich durch eine Ambiguität aus, die mit Bezug auf spezifische DH-Lehrkontexte zunächst die Frage aufwirft, was vermittelt werden soll bzw. welche Felder und Bereiche reflexiver Tätigkeit überhaupt dem Bereich „Theorie in der DH-Lehre“ zugerechnet werden sollen. Was kann unter einer DH-Theoriekompetenz verstanden werden? Gehören beispielsweise Grundlagen der Mathematik und der Statistik dazu? Welche Rolle spielen interdisziplinäre Dialogfähigkeit und die Erprobung und Reflexion kooperativer Forschung (vgl. Wuttke 2022, 50)? Wie unterscheidet sich die Integration von DH-Theorie in die Lehre im internationalen Vergleich? Es existieren di-

verse Konzeptionen von „Theorie“ bzw. „Theoretisieren“: Das Theoretisieren oder auch *doing theory* wird von Spoerhase und Martus als reflexive „Praktik“ (2022, 170) gefasst. Theoretisieren sei Teil eines Ensembles von Praktiken, wie z. B. dem Deuten, Beschreiben und Erklären. Anders hat jüngst Kleymann (2023) drei Perspektiven auf den Theoriebegriff vorgeschlagen: „1. Theorie als Form der Reflexion, 2. Theorie als Praxeologie und Teil der Datenmodellierung, 3. Theorie als referenzierbare Sammlung von Texten, Lektüren und Konzepten“. In dieser dritten Hinsicht erarbeitet die AG aktuell einen Reader „DH-Theorie“, der eine kuratierte und kommentierte Auswahl von Items zur Einführung in die Theorie der DH darstellt. Fragen nach Zugehörigkeit und Abgrenzung, Disziplinspezifik und Generalisierbarkeit und nicht zuletzt auch nach Kanonizität durch häufige Referenz bestimmen den Auswahlprozess bei der Zusammenstellung der Texte.

Eine reflexive Praktik des Theoretisierens kann und sollte auch in Bezug auf den Umgang mit DH-Tools beleuchtet werden. Diskutieren lässt sich hierbei, ob und in welchem Grad die Verfügbarkeit technischer Instrumente einer Theoretisierung im Wege stand bzw. steht, weil KWIC-Tools, Stylo, Machine-Learning-Bibliotheken in Python etc. es erlauben, eine Auseinandersetzung mit statistischen Grundlagen zu vermeiden. Besonders dringend ist diese Frage bei den aktuellen generativen Sprachmodellen, die lediglich eine Prompting-Kompetenz zu erfordern scheinen. Gerade im Umgang mit 'Tools' gibt es ein potenzielles dialektisches Verhältnis zwischen einem 'Bedarf nach mehr Theoretisierung bei gleichzeitiger Praxis ohne theoretischen Boden' zu diskutieren (vgl. Kemman 2022). Zu diskutieren ist hier auch die Frage nach dem Verhältnis von Theorie im Sinn der mathematisch/statistisch/technischen Grundlagen, und Theorie im Sinn einer allgemeineren Wissenschaftstheorie der digitalen Geisteswissenschaften. Einer Klärung bedarf etwa die Frage, ob und wie in der DH-Lehre diese Verschränkung von unterschiedlichen Theorie-Dimensionen geleistet werden kann oder soll. Den Hintergrund dieser Diskussionen bildet auch eine Öffnung des pädagogischen Dialogs im Umgang mit digitalen Technologien, wie sie bereits von Alan Liu (2009, 20) beschrieben wurde: „One of the most remarkable differences of teaching with the new technologies is that they supplement the usual closed discursive circuit of the instructor-talking-to-the-student (and vice versa) with an open circuit of the instructor-and-student talking to others.“

Studiengänge und Modulhandbücher

Studiengänge bzw. deren Aufbau, Formate und Inhalte können als Indikatoren für die Entwicklung eines wissenschaftlichen Feldes gesehen und analysiert werden. Im Hinblick auf die Theorie-Anteile in aktuellen (deutschen) DH-Studiengängen zeigen sich ein heterogenes Bild, aber auch wiederkehrende Muster. Die Heterogenität der DH als Forschungsfeld spiegelt sich in den Studiengängen wieder (begrifflich und vom Aufbau her), d. h. es finden sich Studiengänge, die eine fachliche Verortung („Digital Humanities“) prominent im Titel tragen (z. B. an der TU Dresden oder an der Friedrich-Alexander-Universität Erlangen-Nürnberg), andere Studiengänge werden dem Feld zwar zugeordnet, verbergen ihre Module aber unter einem eher technischen oder allgemeinen Studiengangstitel (z. B. „Informationsverarbeitung“ an der Universität Köln).[6] Diese Heterogenität setzt sich auf Ebene der inhaltlichen Struktur fort. Theorieanteile setzen implizit natürlich ein bestimmtes Verständnis von Theorie voraus, doch auch in einem sehr weiten Sinne (Theorie in den DH als Historisierung, Wissenschaftstheorie, Soziologisierung etc.) lassen sich nicht viele theorieorientierte Studieninhalte finden. Die meisten Studiengänge verfügen über ein Einführungsmodul (Vorlesung oder Seminar) „Digital Humanities“, in dem ein Überblick über das Feld, seine Gegenstände, Methoden und Forschungsfragen gegeben wird. Zusätzlich oder stattdessen lassen sich auch oft spezifischere Einführungsveranstaltungen finden, beispielsweise das Einführungsmodul „Digitale Philologie“ an der TU Darmstadt. Weiterhin gibt es einige wenige Beispiele für Studiengänge mit einem Modul, das sich mit aktuellen Trends in den DH beschäftigt (beispielsweise im „Digital Humanities“-Studiengang an der Friedrich-Alexander-Universität Erlangen-Nürnberg). Insgesamt dominieren in den Studiengängen methodische und informatikorientierte Inhalte (z. B. zu Datentypen, Algorithmen, Statistik, Human-Computer-Interaction, Datenbanken). Häufig wird die Thematisierung der Entwicklung der DH in Modulbeschreibungen erwähnt (z. B. im Modul „Humanities Computing“ des Studiengangs Informationsverarbeitung an der Universität Köln, dem Studiengang „Digital Humanities“ an der Universität Trier oder dem Studiengang „Digital Humanities“ an der Universität Leipzig). Selten finden sich explizit theorie-orientierte Studienangebote, wie beispielsweise das Modul „Philosophy of Science and of Research in the Humanities“ im Studiengang „Data and Discourse Studies“ an der TU Darmstadt. Konkrete Angaben zu den Themen und Inhalte der Module sowie etwaige Literaturhinweise lassen sich in den Modulhandbüchern der Studiengänge quasi nicht finden, sodass eine Analyse der Modulhandbücher schnell an seine Grenzen stößt und zur weiteren Ergründung der Theorieanteile in DH-Studiengängen zusätzliche Formate und Zugriffsarten notwendig werden.

Methodik und Ablauf des Workshops

Der Workshop startet mit einer Einführung in die Thematik und zentrale Fragestellungen. Im Anschluss wird der Reader der AG DH Theorie in seiner dann aktuellen Form vorgestellt und es wird Zeit für Fragen und Antworten eingeräumt.

Hieran schließt sich eine Paneldiskussion an, deren Funktion das breite Auffächern der Thematik einerseits und andererseits die Vermittlung von Erfahrungswissen aus der Praxis von Studiengangskoordinator*innen darstellt. Die Dauer der Diskussion ist (inkl. Pitches der zentralen Thesen der Panelist*innen) auf 60 Minuten angesetzt. Einladungen an die potenziellen Panelist*innen wurden bereits versendet, Zusagen liegen derzeit von Prof. Dr. Christof Schöch (Universität Trier), Dr. Agnes Thomas (Hochschule Mainz und Mainzer Zentrum für Digitalität in den Geistes- und Kulturwissenschaften), Dr. Dominik Kremer (FAU Erlangen-Nürnberg) und Prof. Dr. Julianne Nyhan (TU Darmstadt) vor. Leitfragen für die Paneldiskussion beziehen sich unter anderem auf die aktuelle Situation der Rolle von Theorien und theorieorientierten kritischen Reflexionen in DH-Studiengängen, der didaktischen Einbettung sowie die derzeitigen Bedarfe.

Nach der Paneldiskussion und einer Pause werden die gesammelten Impulse und Ideen der Teilnehmenden durch ein Open-Space-Format gesammelt. Gemeinsam im Plenum werden wir Themenbereiche diskutieren und für Thementische fixieren. Die Unterteilung des Themenkomplexes erfolgt also entlang der akuten Gesprächs- und Diskussionsbedarfe der Teilnehmenden. Anschließend haben diese die Möglichkeit, sich nach eigenen Interessen und Kompetenzen innerhalb von drei Runden à 20 Minuten in die Diskussionen der Thementische einzubringen. Wie bei Unkonferenzformaten üblich gelten die Prinzipien der Offenheit, Hierarchielosigkeit und des Einbringens nach eigenem Interesse.

Im Anschluss an den Open Space werden die Ergebnisse der Thementische im Plenum vorgestellt, abschließend zusammengefasst und eine These auf die Frage nach dem aktuellen Stand und der weiteren Entwicklung der Rolle von Theorie in DH-Studiengängen (*quo vadis?*) gewagt. Der Workshop wird für eine hybride Teilnahme konzipiert: Einführung, Paneldiskussion, Vorstellung der Thementisch-Ergebnisse sowie Abschlussdiskussion eignen sich ebenfalls für eine virtuelle Teilnahme. Die Gruppenarbeit an den Thementischen wird ausschließlich für die Teilnehmenden in Präsenz zugänglich sein. Die Gesamtdauer des Workshops wird auf 4 Stunden veranschlagt.

Zielpublikum und Teilnehmer*innenzahl

Am Workshop können bis zu 25 Theorie-interessierte Personen aus allen Teildisziplinen der DH teilnehmen. Dabei richtet sich der Workshop vor allem an Lehrende und Studierende in den DH, um Erfahrungen, Wünsche und Herausforderungen auszutauschen.

Beteiligte und Forschungsinteressen

Jonathan D. Geiger arbeitet an der Akademie der Wissenschaften und der Literatur | Mainz im Infrastrukturprojekt NFDI4Culture und im NFDI-Projekt zu offenen Bildungsressourcen DALIA. Seine Forschungsschwerpunkte liegen auf der philosophisch-theoretischen Reflexion digitaler Forschungsmethoden in den Geisteswissenschaften und der Digitalität insgesamt.

Kontakt: jonathan.geiger@adwmainz.de , Digitale Akademie, Akademie der Wissenschaften und der Literatur | Mainz, Geschwister-Scholl-Straße 2, 55131 Mainz

Jan Horstmann leitet das Service Center for Digital Humanities (SCDH) an der Universitäts- und Landesbibliothek der Universität Münster. Seine Interessenschwerpunkte liegen im Bereich der digitalen Methodologie mit besonderem Fokus auf Textannotation, -analyse und Visualisierung im Bereich der computationellen Literaturwissenschaft.

Kontakt: jan.horstmann@uni-muenster.de , Universität Münster, Universitäts- und Landesbibliothek, Service Center for Digital Humanities (SCDH), Krummer Timpen 3, 48143 Münster

Rabea Kleymann ist Juniorprofessorin für Digital Humanities an der Technischen Universität Chemnitz. Ihre Forschungsinteressen liegen im Bereich der Wissenschaftstheorie, Critical Data Studies und Mixed-Methods-Forschung.

Kontakt: rabea.kleymann@phil.tu-chemnitz.de , Technische Universität Chemnitz, Reichenhainer Straße 41, C47.022, 09126 Chemnitz

Julian Schröter ist Professor für Digitale Literaturwissenschaften an der Ludwig-Maximilians-Universität München. Seine Forschungsinteressen liegen im Bereich der Theorie und Praxis quantitativer Literaturgeschichtsschreibung, der Literaturtheorie und der Methodologie der Computational Literary Studies.

Kontakt: j.schroeter@lmu.de Julian Schröter, Dep 13/I, LMU München, Schellingstr. 3 RG, 80799 München

Benötigte technische Ausstattung

Wir benötigen für den Workshop einen Raum mit Beamer für die Teile des Workshops, die im Plenum stattfinden, sowie eine flexible Bestuhlung und Tische, die sowohl für die Podiumsdiskussion als auch für den Open Space (Gruppentische) rearrangiert werden können. Darüber hinaus benötigen wir für das Konferenzformat zwei Moderationskoffer und vier Stell(pinn)wände.

Fußnoten

1. Vgl. https://dhcr.clarin-dariah.eu/courses (zugegriffen: 18.07.2023).

2. Vgl. https://programminghistorian.org/ (zugegriffen: 18.07.2023).
3. Vgl. https://teach.dariah.eu/ (zugegriffen: 18.07.2023).
4. Vgl. https://ranke2.uni.lu/de/ (zugegriffen 17.11.2023).
5. Der Titel des Workshops spielt auf Wolfgang Isers Studie „How to do theory“ an. Der Workshop möchte erneut das „Wie“ der Theoriebildung in den Fokus rücken. Während sich Iser jedoch mit der Frage beschäftigt, „how theory is done“ (2006, 12), widmet sich der Workshop konkreten Voraussetzungen und Bedingungen der Theorievermittlung in der Lehre.
6. Vgl. Portal Kleine Fächer, https://www.kleinefaecher.de/kartierung/kleine-faecher-von-a-z?tx_dmdb_monitoring%5Baction%5D=showByLocations&tx_dmdb_monitoring%5Bcontroller%5D=DisciplineTaxonomy&tx_dmdb_monitoring%5BdisciplineTaxonomy%5D=140&cHash=cd0a45fc052773fea946b7bf2872c1c1 (zugegriffen: 18.07.2023).

Bibliographie

Iser, Wolfgang. 2006. *How to do theory*. Malden: Blackwell.

Kemman, Max. 2022. „Tool Criticism through Playful Digital Humanities Pedagogy“. In *The Bloomsbury Handbook to the Digital Humanities*, hg. von James O’Sullivan, 287–294. Bloomsbury Academic. DOI: 10.5040/9781350232143.ch-13 .

Kleymann, Rabea. 2023. „Theorie“. In *Begriffe der Digital Humanities. Ein diskursives Glossar*, hg. von AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum e. V. (= Zeitschrift für digitale Geisteswissenschaften / Working Papers, 2). Wolfenbüttel. DOI: 10.17175/wp_2023_013 .

Liu, Alan. 2009. „Digital Humanities and Academic Change“. In *English Language Notes* 47 (1): 17–35. DOI: 10.1215/00138282-47.1.17 .

Martus, Steffen und Carlos Spoerhase. 2022. *Geistesarbeit: eine Praxeologie der Geisteswissenschaften*. Berlin: Suhrkamp.

Schmidt, Benjamin. 2016. „Do Digital Humanists need to Understand Algortihms? In *Debates in the Digital Humanities 2016*, hg. von Matthew K. Gold und Lauren F. Klein. Minneapolis.

Wuttke, Ulrike. 2022. „Wege bereiten, vermitteln und Denkräume schaffen! Reflexionen zu institutionellen und infrastrukturellen Erfolgsfaktoren für Digital Humanities an deutschen Universitäten auf Grundlage von Expert*inneninterviews.“ In *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel. DOI: 10.17175/2022_006 .

Literatur im Wikiversum – Eine praktische Annäherung über API-Abfragen und Wikipedia-Metriken

Illmer, Viktor J.
v.illmer@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0002-7334-781X

Soethaert, Bart
bart.soethaert@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0002-3845-605X

Welz, Lilly
l.welz@fu-berlin.de
Freie Universität Berlin, Deutschland

Fischer, Frank
fr.fischer@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0003-2419-6629

Jäschke, Robert
robert.jaeschke@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID: 0000-0003-3271-9653

Wikipedia als Repräsentant für populäres Wissen über Literatur

Die kollaborativ erstellte Online-Enzyklopädie Wikipedia bietet mit derzeit über 60 Millionen Artikeln in über 300 Sprachversionen (Wikimedia, 2023; Wikipedia-Beitragende, 2023) Informationen zu den unterschiedlichsten Wissensbereichen. Nach dem Wiki-Prinzip werden die Beiträge von einer weltweiten Gemeinschaft freiwilliger Redakteur*innen erstellt, bearbeitet und gepflegt. Diese Community überprüft aktiv die Referenzen und Änderungen in bestehenden Artikeln sowie die Veröffentlichung neuer Artikel auf die Einhaltung der redaktionellen Richtlinien. Aufgrund des offenen Redaktionsmodells ist Wikipedia ein enzyklopädisches Gemeinschaftsprojekt, das seinesgleichen sucht (Danowski et al. 2005; Lovink et al. 2012; Tkacz, 2015).

Neben der individuellen Lektüre der Fließtexte (und ihrer jeweiligen Versionshistorie) über die Website bietet die Online-Plattform über eine API weitere Möglichkeiten zur Analyse der enzyklopädischen Inhalte und des Community-Engagements. Die Vielzahl an Metadaten, sowohl zu den einzelnen Themen selbst als auch zur Bearbeitung und Nutzung durch die aktiv partizipierende Community bzw. die Leser*innen sowie die semantischen Verknüpfungen lassen sich auch mit digitalen Methoden sammeln, quantifizieren und auswerten.

Auch die rezeptionsorientierte Literaturwissenschaft hat das Projekt inzwischen als Forschungsgegenstand und Datenressource entdeckt (vgl. Hube et al., 2017; Chiu, 2022; Fischer et al. 2023b), da es viele enzyklopädische Beiträge und Metadaten zur Literatur und zum literarischen Leben versammelt, zu Autor*innen, literarischen Werken, Genres, Epochen und weiteren literaturgeschichtlich relevanten Kategorien.

Jüngste Untersuchungen in diesem Bereich werten die inhaltliche Reichweite von Wikipedia etwa im Hinblick auf die Aufnahme und Darstellung von einzelnen Autor*innen (Blakesley, 2018; Bronner, 2018; Fischer et al., 2019; Blakesley, 2022b), Gruppen (Blakesley, 2020; Carrillo-Jara, 2023), literarischen Werken (Blakesley, 2022a), literarischen Figuren (Picard et al., 2023; Wojcik et al., 2023), Gattungen (Figlerowicz, 2023) und Kanones (Miller et al., 2016; van der Deijl et al., 2018; Wojcik et. al, 2019; Lippolis, 2023) aus. Unterschiede in der Verteilung enzyklopädischer Artikel zu bestimmten Themen in verschiedenen Sprachen geben Aufschluss über das unterschiedliche Interesse und die attestierte Relevanz dieser Themen für bestimmte Sprachgemeinschaften. Darüber hinaus können Veränderungen der Seitenaufrufe, der Überarbeitungen und der Beitragenden auch im Zeitverlauf analysiert werden, um das sich entwickelnde Interesse an und die Auseinandersetzung mit bestimmten literarischen Autor*innen und Werken zu verfolgen. Die datenanalytische Auswertung anhand solcher Wikipedia-Metriken ermöglicht es somit, die Auseinandersetzung mit Literatur in Wikipedia evaluierbar zu machen und Aussagen über literarische Kanonizität, Wertungspraktiken und Popularität im Kontext offener Enzyklopädieprojekte weiter zu diversifizieren. In kritischer Auseinandersetzung mit der Kanon- und Popularitätsforschung in globaler Perspektive wird unter anderem besonders deutlich, dass sich in der Wikipedia kein monolithischer Kanon zeigt, sondern viele, sich zudem dynamisch verändernde Kanones manifestieren.

Im Zentrum des Hands-On-Workshops steht die Wikipedia-API, mit deren Funktionsweise die Teilnehmer*innen vertraut gemacht werden. Sukzessive werden Abfrageskripte in Form eines Jupyter Notebooks erarbeitet. Um eine benutzerfreundliche Programmierumgebung anzubieten und langwierige Installationsprozesse zu umgehen, wird für das Ausführen des Notebooks auf Google Colaboratory zurückgegriffen. Im Folgenden werden drei Typen von Abfragen kurz vorgestellt, die im Workshop jeweils im Hinblick auf eigene, von den Teilnehmer*innen mitgebrachte Fragestellungen und Forschungsinteressen modifiziert werden können. Der Programmiercode wurde um Annotationen ergänzt, die es auch Python-Anfänger*innen ermöglichen, über die bereitgestellten Formulare eigene Anfragen auszuführen.

Autor*innenzentrierte Abfragen

Eine mögliche Abfragestrategie ist die autor*innenzentrierte Abfrage, wie am Beispiel von Theodor Fontane demonstriert werden soll.

Über die Wikipedia-API lässt sich herausfinden, wie viele der über 300 Sprachversionen der Wikipedia einen eigenen Artikel über den Autor bereithalten – die Anzahl dieser Sitelinks gilt in der Forschung als »a simple measure of canonicity« (Kukkonen 2020). Diese können dann etwa diagrammatisch auf ihre Artikelgröße hin verglichen werden (Abb. 1). Auf diese Weise können ebenfalls die Anzahl der Überarbeitungen des Artikels, die Anzahl der Bearbeiter*innen, die Backlinks oder das Datum der Artikelerstellung untersucht werden. Diese Datenpunkte können sprachübergreifend Aufschluss über etwaige Konjunkturen der Fontane-Rezeption geben. Es lassen sich Rückschlüsse auf Anlässe ziehen, die eine Erweiterung der Informationsbasis in der Wikipedia ausgelöst haben könnten (Preise, Jubiläen, Übersetzungen, Schulstoff).

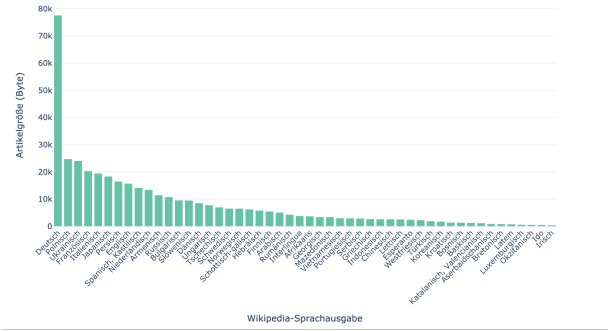


Abb. 1. Größe des Artikels *Theodor Fontane* nach Wikipedia-Sprachausgabe.

Werkzentrierte Abfragen und breitere Kontextualisierung

In ähnlicher Weise kann vorgegangen werden, wenn das Augenmerk auf die Werke Fontanes gerichtet wird. Auch hier können die Anzahl der Überarbeitungen des Artikels, die Anzahl der Bearbeiter*innen, der Backlinks oder das Datum der Artikelerstellung abgefragt und visualisiert werden.

Im Folgenden kann als digitale Entsprechung der literaturhistorischen Praxis ein Vergleichspool anderer Autor*innen zusammengestellt werden, um Fontane und sein li-

terarisches Werk mit denen zeitgenössischer Kolleg*innen zu vergleichen.

Fontanes Geburtsjahr ist 1819, eine mögliche Operationalisierung von Zeitgenossenschaft wäre etwa die Zusammenstellung anderer deutschsprachiger Autor*innen, die bis zu 20 Jahre vor und nach Fontane geboren wurden. Diese Festlegung ist natürlich kontingent und kann individuellen Informationsbedürfnissen angepasst werden.

Die Visualisierung der Artikellänge und ein entsprechendes Ranking ergeben dann beispielsweise, dass Wilhelm Busch unter den zeitgenössischen literarisch schreibenden Autor*innen den aktuell umfangreichsten Artikel vorweisen kann, Fontane aber immerhin in den Top-5 rangiert (Abb. 2). Auch wenn sich diese Artikellängen, die oft über mehr als 20 Jahre gewachsen sind, größenordnungsmäßig nicht so schnell ändern, sind diese Werte in einer communitybetriebenen digitalen Enzyklopädie natürlich durch die Zeit variabel.

Dass Wilhelm Wundt, als Psychologe und Philosoph ebenfalls einflussreicher Autor, bei diesem Ranking mit dem umfangreichsten Artikel ganz vorn steht, zeigt auch, dass es eines weiteren Schritts bedürfte, wollte man die Ergebnisliste auf vorderhand literarisch schreibende Zeitgenoss*innen eingrenzen. Außerdem zeigt sich ein systematischer Bias in der Artikellänge: Zu umfangreiche »Werk«-Abschnitte werden oft in eigene Artikel für Werke ausgegliedert, während die Werke »kleinerer« Autor*innen oft Teil der Personenartikel bleiben. Die genaue Kenntnis der Gepflogenheiten innerhalb von Wikipedia erweist sich daher als Voraussetzung für eine sinnvolle Einschätzung der Quantifizierungen.

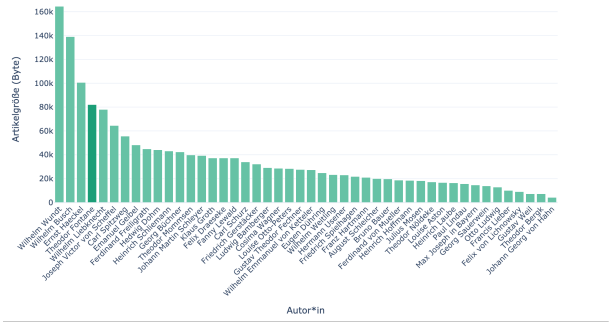


Abb. 2. Umfang der Artikel zu Fontanes Zeitgenoss*innen in der deutschsprachigen Wikipedia.

Abfragen zur Ermittlung der relativen Bedeutung von Artikeln innerhalb der Wikipedia

Neben den inhaltsbezogenen Informationen, die über Wikipedia direkt bezogen werden können, lassen sich über die internen Verweise zwischen Artikeln (Wikilinks) auch Netzwerkmetriken wie der PageRank berechnen (vgl. Thalhammer, 2016; Hube et al., 2017). Ähnlich wie dies Suchmaschinenalgorithmen zur Bestimmung der Rangfolge von Ergebnissen tun, können diese Zahlen dazu verwendet werden, den Wert und die Bedeutung eines Wikipedia-Artikels und seines Themas innerhalb des Hyperlink-Netzwerks der Plattform einzuschätzen. Der Wert, der einem Wikipedia-Artikel auf der Grundlage dieser Metriken zugewiesen wird, gibt Aufschluss über die relative Bedeutung und Konnektivität eines Themas innerhalb der vernetzten Informationen von Wikipedia.

Ausblick

Die Abbildung der Literatur in Wikipedia und in darauf aufbauenden oder damit verwandten Projekten (DBpedia, Wikidata) wird in den Literaturwissenschaften zunehmend als möglicher Forschungsgegenstand wahrgenommen. Einen repräsentativen Überblick über solche Zugänge und Fragestellungen bietet der Sonderband des *Journal of Cultural Analytics*, »Wikipedia, Wikidata, and World Literature« (Fischer et al., 2023a). Die Teilnehmer*innen des Workshops sollen in die Lage versetzt werden, daran anzuschließen und diese entstehende Praxis für ihre eigenen Forschungsfragen und -interessen produktiv zu machen. Während des interaktiven Workshops werden die Teilnehmer*innen praktisch erfahren und üben, wie sie auf relevante Metadaten in der Wikipedia zugreifen und eigenständig API-Abfragen durchführen können, um Einblicke in die Prozesse, den Umfang und die Inhalte der kollaborativen und selbstregulierten Informationsaggregation über Literatur im Wikiversum für die eigene Forschung nutzen zu können.

Format und Aufbau des Workshops (4h)

- 00:00–00:30 Einführung zum Thema anhand aktueller Forschungsbeiträge
- 00:30–01:00 Einführung in die Wikipedia-API
- 01:00–01:45 Allgemeine Erarbeitung des Jupyter Notebooks
- 01:45–02:00 Pause
- 02:00–03:00 Individuelle Anpassungen des Jupyter Notebooks
- 03:00–03:30 Präsentationen der Ergebnisse
- 03:30–04:00 Abschlussdiskussion und Perspektiven

Zielgruppe und notwendiges Vorwissen

Der Workshop zielt auf Literaturwissenschaftler*innen, aber auch auf Kolleg*innen angrenzender Gebiete. Vor-

kenntnisse der Programmiersprache Python sowie zu Programmierschnittstellen (APIs) sind hilfreich, aber keine Voraussetzung zur Teilnahme am Workshop. Den Teilnehmer*innen soll das nötige Praxiswissen vermittelt werden, um eigenständig weiterzuarbeiten.

Zahl der möglichen Teilnehmer:innen

max. 20 Personen

Benötigte technische Ausstattung

Die Teilnehmer*innen benötigen einen eigenen Laptop . Für die Durchführung des Workshops wird ein Beamer benötigt; Flipchart und Stifte wären hilfreich. Als Grundlage für eigene Abfragen können Workshopteilnehmer*innen eine Liste mit Artikelnamen einer einzelnen Wikipedia-Sprachausgabe oder eine Liste mit Wikidata-IDs mitbringen. Der Raum sollte über ausreichend Lademöglichkeiten verfügen und ein verlässliches und schnelles WLAN bieten.

Beitragende

- Viktor J. Illmer
- Bart Soethaert
- Lilly Welz
- Frank Fischer
- Robert Jäschke

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.

Bibliographie

Blakesley, Jacob. 2018. » The Global Popularity of William Shakespeare in 303 Wikipedias « . *Memoria Di Shakespeare. A Journal of Shakespearean Studies* 5 (Dezember): 149–71. https://doi.org/10.13133/2283-8759/14509.

———. 2020. » World Literature According to Wikipedia Popularity and Book Translations: The Case of Modern Italian Poets « . *Comparative Critical Studies* 17 (3): 433–58. https://doi.org/10.3366/ccs.2020.0373.

———. 2022a. » The Global Popularity of Dante's ‚Divina Commedia': Translations, Libraries, Wikipedia « . *Journal of Dante Studies* 5. https://repository.upenn.edu/bibdant/vol5/iss1/8.

———. 2022b. » The Wikipedia Popularity of James Joyce « . *James Joyce Quarterly* 59 (2): 289–313. https://doi.org/10.1353/jjq.0.0164.

Bronner, Friedrich Georg. 2018. »WHWP – Walter Höllerer bei WikiPedia«. Doctoral Dissertation, Berlin: Technische Universität Berlin. https://doi.org/10.14279/depositonce-6865.

Carrillo-Jara, Daniel. 2023. » Escritor / Qillqaq: The Representation of Peruvian Literature in the Spanish and Quechua Wikipedias « . *Journal of Cultural Analytics* 8 (2). https://doi.org/10.22148/001c.73258.

Chiu, Kuei-fen. 2022. » World Literature in an Age of Digital Technologies: Digital Archive, Wikipedia, and Goodreads.com « . In *The Making of Chinese-Sinophone Literatures as World Literature* , herausgegeben von Kuei-fen Chiu und Yingjin Zhang, 1. Aufl., 217–36. Hong Kong University Press. https://doi.org/10.2307/j.ctv2j6xdjn.

Danowski, Patrick und Voss, Jakob. 2005. » Das Wissen der Welt – die Wikipedia « . In *Open Source Jahrbuch 2005* , herausgegeben von Matthias Bärwolff, Robert A. Gehring, und Bernd Lutterbeck, 393–405. Berlin: Lehmanns Media. http://www.opensourcejahrbuch.de/download/jb2005/chapter_06/osjb2005-06-05-danowskivoss.

van der Deijl, Lucas und Smeets, Roel. 2018. » The Canon of Dutch Literature According to Wikipedia: A Network Analysis of 2286 Wikipedia-Entries on Dutch Authors. « In *Proceedings of EADH2018: » Data in Digital Humanities «* . Galway: National University of Ireland. https://eadh2018.exordo.com/files/papers/104/final_draft/Abstract_Long_Paper_EADH_2018_Van_der_Deijl___-Smeets_final_version.pdf.

Figlerowicz, Matylda und Mertehikian, Lucas. 2023. » An Ever-Expanding World Literary Genre: Defining Magic Realism on Wikipedia « . *Journal of Cultural Analytics* , 8 (2). https://doi.org/10.22148/001c.73249.

Fischer, Frank, Jacob Blakesley, Paula Wojcik, und Robert Jäschke, Hrsg. 2023a. *Wikipedia, Wikidata, and World Literature.* Special Issue of the *Journal of Cultural Analytics* , 8 (2). https://culturalanalytics.org/issue/7259.

Fischer, Frank, Jacob Blakesley, Paula Wojcik, und Robert Jäschke. 2023b. » Preface: World Literature in an Expanding Digital Space « . Journal of Cultural Analytics , 8 (2). https://doi.org/10.22148/001c.74598.

Fischer, Frank, und Robert Jäschke. 2019. »Fontane im Wikiversum. Ein Beitrag zur digitalen Rezeptionsgeschichte«. Conference Talk gehalten auf der Kongress »Fontanes Medien«, Potsdam, Juni 14. bit.ly/2wShKUT.

Hube, Christoph, Frank Fischer, Robert Jäschke, Gerhard Lauer, und Mads Rosendahl Thomsen. 2017. » World Literature According to Wikipedia: Introduction

to a DBpedia-Based Framework « . *ArXiv* . https://doi.org/10.48550/ARXIV.1701.00991.

Kukkonen, Karin. 2020. » Does Cognition Translate? « *Poetics Today* , 41 (2): 243–59. https://doi.org/10.1215/03335372-8172556.

Lippolis, Anna Sofia. 2023. » Italian Nostalgia: National and Global Identities of the Italian Novel « . *Journal of Cultural Analytics* , 8 (2). https://doi.org/10.22148/001c.68341.

Lovink, Geert, und Nathaniel Tkacz, Hrsg. 2012. *Critical Point of View: A Wikipedia Reader* . INC Reader 7. Amsterdam: Institute of Network Cultures. https://doi.org/10.2139/ssrn.2075015.

Miller, Ben, Cindy Berger, Sayan Bhattcharyya, Tommaso Caselli, David Kelman, Jennifer Olive, und Jay Rajiva. o. J. » Contextualizing Receptions of World Literature by Mining Multilingual Wikipedias « . In *Digital Humanities 2016: Conference Abstracts* , 282–85. Kraków: Jagiellonian University & Pedagogical University. https://dh2016.adho.org/abstracts/83.

Picard, Sophie, Paula Wojcik, und Sina Zarrieß. 2023. » Zirkulation und Wertschöpfung am Beispiel literarischer Figuren « . In *Der Wert der literarischen Zirkulation / The Value of Literary Circulation* , herausgegeben von Michael Gamper, Jutta Müller-Tamm, David Wachter, und Jasmin Wrobel, 431–53. Globalisierte Literaturen. Theorie und Geschichte transnationaler Buchkultur / Globalized Literatures. Theory and History of Transnational Book Culture 3. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-662-65544-3_26.

Thalhammer, Andreas, und Achim Rettinger. 2016. » PageRank on Wikipedia: Towards General Importance Scores for Entities « . In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016, Revised Selected Papers* , herausgegeben von Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenic, Sören Auer, und Christoph Lange, 227–40. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47602-5_41.

Tkacz, Nathaniel. 2015. *Wikipedia and the Politics of Openness* . Chicago/London: University of Chicago Press.

Wikimedia. 2023. »Wikimedia Statistics – All Wikipedias – Pages to date«. 2023. https://stats.wikimedia.org/#/all-wikipedia-projects/content/pages-to-date/normal|line|2-year|page_type~content|monthly.

Wikipedia-Beitragende. 2023. » List of Wikipedias « . In *Wikipedia* . https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=1161354568.

Wojcik, Paula, Bastian Bunzeck, und Sina Zarrieß. 2023. » The Wikipedia Republic of Literary Characters « . *Journal of Cultural Analytics* , 8 (2). https://doi.org/10.22148/001c.70251.

Wojcik, Paula, und Sophie Picard. 2019. » Klassiker@wikipedia: Klassikforschung und Digital Humanities. Ein Kommentar zur Studie World Literature According to Wikipedia « . In *Klassik als kulturelle Praxis. Funktional, intermedial, transkulturell* , herausgegeben von Paula Wojcik, Stefan Matuschek, Sophie Picard, und Monika Wolting, 149–64. spectrum Literaturwissenschaft / spectrum Literature 62. Berlin/Boston: de Gruyter. https://doi.org/10.1515/9783110615760-010.

Machine Learning to Read Yesterday's News. How semantic enrichments enhance the study of digitised historical newspapers

Bunout, Estelle

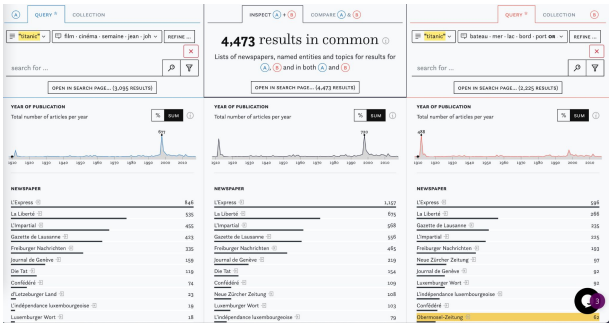
estelle.bunout@uni.lu
Universität Luxemburg, Luxemburg

Düring, Marten

marten.during@uni.lu
Universität Luxemburg, Luxemburg

In this workshop we will use the *impresso* app to explore opportunities and challenges which accompany the semantic enrichment of historical newspapers. We will reflect on the added value of Natural Language Processing techniques such as topic modelling, text reuse detection and word embeddings for historians in conjunction with an introduction and critical assessment of design solutions for the scalable reading of such enriched sources.

We target researchers at all (digital) skill levels.



Historical newspapers shaped and reflect the political, moral, and economic environments in which they were produced. They hold dense, continuous, and multi-level information which can help us reconstruct how contemporaries represented and experienced their present. This makes them

indispensable sources for research and their value is reflected in mass digitisation efforts over the past years.

As a consequence, researchers today face an abundance of materials which can no longer be managed with keyword search and basic content filtering alone even though only a fraction of the overall archival record has yet been processed. Digitisation also transformed analogue sources into highly complex digital objects determined by multiple layers of technical processing. Subsumed under the notion of "digital hermeneutics", scholars have pointed to the epistemological challenges inherent in such documents and called for a critical engagement with data provenance, its processing, and interfaces.

Our goal for this workshop is first for historians to familiarise themselves with the opportunities and pitfalls of semantically enriched sources. Second, we will introduce participants to the complexity of digitised newspaper collections and focus on key operations for their exploration and analysis. Third, we will encourage participants to discuss the capacities of the *impresso* app to support historical research and to offer transparency regarding data processing and interface functions. To this end we will focus on:

- Creation and comparison of user-generated content collections to reveal (dis)similarities (see e.g. screenshot above).
- Word embeddings to reveal synonyms, related terms, (historical) spelling variations and frequent OCR misspellings.
- Content filters based on topic models to include or exclude themes such as "sports", "arts" or "foreign politics".
- Content filters based on linked named entities to reveal the changing contexts in which entities such as persons, institutions and locations appear across time and newspapers.
- Article recommendations to identify potentially relevant content outside a researcher's search scope.
- Dedicated exploration interfaces for text reuse clusters, n-grams and topics to reveal trends over time and newspapers and to assist query-building.
- Image similarity search reveals the distribution of similar images within the corpus.
- Visualisations of gaps, biases in the corpus and confidence scores for OCR and entities to better manage user expectations as to what can be found in the corpus and to judge the value of any finds.
- Educational materials to document data processing and interface functionalities.
- The interplay between these components allows researchers to address generic, yet complex historical questions such as: "How did the news about the Titanic catastrophe travel through the media sphere?" or "What constitutes a "crisis"? And why did it peak in 1932 in the press?"

We propose the following structure, for a group of about 20 participants:

1. Introduction: *impresso* project, newspaper collections and semantic enrichment
2. Small groups: Content retrieval with newspaper interfaces, e.g. Deutsches Zeitungsportal, ANNO, or eLuxemburgensia.
3. Demonstration and hands-on tutorial of the impresso interface with focus on: data criticism, content retrieval, comparison, result representativity, blind spots, user-generated collections.
4. Discussion: Experiences of working with two distinct newspaper interfaces and their search tools, compare and contrast. To which extent do digital tools empower new search and discovery workflows for newspapers? Which new skills do such tools require? How can we trust computationally generated information? What questions could not be answered that should have been (corpus stats, overview…)? What interaction was missing?
5. Outlook: follow-up project, impresso API + notebooks for data-driven research and impresso Powervis experimentation: navigate the graphs to get more general impressions of the corpus, linked to the pre-determined questions

Please note that participants can be at any level of digital literacy, the bigger barrier might rely in the language of the source material (French and German mainly).

Also, the newspapers collections will be handled only via the app, to enable the exploration of the existing collections. In the context of this workshop, the content extraction will consist in searching with keywords, using filters based on semantic enrichments, visualise, comparing the results of queries, etc.

No need to download or install anything to conduct the workshop, all is web-based. The needed material for participants is a laptop with internet connection.

Convenors

Marten Düring, Assistant Professor in Digital History at the Luxembourg Centre for Contemporary and Digital History (C2DH), University of Luxembourg

Estelle Bunout, post-doctoral researcher at the C2DH.

Relevant publications:

- Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology, edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. Studies in Digital History and Hermeneutics. Berlin, Germany: De Gruyter, 2023.
- Düring, Marten, Roman Kalyakin and Daniele Guido. "Impresso Inspect and Compare. Visual Comparison of Semantically Enriched Historical Newspaper Articles." Information 12, no. 9 (September 2021): 348.

- With Ehrmann, Maud, et al. Historical Newspaper User Interfaces: A Review. 2017. library.ifla.org, http://library.ifla.org/2578/.
- « The digitisation of newspapers: how to turn a page », From the archival to the digital turn · Ranke.2, 2019, https://ranke2.uni.lu/lessons/
- « Collections of Digitised Newspapers as Historical Sources – Parthenos training », 2019: https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/

Bibliographie

Alberto Romele, Marta Severo, and Paolo Furia, "Digital Hermeneutics: From Interpreting with Machines to Interpretational Machines," AI & SOCIETY 35, no. 1 (March 2020): 73–86, https://doi.org/10.1007/s00146-018-0856-2;

Andreas Fickers and Juliane Tatarinov, eds., Digital History and Hermeneutics: Between Theory and Practice, Digital History and Hermeneutics, vol. 2 (De Gruyter Oldenbourg, 2022), https://www.degruyter.com/document/isbn/9783110723991/html;

Chiel van den Akker et al., "Digital Hermeneutics: Agora and the Online Understanding of Cultural Heritage," in Proceedings of the 3rd International Web Science Conference, WebSci '11 (Koblenz, Germany: Association for Computing Machinery, 2011), 1–7, https://doi.org/10.1145/2527031.2527039.

Microblogging mit Mastodon: Fediverse, Fedihum und Co. in den Digital Humanities – ein Praxisworkshop

König, Mareike
mkoenig@dhi-paris.fr
Deutsches Historisches Institut Paris, Frankreich
ORCID: 0000-0002-8189-8574

Hermes, Jürgen
hermesj@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland
ORCID: 0000-0002-8367-8073

Schildkamp, Philip
philip.schildkamp@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland
ORCID: 0000-0003-0209-2837

Wolter, Vivien
vivien.wolter@gmx.net
Universität Trier, Deutschland
ORCID: 0000-0002-9982-9382

Wuttke, Ulrike
ulrike.wuttke@fh-potsdam.de
Fachhochschule Potsdam, Deutschland
ORCID: 0000-0002-8217-4025

Einführung

Soziale Medien wie Blogs, Twitter, Facebook und andere haben beim Aufbau der Digital Humanities-Community seit Mitte der 2000er Jahre eine wichtige Rolle gespielt. DH-Forschende gelten als frühe Anwender:innen sozialer Medien für Vernetzung, Austausch und Kommunikation (Kirschenbaum, 2012). Neben Wissenschaftsblogs erschien über viele Jahre der Kurznachrichtendienst Twitter für die DH-Community zentral zu sein: zugleich informative wie soziale Grundlage über sprachliche und räumliche Grenzen (Haustein et al., 2015), Verbreitungsmittel für die eigene Forschung (Quan-Haase et al., 2015), für Wissenschaftskommunikation in die interessierte Öffentlichkeit hinein (Acatech, 2017; European Commission, 2020) und selbst Untersuchungsgegenstand etwa für Netzwerk-, Sprach- oder Inhaltsanalysen (z. B. Wuttke, 2021, König und Ramisch, 2022). Die Forschung zu sozialen Medien in der Wissenschaft hat die Breite der Praktiken in Abhängigkeit von individuellen Vorlieben, der Position in der Wissenschaft, von Geschlecht, Alter und Disziplinen gezeigt (Sugimoto et al., 2017). Parallel dazu florierten erfahrungsbasierte Ratgeber, die mit unterschiedlichen Schwerpunkten die Vorteile von sozialen Medien für die Eigen-PR (etwa das Steigern der Zitationsrate eigener Aufsätze), für Vernetzung, Austausch und zum Einsatz in der Lehre priesen (vgl. z.B. Trapp, 2015, Coté und Darling, 2018, Geier und Gottschling, 2019).

Zugleich stellte sich die Erkenntnis ein, dass die kommerziellen Plattformen vor allem auf Gewinnmaximierung ausgelegt sind und die Kritik an sozialen Medien nahm über die Jahre zu: insbesondere an der zunehmenden Verrohung im Austausch – inklusive Angriffe auf Wissenschaftler:innen – am unklaren Umgang mit der Privatsphäre von Nutzenden sowie an undurchsichtigen Algorithmen für die Sichtbarkeit von Postings. "Today you can talk about your research on social media platforms all you want, but hardly anyone will hear you unless you pay cash money because of

Algorithms" (Mewburn, 2023). Die Übernahme-Saga von Twitter durch Elon Musk zwischen April und November 2022 und in der Folge die technischen, finanziellen und politischen Managemententscheidungen sorgten für mehrere Exodus-Wellen von Twitter. Erst vor kurzem bedeutete die weitgehende Schließung der freien Twitter-API und die Einführung von Lese- und Postlimits für Tweets das Ende sowohl für Forschungsprojekte, die ohne Budget für den Ankauf von Daten Tweets auswerten möchten, als auch für automatisierte Twitterprojekte, die freie Dienste wie autoChirp (Hermes et al., 2017) für das Planen und Verbreiten ihrer Tweets nutzen. Schließlich halten auch ethische und gesellschaftspolitische Motivationen Forschende dazu an, Twitter zu verlassen und sich dem dezentralen Fediverse (eine Kombination aus *federation* und *universe*) zuzuwenden, ein Netzwerk miteinander föderierter dezentraler sozialer Netzwerke und Kommunikationsdienste, basierend auf freier Software und auf dem vom W3C definierten offenen Kommunikationsprotokoll Activity Pub.

Aus der langen Agonie von Twitter sollte gleichwohl nicht auf ein Ende der sozialen Medien und der Wissenschaftskommunikation geschlossen werden, wenn auch das Nachdenken über Nachhaltigkeit und Zugänglichkeit für wissenschaftliche Inhalte auf Social-Media-Plattformen intensiviert wurde (König 2023). Im DHd-Verband hat sich letztes Jahr eine AG Digitale Wissenschaftskommunikation und Public Humanities gegründet, die Theoretisierung dieses Bereichs mit seinen vielfältigen Praktiken und Facetten ist weiter im Gange (vgl. z. B. Seltmann, 2023), das Fediverse erlebt Auftrieb und Forschungseinrichtungen fördern Wissenschaftskommunikation angesichts einer zunehmenden Wissenschaftsfeindlichkeit und Polarisierung in der Gesellschaft (z. B. die Initiative Wissenschaftskommunikation.de) . Kurz: Wer den Einstieg in die Nutzung sozialer Medien für die DH-Forschung und -Lehre plant, kommt zum richtigen Zeitpunkt, benötigt jedoch anderes Wissen und andere Strategieplanungen als noch vor einigen Jahren.

Die Fedihum-Community auf Mastodon

Angesichts der massiven Probleme von Twitter und ermutigt durch die Community hat der DHd-Verband am 14.11.2022 mit Fedihum eine Mastodon-Instanz initiiert (Wuttke 2022; eine Anleitung bei König, 2022). Mit Stand 17.7.2023 haben sich dort 236 Nutzer:innen angemeldet. Unterstützt wurde die Einrichtung des neuen Servers von Ralf Stockmann (Staatsbibliothek zu Berlin), was zugleich ein Beispiel für den kooperativen Spirit des Fediverse ist. Fedihum wurde mit dem Ziel initiiert, DH-Aficionados eine Alternative für Twitter zu bieten und den Austausch der DH(d)-Community im Fediverse zu fördern. Die Plattform steht DH-Forschenden in allen Sprachen unabhängig von einer DHd-Mitgliedschaft offen, solange sie sich an die Serverregeln halten. Im Sinne der DHd-AG Greening DH wurde bei der Auswahl des Hosting-Anbieters darauf geachtet, dass dieser mit Ökostrom arbeitet. Administriert und moderiert wird der Server von der DHd-Communication Fellow Vivien Wolter und weiteren Freiwilligen. Das Logo der Mastodoninstanz und das Profilbild des Adminaccounts wurden über einen Community-Wettbewerb ermittelt.

Auch wenn Zuspruch und Aktivitäten auf Fedihum nach den ersten Wechselwellen gut sind, sehen wir Bedarf an einem Mastodon-Workshop bei 1) Personen, die Startschwierigkeiten im Fediverse verspüren (Wie fülle ich meine Timeline? Welche spannenden Funktionen und Einstellungen gibt es?), bei 2) Personen, die kommerzielle Social Media abgelehnt haben und sich nun dem freien Fediverse anvertrauen möchten und 3) bei Forschenden, die automatisierte Publikationen von Postings für die Lehre oder ihre Projekte planen.

Automatisierte Veröffentlichung von Social-Media-Posts

Die oben beschriebenen Änderungen bei Twitter führten dazu, dass eine ganze Reihe von Twitter-basierten wissenschaftlichen DataScience-Projekten oder auch automatisierte Accounts eingestellt werden mussten. Darunter ist der von der Universität Köln betriebene Service autoChirp, der die Grenze von maximal 50 erlaubten Tweets pro Tag/Service überschritt und dementsprechend suspendiert wurde. Bis zuletzt wurde eine zweistellige Anzahl von Projekten über autoChirp realisiert – Beispiele sind @jeanpaultoday , @UniBielefeld50 und @satzomat (Hermes, 2021) –, so dass durchschnittlich jedes Projekt täglich weniger als fünf Tweets zur Verfügung hatte, was unzureichend ist.

Die automatisierte und auf Kurznachrichtendienste abgestimmte Veröffentlichung von Forschungsdaten, u. a. auch in Lehrprojekten, war ein kleiner, aber beliebter Bestandteil der deutschen DH-Community auf Twitter (Hermes et al., 2020). Daher haben sich die Entwickler des autoChirp-Service entschlossen, eine analoge Funktionalität für Mastodon zur Verfügung zu stellen (Projektname autodone).

Zielsetzung des Workshops

Der Praxis-Workshop verfolgt ein doppeltes Ziel: Er möchte zum einen Einstiegshürden abbauen und die DH-Community mit dem als kompliziert geltenden Fediverse und insbesondere mit dem DH-Server Fedihum vertraut machen. Zum anderen vermittelt er Kenntnisse für das automatisierte Posten bei Mastodon im Rahmen von Forschungs- und Lehrprojekten mit autodone.

Didaktischer Zugang und Ablauf des Workshops

Der halbtägige Praxisworkshop (4 h, inkl. 30 min. Pause) gliedert sich in drei Teile: Im ersten Teil (60 min) werden als Input in einem Rundgespräch der Hosts ggf. mit Special Guests grundlegende Fragen der Wissenschaftskommunikation in den sozialen Medien diskutiert, darunter: Welche Plattform eignet sich für welches kommunikative Ziel, was ist und wie funktioniert das Fediverse, wie finde ich dort meine Community, was soll/kann/darf ich posten, welche Sprache, welcher Stil sind angemessen, wie gehe ich mit negativen Kommentaren um, was ist rechtlich zu beachten und wie poste ich barrierefrei?

Die beiden anschließenden Teile dienen der praktischen Übung: Im zweiten Teil (75 min) werden grundlegende Einstellungen von Accounts durchgesprochen und ausprobiert sowie Schritte für Instanzwechsel und Import aus Twitter vorgestellt. Außerdem werden Tipps gegeben für die Sichtbarkeit von Tröts, Content Warnings, Threads, Alt-Text bei Bildern, Aufbau und die Pflege von Communities und für das Bespielen mehrerer Accounts. Abgerundet wird der Teil mit einer interaktiven Feedbackrunde zu Fedihum (Bedarfe, Serverlizenz). Im dritten Teil (75 min) geht es um die praktische Nutzung von autodone. Dabei wird auf die Erstellung eigener Projekte mit Diskussion von Themenwahl, Post-Aufbau, Bildern, Veröffentlichungsstrategien etc. eingegangen. Daten für autodone werden bereit gestellt, können aber auch selbst mitgebracht werden.

Die Workshopinhalte werden dokumentiert und für die spätere Nachnutzung online zur Verfügung gestellt.

Zielpublikum

Der Workshop richtet sich zugleich an Social-Media-Neulinge und Umsteigewillige von anderen Mastodon-Instanzen oder Social-Media-Plattformen. Er zielt auf die DH-Community und dort gleichermaßen auf Studierende wie auf Hochschullehrende, die Fedihum für sich, für Projekte oder in der Lehre einsetzen möchten. Es bestehen keine Voraussetzungen für die Teilnahme mit Blick auf technisches Vorwissen. Teilnehmende sind eingeladen, sich vorab einen Mastodon-Account anzulegen, wofür der Fedihum-Server zur Verfügung steht. Die Gruppengröße ist auf 20-25 Teilnehmende beschränkt.

Benötigte technische Ausstattung

Teilnehmende sollten ein eigenes Laptop oder Tablet mitbringen, die sich für die praktischen Anteile besser eignen als Smartphones.

Für die technische Raumausstattung werden Internetzugang und ein Beamer benötigt sowie ein Whiteboard.

Beteiligte und ihre Forschungsinteressen

Dr. Jürgen Hermes (Rolle nach CRediT: Writing – original draft, ID: 43ebbd94-98b4-42f1-866b-c930cef228ca), Geschäftsführer am Institut für Digital Humanities, Universität zu Köln. Wissenschaftliche Schwerpunkte: Konzeption und Entwicklung von Softwaretools für die Public Humanities; Einsatz computerlinguistischer Methoden in den Geisteswissenschaften. ORCID https://orcid.org/0000-0002-8367-8073
Institut für Digital Humanities, Universität zu Köln, Albertus Magnus Platz 1, 50923 Köln, hermesj@uni-koeln.de .

Dr. Mareike König (Rolle nach CRediT: Writing – original draft, ID: 43ebbd94-98b4-42f1-866b-c930cef228ca), stellvertretende Direktorin am Deutschen Historischen Institut Paris. Sie leitet das Blogportal de.hypotheses. Forschungsinteressen: Digitale Geschichtswissenschaft, Wissenschaftskommunikation mit sozialen Medien (insbes. Blogs), Open Science. ORCID https://orcid.org/0000-0002-8189-8574
Deutsches Historisches Institut Paris, 8, rue du Parc Royal, FR-75003 Paris, mkoenig@dhi-paris.fr .

Philip Schildkamp (Rolle nach CRediT: Software – ID: f89c5233-01b0-4778-93e9-cc7d107aa2c8) arbeitet mit den Schwerpunkten Systemadministration/-integration, Entwicklung und Provisionierung am Cologne Center for eHumanities (CCeH) & Institut für Digital Humanities (IDH). ORCID: https://orcid.org/0000-0003-0209-2837
Universität zu Köln, philip.schildkamp@uni-koeln.de .

Vivien Wolter (Rolle nach CRediT: Writing – review & editing, ID: d3aead86-f2a2-47f7-bb99-79de6421164d), Studentin der Digital Humanities an der Universität Trier, wissenschaftliche Hilfskraft am TCDH, Communication Fellow des DHd-Verbands und eine der Gründer:innen von Fedihum. Forschungsinteressen: Digitale Literatur- und Theaterwissenschaft und Wissenschaftskommunikation. ORCID: https://orcid.org/0000-0002-9982-9382
Universität Trier, vivien.wolter@gmx.net .

Prof. Dr. Ulrike Wuttke (Rolle nach CRediT: Writing - original draft, ID: 43ebbd94-98b4-42f1-866b-c930cef228ca), Professorin für Bibliothekswissenschaft - Strategien, Serviceentwicklung und Wissenschaftskommunikation an der Fachhochschule Potsdam. Sie ist DHd-Vorstandsmitglied (Beauftragte für Community und Communication) und eine der Gründer:innen von Fedihum. Forschungsinteressen: Open Science, Forschungsdaten und Wissenschaftskommunikation. ORCID: 0000-0002-8217-4025
Fachhochschule Potsdam, ulrike.wuttke@fh-potsdam.de .

Bibliographie

acatech - Deutsche Akademie der Technikwissenschaften et al. 2017. *Social*

Media und digitale Wissenschaftskommunikation. Analyse und Empfehlungen zum Umgang mit Chancen und Risiken in der Demokratie . München . https://www.leopoldina.org/uploads/tx_leopublication/2017_Stellungnahme_WOeM_web.pdf (zugegriffen: 18. Juli 2023).

Côté, Isabelle M. und Emily S. Darling . 2018. "Scientists on Twitter: Preaching to the choir or singing from the rooftops?" *FACETS* 3/1: 682–694. 10.1139/facets-2018-0002 (zugegriffen: 18. Juli 2023).

Geier, Andrea und Markus Gottschling . 2019. "Wissenschaftskommunikation auf Twitter? Eine Chance für die Geisteswissenschaften!" *Mitteilungen des Deutschen Germanistenverbandes* 66/3: 282–291. 10.14220/mdge.2019.66.3.282 (zugegriffen: 18. Juli 2023).

Haustein, Stefanie, Cassidy Sugimoto und Vincent Larivière . 2015. "Guest Editorial: Social Media in Scholarly Communication." *Aslib Journal of Information Management* 67/3. 10.1108/AJIM-03-2015-0047 (zugegriffen: 18. Juli 2023).

Hermes, Jürgen . 2021. "Chirpy Humanities." *Public Humanities* 5. Juli. https://publicdh.hypotheses.org/42 (zugegriffen: 18. Juli 2023).

Hermes, Jürgen, Øyvind Eide, Moritz Hoffmann, Alena Geduldig und Philip Schildkamp . 2017. "Twhistory mit autoChirp, Social Media Tools für die Geschichtsvermittlung." In *DHd 2017 Digitale Nachhaltigkeit. 4. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum"* . Bern. https://doi.org/10.5281/zenodo.4622681 (zugegriffen: 18. Juli 2023).

Hermes, Jürgen, Harald Klinke und Dennis Demmer. 2020. "Public Humanities Tools: Der Bedarf an niederschwelligen Services." In *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum"* . Paderborn. 10.5281/zenodo.4621842 (zugegriffen: 18. Juli 2023).

European Commission . 2018. *H2020 Programme Guidance: Social Media Guide for EU Funded R&I Projects. European Commission* . 4. Juni. ec.europa.eu/research/participants/data/ref/h2020/other/grants_manual/amga/soc-med-guide_en.pdf (zugegriffen: 18. Juli 2023).

Kirschenbaum, Matthew . 2012. "Digital Humanities As/Is a Tactical Term." In *Debates in the Digital Humanities*, hg. v. Matthew K. Gold: 416–417. Minneapolis: University of Minnesota Press.

König, Mareike . 2023. "Bei der nächsten Social-Media-Plattform wird alles anders. Oder: warum ich mich nicht bei Bluesky anmelde." *Zeitgeschichte Online* 13. November. https://zeitgeschichte-online.de/node/70045 (zugegriffen: 30. November 2023).

König, Mareike . 2022. "Tröten über Droysen: ein Mastodon-Leitfaden für Historiker:innen." *Digital Humanities am DHIP* 20. November. https://dhdhi.hypotheses.org/7205 .

König, Mareike und Paul Ramisch . 2022. "Die twitternde Zunft. Historikertage auf Twitter (2012–2018)." In *Digital History: Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft* , hg. v. Karoline Dominika Döring, Stefan Haas, Mareike König und Jörg Wettlaufer: 319-346. Berlin, Boston: De Gruyter Oldenbourg, 10.1515/9783110757101-017 (zugegriffen: 18. Juli 2023).

Mewburn, Inger. 2023. "The enshittification of academic social media." *The Thesis Whisperer* 7. Juli. https://thesiswhisperer.com/2023/07/10/academicenshittification/ (zugegriffen: 18. Juli 2023).

Quan-Haase, Anabel, Kim Martin und Lori McCay-Peet. 2015. "Networks of Digital Humanities Scholars: The Informational and Social Uses and Gratifications of Twitter." *Big Data & Society* 2/1. 10.1177/2053951715589417 (zugegriffen: 18. Juli 2023).

Seltmann, Melanie. 2023. "#PublicDH oder doch nur #WissKomm?" In *DHd2023: Open Humanities, Open Culture. 9. Tagung des Verbands " Digital Humanities im deutschsprachigen Raum "*. Belval/Trier. 10.5281/zenodo.7688632 (zugegriffen: 18. Juli 2023).

Sugimoto, Cassidy R., Sam Work, Vincent Larivière und Stefanie Haustein . 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.

Trapp, Markus . 2015. "Veränderungsmanagement bei der Implementation einer Social-Media-Strategie." *o-bib. Das offene Bibliotheksjournal* 2/4: 54–64. 10.5282/O-BIB/2015H4S54-64 (zugegriffen: 18. Juli 2023).

Wuttke, Ulrike . 2021. "#twitter101dh: Super-Experiment zu Twitter, Bibliotheken und COVID-19. Exploration der Twitter-Daten der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz und der Bayerischen Staatsbibliothek im Rahmen des vDHd2021-Events." *B.I.T. online* 24/4: 389–398. https://www.b-i-t-online.de/heft/2021-04-fachbeitrag-wuttke.pdf (zugegriffen: 18. Juli 2023).

Wuttke, Ulrike . 2022. "DHd-Verband initiiert mit Fedihum.org neue Mastodon-Instanz für alle Digital Humanities-Aficionados." *DHd: Digital Humanities im deutschsprachigen Raum* 14. November. https://dig-hum.de/aktuelles/dhd-verband-initiiert-fedihumorg-neue-mastodon-instanz-f%C3%BCr-alle-digital-humanities (zugegriffen: 18. Juli 2023).

not opaque flow – Workflows zur Aufbereitung und Auswertung historischer Dokumente

Weber, Dominic
dominic.weber@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz
ORCID: 0000-0002-9265-3388

Schwandt, Silke
silke.schwandt@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland
ORCID: 0000-0001-8303-4668

Huang, Angela
alhuang@fgho.eu
Forschungsstelle für die Geschichte der Hanse und des Ostseeraums, Europäisches Hansemuseum Lübeck, Deutschland
ORCID: 0000-0002-5321-9888

Hodel, Tobias
tobias.hodel@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz
ORCID: 0000-0002-2071-6407

Tolino, Serena
serena.tolino@unibe.ch
Institut für Studien zum Nahen Osten und zu muslimischen Gesellschaften, Universität Bern, Schweiz
ORCID: 0000-0001-7740-5805

Kuhlmann, Christopher
christopher.kuhlmann@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Meyer, Dana
dameyer@techfak.uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Wilde, Melvin
melvin.wilde@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Kirschnick, Inga
inga.kirschnick@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Jentsch, Patrick
p.jentsch@uni-bielefeld.de
Digital History, Universität Bielefeld, Deutschland

Hostettler, Myrjam
myrjam.hostettler@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz
ORCID: 0000-0001-9316-6330

Widmer, Jonas
jonas.widmer@unibe.ch
Digital Humanities, Walter Benjamin Kolleg, Universität Bern, Schweiz

Lange, Inga
ilange@fgho.eu
Forschungsstelle für die Geschichte der Hanse und des Ostseeraums, Europäisches Hansemuseum Lübeck, Deutschland

Popken, Vivien
vpopken@fgho.eu
Forschungsstelle für die Geschichte der Hanse und des Ostseeraums, Europäisches Hansemuseum Lübeck, Deutschland

Ausgangslage

Jüngste Entwicklungen im Bereich des *Deep Learning* haben große Fortschritte in vielen Anwendungsbereichen mit sich gebracht. Automatisches Erkennen von (Hand-)Schriften, das Identifizieren von Entitäten, Events und Beziehungen sowie *Topic Modeling* sind dabei für die *Digital Humanities* von besonderem Interesse. Während für moderne Sprachformen und Sprachen des globalen Nordens bereits eine Vielzahl von Lösungen existieren, müssen sie für vormoderne Sprachen, Sprachstufen und Sprachen des globalen Südens, noch ausgiebig getestet und angepasst werden. Auch den ihnen zugrunde liegenden Algorithmen und Datensätze müssen erweitert und kritisch untersucht werden.

Das Projekt ‘The Flow – From Deep Learning to Digital Analysis and the Role in the Humanities’ nimmt sich diesen Chancen und Herausforderungen aus einer dezidiert historischen Perspektive anhand vier sehr unterschiedlicher Korpora aus verschiedenen Zeiten und Räumen an und versucht generalisierbare und bewegliche Lösungen zu

entwickeln. Ausgehend von der an der Universität Bielefeld lancierten Applikation *nopaque* wird der gesamte Weg vom digitalisierten Quellenkorpus zu den für die Forschung nutzbaren Daten in niederschwellig, wiederverwendbare Workflows übertragen (Jentsch und Porada 2022). Das Ziel ist es, die Vorteile maschineller Lernverfahren für Geisteswissenschaftler:innen zugänglicher zu machen und ihre Herausforderungen und Grenzen in den historisch arbeitenden Geisteswissenschaften hervorzuheben. Gleichzeitig sollen die Ansätze transparent und kritisch diskutiert und somit das methodologische Instrumentarium der Fachwissenschaften geschärft werden.

Alle bereits erwähnten Schritte von der Quelle zu den Daten sind mit einigem Anpassungsaufwand und den entsprechenden Kenntnissen auf historische Korpora anwendbar (siehe Bspw. für Latein und Französisch Torres Aguilar und Stutzmann, 2021; Cafiero u. a., 2021). Layoutanalyse und Handschriftenerkennung werden von Transkribus, eScriptorium und ähnlichen Anwendungen abgedeckt (Muehlberger u. a., 2019; Kiessling u. a., 2019). Für (*Named) Entity Recognition* und *Event Extraction* existieren Programmbibliotheken und bereits trainierte Modelle, auf denen aufgebaut werden kann (Akbik u. a. 2019; Vasiliev 2020; Brunner u. a. 2020).

Unser Ausgangspunkt bildet nopaque, eine Anwendung der Universität Bielefeld zur Prozessierung von Textkorpora. Derzeit kann die Webanwendung nopaque aus einer Reihe von Digitalisaten ein Korpus erstellen, mit Tesseract OCR (Metzger und Weil, 2019) oder Transkribus, basierend auf TrHTR und pyLaia (Puigcerver [2017] 2022) den Text erkennen sowie mit spacy tokenisieren, lemmatisieren, Part-of-Speech-Tagging und Named Entity Recognition durchführen. Dies alles wird ermöglicht durch eine Applikation mit intuitiver graphischer Benutzeroberfläche, die über einen Browser aufgerufen werden kann.

Im Rahmen des Projekts “The Flow” werden die genutzten Algorithmen erweitert und zusätzliche Formen der Aufbereitung in nopaque integriert. Aktuell sind in unserem Projekt in der Quellenerschließung und -analyse drei Aufbereitungsschritte vorgesehen, die alle jeweils unabhängig voneinander stehen und als *open source* Pakete publiziert werden. Über die nopaque-Oberfläche werden alle Teile verbunden und über ein Jobmanagement zu einem Workflow kombiniert.

Prozess 1: OCR/HTR

Die neuesten Entwicklungen in *Computer Vision* und Texterkennung mit Transformers versprechen auch für vormoderne und nicht-lateinische Texte höhere Präzision und bessere Wiederverwendbarkeit von Modellen (Ströbel u. a. 2022). Dieses Versprechen gilt es im Rahmen des “The Flow”-Projekts zu überprüfen und große Modelle zu integrieren bzw. die Integration über HuggingFace zu erleichtern.

In diesem Arbeitsschritt soll insbesondere die Publikation von Ground Truth Daten mit dem Training (im Sinne von *Fine-Tuning* großer Modelle) verknüpft werden. Der Schritt resultiert daher in neuen HTR-Modellen für individuelle Handschriften und der Generierung einer Datenbasis für spezifische Sprachformen.

Prozess 2: Entity, Event und Relation Extraction

Automatisiertes Identifizieren von Entitäten, Ereignissen und Relationen sind klassische Aufgaben des *Natural Language Processing*. Für diese Aufgaben, wie auch für das Part-of-speech-Tagging sind stabile Language Models, die entweder enorm groß sind (im Sinne von Large Language Models) oder die spezifisch für eine bestimmte Domäne oder Zeit entwickelt wurden, erforderlich sind. Wiederum geht es um die Generierung großer Modelle und die Nachnutzung bestehender Modelle, die danach durch *Tagger* oder anderer Sequenz-zu-Sequenz Annotationsformen ausgezeichnet werden.

Für vormoderne Sprachen haben sich, wie für moderne Sprachstufen, insbesondere große oder domänenspezifische Sprachmodelle als hilfreich herausgestellt (Hodel, Prada Ziegler, und Schneider, 2023), sodass keine Normalisierungen mehr notwendig sind. Dies bedeutet gleichzeitig, dass Grundlagen (Sprachmodelle) erarbeitet und best-practices (beispielsweise für Annotationen) definiert werden müssen.

Prozess 3: Topic Modeling, Clustering und Vergleichsstudien

In diesem Schritt sollen auch die in den vorherigen Schritten trainierten Modelle für das thematische Clustern von Dokumenten innerhalb der jeweiligen Korpora verwendet werden. Auf Basis der Sprachmodelle können Abschnitte oder ganze Dokumente vektorisiert und mittels Clusteringverfahren miteinander verglichen werden. Die Resultate daraus können mit bereits intensiv genutzten Verfahren der Themenextraktion, insbesondere klassische Topic Modeling Verfahren mit LDA, abgeglichen werden (Graham, Weingart, und Milligan, 2012; Hodel, Möbus, und Serif, 2022; Schöch, 2017), was Vergleiche und die Kombination der Ergebnisse ermöglicht.

Dadurch wird ein Fundament gelegt, um die Quellenkritik, *close-reading* sowie Methoden und Standards der Geschichtswissenschaft wieder einzubringen und die methodologische Brücke zwischen *Digital* und *Humanities* zu schlagen.

Workflow Prozesse: Implementierung in *nopaque*

Aus den oben eingeführten Prozessen der Datenaufbereitung und -modellierung lassen sich wiederverwendbare Workflows entwickeln. Diese werden modular in *nopaque* implementiert, sodass sie auch von Forschenden eingesetzt werden können, die vorwiegend an der Anwendung interessiert sind.

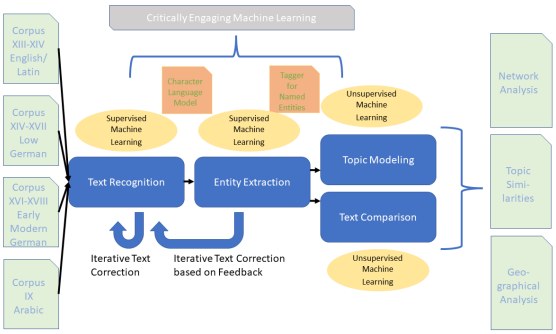


Abbildung 1: Schema der Workflows in einzelne Schritte zerlegt.

Die einzelnen Stationen dieser Workflows sind als Mikroprozesse zu verstehen, die jeweils eine bestimmte (nicht unbedingt kleine) Aufgabe erfüllen. Dazu gehört neben den oben eingeführten Prozessen beispielsweise eine Indizierung der Texte. Diese Technologie ermöglicht schnelleren und zuverlässigen Zugriff auf die Texte, wodurch das Trainieren und Anwenden von Machine-Learning-Modellen effizienter gestaltet werden kann.

Die Abläufe und die Interaktion zwischen den einzelnen Services werden über APIs (Programmierschnittstellen) abgewickelt, welche von einem von außen zugänglichen API-Gateway zusammengehalten und koordiniert werden. So werden unter anderem domänenspezifische Anpassungen und die Nachnutzbarkeit des Outputs in beliebige Analyse- und Auswertungsumgebungen vereinfacht. Das Zusammenspiel von Microservices, internen und externen APIs stellt die Skalierbarkeit sicher und erhöht die Flexibilität der Workflows.

Außerdem können so nach Bedarf zusätzliche Microservices zwischengeschaltet werden. Die intuitive graphische Benutzeroberfläche von nopaque muss dadurch nicht die einzige mögliche Anwendungsform bleiben. Die einzelnen Module sind zwar für die Anwendung in nopaque intendiert und optimiert, sind davon aber grundsätzlich unabhängig und können einzeln angesteuert werden.

Die modulare Struktur schafft großes Entwicklungspotenzial und bereitet *nopaque* auf längerfristige Weiterentwicklung und Anpassung an künftige technische Innovationen vor. Gleichzeitig kann auch das Projekt ‘The Flow’ derzeit noch offene Fragen in Zukunft einfacher in die Workflows einbinden. Dies betrifft beispielsweise die Anbindung an HPC-Clusters (High Performance Computing), was vor allem für aufwendige Prozesse wie das Training von HTR- und Sprachmodellen interessant ist.

Ziel des Workshops

Der Workshop verfolgt zwei Ziele: Erstens wird der aktuelle Stand von *nopaque* und seine Module der Community vorgestellt. Zweitens wird in einem geführten Brainstorming überlegt, welche Ansätze und Methoden in den Prozessen mitgedacht werden können. Das heißt, die Teilnehmenden haben die Möglichkeit sich in die weitere Entwicklung einzubringen und eigene Erfahrungen zu tauschen. Darum wird für den Workshop ein Call for Participation veröffentlicht (siehe Call anbei).

Ziel des Brainstorming-Teils ist es, die Diversität der historisch arbeitenden digitalen Geisteswissenschaften ernst zu nehmen und gemeinsam mit Critical Friends nach Ansätzen und Auswertungsformen zu suchen, die im Projektteam nicht abgedeckt sind und blinde Flecken darstellen. Aus diesem Grund sind Teilnehmende eingeladen, ihre Arbeitsweisen mit Auswertungs- und Analysetools zu demonstrieren.

Auf dieser Grundlage sollen in einem abschließenden Teil, die jeweiligen Implikationen für wiederverwendbare Workflows und mögliche Implementierungen und Integrationen in nopaque diskutiert werden. Ein besonderer Fokus soll dabei auf aufgrund ungewöhnlicher Layouts, Schriften oder Sprachen auftretender Bedürfnisse an Tools – und wo diese nicht bedient werden – liegen.

Format des Workshops

Der Workshop wird in vier Teilen plus Kaffeepause (3h30min) durchgeführt

00:00-00:30 Einführung und Vorstellungsrunde
00:30-01:30 Einführung in nopaque
01:30-01:50 Kaffeepause
01:50-02:20 Pitches der eingereichten Algorithmen und Methoden (aus CFP)
02:20-03:00 Brainstorming-Session
03:00-03:30 Schlussdiskussion

Zielpublikum und notwendiges Vorwissen:

Der Workshop richtet sich an Personen, die mit historischen Dokumenten arbeiten oder aus technischer Warte Prozessierung von Dokumenten anstreben oder verantworten. Vorwissen ist keines notwendig und die Einreichung eines eigenen Beitrags wird auch nicht vorausgesetzt.

Bibliographie

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, und Roland Vollgraf. 2019. „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP“. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4010.

Brunner, Annelen, 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 . CEUR-WS. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/11561.

Cafiero, Florian, Thibault Clérice, Paul Fièvre, Simon Gabay, und Jean-Baptiste Camps. 2021. „Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre“. *Journal of Data Mining & Digital Humanities* 2021 (Februar). https://doi.org/10.46298/jdmdh.6485.

Graham, Shawn, Scott Weingart, und Ian Milligan. 2012. „Getting Started with Topic Modeling and MALLET“. *Programming Historian*. https://doi.org/10.46430/phen0017.

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*, herausgegeben von Selin Gerlek, Sarah Kissler, Thorben Mämecke, Dennis Möbus, Jennifer Eickelmann, Katrin Köppert, Peter Risthaus, und Florian Sprenger, 1. Auflage, 1:181–205. Digitale Kultur. Hagen: Hagen University Press. https://doi.org/10.57813/20220623-153139-0.

Hodel, Tobias, Ismail Prada Ziegler, und Christa Schneider. 2023. „Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern“. Graz, Austria, Juni 30. https://doi.org/10.5281/zenodo.8107616.

Jentsch, Patrick, und Stefan Porada. 2022. „nopaque“. nopaque | from text > to data > to analysis. 2022. https://nopaque.sfb1288.uni-bielefeld.de/.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, und Daniel Stökl Ben Ezra. 2019. „eScriptorium: An Open Source Platform for Historical Document Analysis“. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2:19–19. https://doi.org/10.1109/ICDARW.2019.10032.

Metzger, Noah, und Stefan Weil. 2019. „Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow“. Workshop gehalten auf der MAD HD, Heidelberg, Juli 30.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, u. a. 2019. „Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study“. *Journal of Documentation* 75 (5): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Puigcerver, Joan. (2017) 2022. „PyLaia“. Python. https://github.com/jpuigcerver/PyLaia.

Schöch, Christof. 2017. „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“. *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Ströbel, Phillip Benjamin, Simon Clematide, Martin Volk, und Tobias Hodel. 2022. „Transformer-based HTR for Historical Documents“. Preprint. ArXiv: https://doi.org/10.48550/arXiv.2203.11008.

Torres Aguilar, Sergio, und Dominique Stutzmann. 2021. „Named Entity Recognition for French medieval charters“. In *Workshop on Natural Language Processing for Digital Humanities*. Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop. Helsinki, Finland. https://hal.archives-ouvertes.fr/hal-03503055.

Vasiliev, Yuli. 2020. *Natural language processing with Python and spaCy: a practical introduction*. San Francisco: No Starch Press.

Offen – frei zugänglich – für alle? Partizipative Ansätze zum barrierefreien Umgang mit Forschungsdaten

Wunsch, Samuel
wunsch@iib.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland

Lehnen, Katrin Anna
lehnen@ub.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland

Henzel, Katrin
henzel@ub.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland
ORCID: 0000-0001-8260-223X

Christ, Andreas
christ@ub.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland
ORCID: 0000-0002-3591-2355

Zielsetzung des Workshops

Open Science, Open Humanities, Open Data – bei den Themen Offene Wissenschaft und Offene Daten wird meist implizit vorausgesetzt, dass die Zugangs- und Nachnutzungsmöglichkeiten für alle Nutzenden unterschiedslos gleich sind, solange die Daten nur frei zugänglich sind. Bei genauerer Betrachtung zeigt sich allerdings, dass dies nicht zutrifft. Vielmehr bestehen vielfach Barrieren bei der Beschaffenheit der Daten selbst als auch bei den sie verarbeitenden technischen Systemen und sogar bei den Möglichkeiten zum Erwerb von Datenkompetenzen. Im Sinne der gesellschaftlichen Teilhabe und Chancengleichheit (ebenfalls zentralen Paradigmen von Open Science) besteht daher für Datenzentren wie Forschende gleichermaßen die Notwendigkeit, Gestaltung, Zugang und Nutzung von Forschungsdaten um partizipative und inklusive Aspekte zu erweitern.

In diesem Workshop möchte die AG Datenzentren in Kooperation mit dem Institut für Inklusive Bildung (IIB) der Christian-Albrechts-Universität zu Kiel bei Forschenden und Forschungsdatenmanager*innen für die bislang noch wenig adressierte Inklusion im Datenmanagement sensibilisieren und ein Bewusstsein für das Thema schaffen. Im Mittelpunkt des Workshops stehen die praktische Erprobung und gemeinsame Bewertungen digitaler Zugänglichkeit. Konkret auftretende Hürden bei der Bereitstellung und Verwaltung sowie Nutzung von Datenrepositorien sollen im Workshop identifiziert und mögliche Lösungsansätze gemeinsam erarbeitet und diskutiert werden.

Darüber hinaus wird der Workshop um die Vorstellung und Diskussion der Ergebnisse eines explorativ angelegten Forschungsprojekts bereichert: Im Rahmen des Digitalisierungsprogramms des Landes Schleswig-Holstein[1] wurde von November 2022 bis März 2023 das Projekt „Forschungsdatenmanagement inklusiv und partizipativ“ an der Christian-Albrechts-Universität zu Kiel gefördert, in dem zwei Fokusgruppen aktiv in die Evaluierung von Datenportalen bzw. -repositorien einbezogen wurden: Ein darauf aufbauendes Folgeprojekt ist in der gleichen Förderlinie im Juli 2023 gestartet. Die hier gewonnenen Erkenntnisse bezüglich subjektiv wahrgenommener Barrieren beim Navigieren durch die Portale zeigen auf, wo besonderer Handlungsbedarf besteht.

Neben dem Erfahrungsbericht werden im Workshop zudem bewährte Praktiken und pädagogische Lösungsansätze vorgestellt, um Menschen mit Behinderungen eine aktive Teilhabe an Forschungsprojekten und -prozessen zu ermöglichen und die Zugänglichkeit von Forschungsdaten zu verbessern. Diese Ansätze basieren auf aktuellen Erfahrungen und können als Grundlage für weitere Diskussionen und konkrete Maßnahmen dienen. Von zentraler Bedeutung ist dabei die Partizipation. Partizipative Forschung hat „die Beteiligung von gesellschaftlichen Akteuren als Co-Forscher/innen sowie Maßnahmen zur individuellen und kollektiven Selbstbefähigung und Ermächtigung der Partner/innen (Empowerment)“ (von Unger, 2014: 1) zum Ziel. Das bedeutet auch, dass „[w]ichtige Entscheidungen […] gemeinsam in der Forschungsgruppe getroffen“ werden (Schwörer et al., 2022: 225). Das gilt konsequenterweise auch für das Datenmanagement. Partizipation bedeutet demnach größtmögliche Transparenz für alle Projektbeteiligten in allen Entscheidungen zum Datenmanagement in allen Phasen des Forschungsdatenlebenszyklus (vgl. Henzel, 2022: 86). Um Menschen mit Behinderungen einen umfassenden Zugang zu Daten zu ermöglichen und sie als Mitgestaltende aktiv einbinden zu können, bedarf es daher einer Umgestaltung der Infrastrukturlandschaft für Forschung, Lehre und Management: Barrieren des Datenzugangs wie der technischen Systeme sind abzubauen oder zu vermeiden, Awareness ist zu schaffen. Dieses Ziel stellt die praktische Ausgestaltung von Artikel 9 der UN-Behindertenrechtskonvention (United Nations, 2006) dar, der die Bedeutung der Barrierefreiheit von Informationen und Kommunikationstechnologien betont, um Menschen mit Behinderungen einen gleichberechtigten Zugang zu ermöglichen.

Status Quo eines inklusiven Forschungsdatenmanagements

Ein in die Breite wirkendes inklusives Forschungsdatenmanagement stellt derzeit noch ein Desiderat dar. Es gibt aber Initiativen, Gruppen und Verbände, die zu diesem Thema arbeiten, konkreten Support anbieten und so Awareness schaffen. In erster Linie sind hier die AG „Barrierefreiheit in Bibliotheken“[2] in der Vereinigung österreichischer Bibliothekarinnen und Bibliothekare (VÖB) sowie das „Netzwerk für Repositorienmanager*innen“ (RepManNet)[3] zu nennen. Mittels Publikationen, Dokumentationen und Handreichungen – hier sind v. a. die Guidelines von Oxford/Woodbrook (2023) zu nennen – wird so ein praktisches Angebot für die Umsetzung rechtlicher Vorgaben von Barrierefreiheit geschaffen. Dies betrifft insbesondere die barrierefreie Gestaltung von (Meta-)Daten, graphischen Nutzeroberflächen (v. a. Webseiten) und Repositorien (Blumesberger et al., 2022; Blumesberger, 2019). In diesem Kontext relevante Tools zur Messung eines barrierefreien Webdesigns werden gesammelt und vorgestellt (etwa Andrae et al., 2020: 270f.). Sich die Entwicklungen des Forschungsdatenmanagements in Österreich zum Vorbild nehmend, hat sich auch in Deutschland mit der AG „Inklusion im Forschungsdatenmanagement“ im Verbund von GO- Unite! zwischenzeitlich eine Gruppe formiert, die sich zum Ziel gesetzt hat, „den Zugang zu und den Umgang mit Forschungsdaten barrierearm zu gestalten“ und „Inklusion als Teil des FAIRen Umgangs mit Daten sichtbar [zu] machen.“[4] Die AG beteiligte sich mit einer Stellungnahme aus der öffentlichen Konsultation zum Forschungsdatengesetz (AG Inklusion im Forschungsdatenmanagement im Verbund von GO-Unite!,

2023), zu der das BMBF im Frühjahr 2023 aufgerufen hatte. Wünschenswert sind darüber hinaus weitere Angebote in Form von Best-Practice-Beispielen und User Stories, um der Herausforderung aus der Perspektive des Forschungsdatenmanagements zu begegnen.

Ein grundlegendes Defizit besteht allerdings bereits darin, dass es keine belastbaren Aussagen zum Stand der Umsetzung digitaler Barrierefreiheit von Datenportalen und - repositorien gibt und bisher wenige Repositorien überhaupt darauf systematisch getestet wurden. Eine dieser Ausnahmen bildet das Qualitative Data Repository (QDR)[5] an der Syracuse University (NY), das einem intensiven Barrierecheck unterzogen wurde. Eine ausführliche Beschreibung der Vorgehensweise, der festgestellten Defizite und weiterer nötiger Schritte hin zu einem inklusiven Forschungsdatenmanagement liefern Anderson et al. (2022). Auch sie kommen zu dem Schluss, dass ein barrierefreies Datenmanagement nur gemeinsam mit Betroffenen zu gestalten sei: „It is therefore indispensable to work with disabled users, testers, and, if possible, developers to ensure that any measures taken are successful." (Anderson et al., 2022: 7)

Mit dem fehlenden empirischen Material zur digitalen Zugänglichkeit zu Datenrepositorien geht das Problem einher, subjektiv wahrgenommene Hürden nicht einheitlich messen zu können. Zur Erhebung quantitativer Daten mithilfe standardisierter Verfahren gibt es u. a. Software (etwa für die Optimierung von Farbkontrasten), die die rechtlichen Vorgaben und daraus abgeleitete Richtlinien zur Barrierefreiheit von Weboberflächen prüfen kann. Individuell auftretende Barrieren lassen sich damit aber nur sehr eingeschränkt feststellen. Insbesondere das in den Web Content Accessibility Guidelines (WCAG) 2.1 (W3C, 2018: #understandable) formulierte Prinzip der Verständlichkeit (neben Wahrnehmbarkeit, Operabilität und Robustheit das dritte von insgesamt vier Prinzipien für Web-Zugänglichkeit) lässt sich nicht einheitlich messen. Hier bedarf es also anderer Methoden zum Testen digitaler Zugänglichkeit, nämlich partizipativ gestalteter Verfahren, wie sie auch im Workshop erprobt werden sollen.

Ablauf des Workshops

Der Workshop ist für einen halben Tag geplant und gliedert sich in vier Teile:

Dauer		Inhalt/Thema
15‘		Begrüßung und Ankommen, Programmvorstellung
30‘		Teil 1: Einführung in die rechtlichen Rahmenbedingungen digitaler Zugänglichkeit und Diskussion zu deren Umsetzungsmöglichkeiten
30‘		Teil 2: Erfahrungsbericht (Samuel Wunsch) und anschließende Diskussion zu partizipativen Formaten im Datenmanagement und in Forschungsprojekten
15‘		Pause
45‘		Teil 3: Testen eines Datenrepositoriums auf Verständlichkeit, in Gruppenarbeit mit Leitfaden
30‘		Zusammentragen der Beobachtungen (Dokumentation auf Pinnwänden) und Diskussion im Plenum
15‘		Pause
45‘		Teil 4: Zusammenführen der Ergebnisse der Gruppenarbeit (Teil 2) mit dem vorgestellten Erfahrungsbericht (Teil 3) für gemeinsame Diskussion von partizipativen Lösungsansätzen und weiteren Schritten für eine Anpassung des untersuchten Datenrepositoriums, schriftliche Fixierung der Diskussionsergebnisse
15‘		Zusammenfassung, Evaluierung, Verabschiedung

Im ersten Teil des Workshops werden Bedeutung und Relevanz von Artikel 9 (Zugänglichkeit) der UN-Behindertenrechtskonvention für die Nutzung und Verwaltung von Forschungsdaten diskutiert. Es erfolgt eine Einführung in die rechtlichen Rahmenbedingungen (v. a. EAA, BFSG, EN 301 549, BITV 2.0) und aktuellen Entwicklungen von Barrierefreiheit in Deutschland, wobei auch ein kritischer Blick auf bestehende Fördervorgaben, Richtlinien und Empfehlungen des Datenmanagements für die Umsetzung der rechtlichen Vorgaben geworfen werden soll. Gemeinsam diskutiert werden im Anschluss mögliche Lösungsansätze auf institutioneller Ebene. Dabei steht die Frage im Mittelpunkt, wie vor Ort unter jeweils unterschiedlichen lokalen Rahmenbedingungen partizipative Ansätze eines inklusiven Datenmanagements auf hochschulinstitutioneller Ebene umgesetzt werden können. In diesem Kontext wird ein mögliches Modell vorgestellt: Mit dem Institut für Inklusive Bildung (IIB) besitzt die Kieler Universität eine zentrale Einrichtung, die Menschen mit Behinderungen zu Bildungsfachkräften qualifiziert hat, welche nunmehr als Expert*innen in eigener Sache vielfältige Bildungsangebote realisieren sowie Beratung, Unterstützung und Vernetzung für Menschen in Fach- und Hochschulen, Politik, Verwaltung, Verbänden und Unternehmen, die Inklusion umsetzen wollen, anbieten. Einen Impuls aus der Praxis liefert im zweiten Teil des Workshops Samuel Wunsch, Bildungsfachkraft des Kieler IIB. Er berichtet über seine persönlichen Erfahrungen in partizipativ ausgerichteten Projekten (s. o.). Im Mittelpunkt des Berichts stehen die Identifizierung von Barrieren und Herausforderungen bei der Erstellung, Bearbeitung und Nachnutzung von Forschungsdaten, die untrennbar mit Fragen des Projekt- und Teammanagements sowie der Wissenschaftskommunikation verbunden sind. Der Erfahrungsbericht soll den Teilnehmenden des Workshops als Grundlage für eine wei-

tergehende Diskussion zu konkreten Lösungsansätzen für die Zugänglichkeit von Forschungsdaten insbesondere in projektbezogener Arbeit dienen. Die Teilnehmenden des Workshops sollen dazu ermuntert werden, eigene Erfahrungen bei der Umsetzung digitaler Zugänglichkeit zu teilen. Der dritte Teil des Workshops beinhaltet das Testen von Zugänglichkeit von Forschungsdatenportalen in Kleingruppen. Ein Repositorium[6] für Daten aus den Geisteswissenschaften soll hierbei gezielt auf Verständlichkeit untersucht werden. Eine Checkliste mit zu berücksichtigenden Kriterien wird hierfür bereitgestellt. Auftretende Hürden und subjektiv wahrgenommene Barrieren gilt es in der Gruppenarbeit zu dokumentieren. Anschließend werden die Ergebnisse zusammengetragen und im Plenum gemeinsam diskutiert. Im abschließenden vierten Teil des Workshops wird gemeinsam nach Lösungen und weiteren Schritten für eine Anpassung des untersuchten Datenrepositoriums gesucht. Die im Impulsvortrag (zweiter Teil des Workshops) vorgestellten Erfahrungen und Ergebnisse werden dabei vergleichend herangezogen, um den Aspekt des Partizipativen noch stärker zu machen. Die Ergebnisse aus dem vierten Teil des Workshops sollen in Form eines Berichts unter aktiver Einbindung interessierter Teilnehmender des Workshops in den DHdBlog[7] überführt werden.

Workshop-Format und Ausstattung

Halber Tag/4 h inkl. Pausen. Benötigte Materialien und Ausstattung: Beamer, Leinwand, WLAN, Moderationskoffer, Magnet- oder Pinnwand, für die Gruppenarbeit wäre ein zweiter Raum ideal

Zielgruppe

Der Workshop richtet sich an alle Interessierten aus Forschung, Lehre, Studium und Forschungsdatenmanagement, die gemeinsam mit qualifizierten Menschen mit Behinderungserfahrungen bestehende Probleme der digitalen Zugänglichkeit von Forschungsdatenrepositorien für die Geisteswissenschaften diskutieren und konstruktive Lösungsansätze für ein inklusives Datenmanagement erproben möchten. Für die Teilnahme ist kein Vorwissen nötig, es zählt das Interesse am Thema. Für die praktische Arbeit ist ein eigener Laptop wünschenswert.

Die Zahl der Teilnehmenden soll aus methodischen Gründen 15 nicht überschreiten.

Zusammensetzung des Teams

Samuel Wunsch, Bildungsfachkraft am Institut für Inklusive Bildung der Christian-Albrechts-Universität zu Kiel, Bildungsfachkräfte sind sowohl Botschafter*innen für Inklusion als auch Expert*innen für Behinderungserfahrungen

Katrin Anna Lehnen, Mitarbeiterin im Projekt „Inklusives Datenmanagement partizipativ gestaltet“, pädagogisch wissenschaftliche Mitarbeiterin im Bereich Inklusionstheorie und -praxis und Kompetenzforschung an der Christian-Albrechts-Universität zu Kiel

Katrin Henzel, Mitarbeiterin des Bereichs Digital Humanities & Forschungsdaten an der Universitätsbibliothek und im Zentralen Forschungsdatenmanagement der Christian-Albrechts-Universität zu Kiel, Sprecherin der AG Inklusion im Forschungsdatenmanagement im Verbund von GO-Unite!

Andreas Christ, Leitung Digital Humanities & Forschungsdaten an der Universitätsbibliothek und Mitkoordinator der universitären Forschungsdatenmanagement-Services der Christian-Albrechts-Universität zu Kiel, stellv. Convenor der AG Datenzentren im DHd

Fußnoten

1. https://www.schleswig-holstein.de/DE/landesregierung/themen/digitalisierung/digitalisierung-zukunftsthema/digitalisierung-zukunftsthema_node.html (zugegriffen: 19. Juli 2023).
2. https://voeb-b.at/voeb-kommissionen/ag-barrierefreiheit-in-bibliotheken#section1 (zugegriffen: 19. Juli 2023).
3. https://datamanagement.univie.ac.at/forschungsdatenmanagement/netzwerk-fuer-repositorienmanagerinnen-repmannet/ (zugegriffen: 19. Juli 2023).
4. https://go-unite.de/index.php/ag-inklusion-im-forschungsdatenmanagement/ , Abs. 1 (zugegriffen: 19. Juli 2023).
5. https://qdr.syr.edu/ (zugegriffen: 19. Juli 2023).
6. Die Auswahl des zu untersuchenden Forschungsdatenrepositoriums erfolgt kurz vor der Durchführung des Workshops, da Repositorien i. d. R. immer wieder Änderungen unterliegen und das gewählte Repositorium möglichst passgenau auf die Bedarfe des Workshops (Ziele und v. a. Methodik) abgestimmt werden soll.
7. https://dhd-blog.org/ (zugegriffen: 19. Juli 2023).

Bibliographie

AG Inklusion im Forschungsdatenmanagement im Verbund von GO-Unite! 2023. *Öffentliche Konsultation zum Forschungsdatengesetz – Fragebogen*, hg. vom Bundesministerium für Bildung und Forschung, 10.03.2023. https://www.bmbf.de/bmbf/shareddocs/downloads/files/Forschungsdatengesetz/AGInklusion.html (zugegriffen: 19. Juli 2023).

Anderson, Theresa, Randy D. Colón, Abigail Goben und Sebastian Karcher. 2022. “Curating for Accessibility.” *International Journal of Digital Curation* 17/1. 10.2218/ijdc.v17i1.837.

Andrae, Magdalena, Susanne Blumesberger, Sonja Edler, Julia Ernst, Sarah Fiedler, Doris Haslinger, Gerhard Neustätter und Denise Trieb. 2020. „Barrierefreiheit für Repositorien. Ein Überblick über

technische und rechtliche Voraussetzungen.“ *Mitteilungen der VÖB* 73/2: 259–277. 10.31263/voebm.v73i2.3640.

Blumesberger, Susanne. 2019. „Barrierefreiheit und Repositorien – Nachdenken über Open Science für alle.“ *b.i.t. online* 22/4: 297–302.

Blumesberger, Susanne, Sonja Edler, Eva Gergely, Doris Haslinger und Denise Trieb. 2022. *Guidelines zur Erstellung barrierearmer Inhalte für Repositorien* . Wien. https://phaidra.univie.ac.at/o:1430148 (zugegriffen: 19. Juli 2023).

Henzel, Katrin. 2022. „Vermittlung auf Augenhöhe – digitale Editionen inklusiv gestaltet.“ *editio* 26: 72–88. 10.1515/editio-2022-0003.

Oxford, Emily und Rachel Woodbrook. 2023. *Accessibility Data Curation Primer. Data Curation Network* . Retrieved from the University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/253392 (zugegriffen: 19. Juli 2023).

Schwörer, Laura, Hannah von Ledden, Pia Algermissen und Mandy Hauser. 2022. „Zusammenarbeit und Mediennutzung in einer Partizipativen Forschungsgruppe“. In *Grenzen.Gänge.Zwischen.Welten. Kontroversen – Entwicklungen – Perspektiven der Inklusionsforschung* , hg. von Bernhard Schimek, Gertraud Kremsner, Michelle Proyer, Rainer Grubich, Florentine Paudel und Regina Grubich-Müller, 223–230. 10.35468/5924-24. Bad Heilbrunn: Julius Klinkhardt.

von Unger, Hella. 2014. *Partizipative Forschung. Einführung in die Forschungspraxis* . Wiesbaden: Springer. 10.1007/978-3-658-01290-8.

United Nations. 2006. *United Nations Convention on the Right of Persons with Disabilities* . 06.12.2006. https://www.un.org/esa/socdev/enable/rights/convtexte.htm (zugegriffen: 19. Juli 2023).

W3C. 2018. *Web Content Accessibility Guidelines (WCAG) 2.1*. 05.06.2018. https://www.w3.org/TR/WCAG21 (zugegriffen: 19. Juli 2023).

Producing & sparqling Open and FAIR data with the Geovistory environment

Hart, Stephen
stephen.hart@unibe.ch
Universität Bern, Schweiz

Knecht, David
david.knecht@kleiolab.ch
KleioLab GmbH, Schweiz

Beretta, Francesco
francesco.beretta@cnrs.fr
LARHRA–CNRS/Université de Lyon/ENS, Frankreich

Schneider, Jonas
jonas.schneider@kleiolab.ch
KleioLab GmbH, Schweiz

Context

Digitization of research becomes an increasingly important topic in scientific disciplines. Agencies like SNSF, ANR, and Horizon Europe ask funded research projects to detail how the produced data respond to the FAIR principles, as well as the publication of research data and metadata in public repositories so that it can be found, accessed and reused (see Swiss National Science Foundation, "Open Research Data."; European Commission, "Open Data, Software and Code Guidelines."; ANR, "La science ouverte."). Those requirements are in line with the movement of Open Science that advocates for better accessibility of scientific research (especially publicly funded) so that knowledge is easily shared with scientists (as well as the rest of society), thus improving the quality, efficiency and responsiveness of research (see UNESCO, "UNESCO Recommendation on Open Science."). Open data is one of the pillars of this movement.

Such requirements, as well as the increasing awareness and presence of digitization in many aspects of academia, are putting pressure on researchers, who need to learn and understand the principles and standards of FAIR data and its impact on research data, but also require them to acquire new methods and knowledge, such as data management workflows and best practices.

Technologies of Open Data, such as RDF, also allow the analysis across multiple datasets through complex research queries, which gives researchers the possibility to ask a wider number of questions to their research data. Acquiring the new technical skills needed to manipulate those technologies would allow researchers to mobilize those new tools and therefore fully benefit from Open Data.

The goal of the Geovistory (http:// geovistory.org) environment is to alleviate the weight for researchers, by providing tools and practices that are easy to use, help them structure their research data according to the FAIR principles, as well as also allow them to mobilize the full potential of their data for research with digital analysis tools.

About Geovistory

Geovisotry is an attractive virtual research and data publication environment designed to strengthen Open Research Data practices. Geovistory is developed for research

projects in the humanities and especially in history according to the participatory method of "user experience design". Geovistory supports researchers with simple and easy-to-use interfaces and allows them to make their research accessible in an attractive way to people interested in history. Geovistory includes:

- The Geovistory Toolbox, which allows to manage and curate projects' research data. The Toolbox is freely accessible for all individual projects. Each research project works on its own data perspective but at the same time automatically contributes to a joint knowledge graph.
- The Geovistory Publication platform Geovistory (http:// geovistory.org), which allows to access data as an external user via the community page or project-specific webpages and its graphical search function or the SPARQL-endpoint.

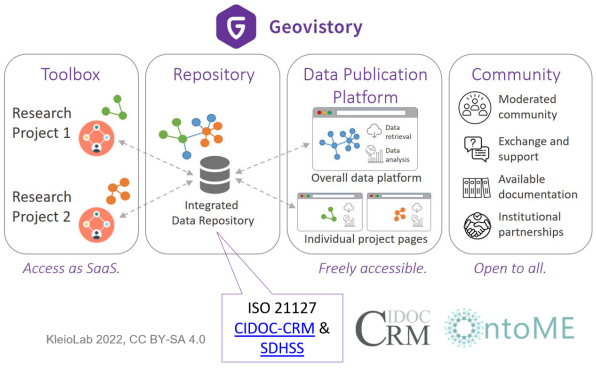


Fig. 1: Geovistory components.

A particular strength of Geovistory is its handling of the challenges of scientific information in the Humanities and Social Sciences: The context-sensitive nature of information and its relation to different research agendas, the wide variations in meaning for the same terms and vocabulary complexities, competing views or gaps and fragmentation of available information. This is what makes the Geovistory graph data model and semantic enrichment capabilities so interesting for the Humanities and Social Sciences research.

The data model is collaboratively managed in the ontome.net (http://ontome.net) application and takes advantage of the ontology ecosystem of the Semantic Data for Humanities and Social Sciences project (https://ontome.net/ns/sdhss/), relying on CIDOC CRM as a high-level ontology. The robust and modular modeling method allows extending the ontology easily and thus using Geovistory as a virtual research environment to researchers in different Humanities and Social Sciences disciplines.

As per current terms of service, all data produced in the information layer of Geovistory are licensed under creative commons BY-SA 4.0. The different infrastructure components are jointly developed by KleioLab, LARHRA, the University of Bern, and welcome other actors joining the Geovistory vision. All the web components and the publication platform have been made available as open source, and before the end of 2023, the toolbox and other components will be made open. The LOD4HSS project (https://www.geovistory.org/lod4hss), co-funded by swissuniversities, structures these efforts.

Geovistory within the DH ecosystem

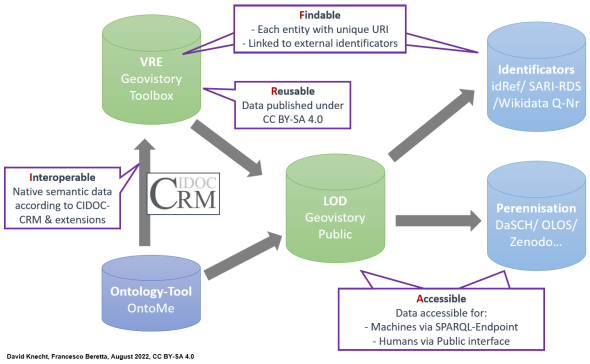


Fig. 2: Geovistory and the FAIR Principles.

First, Geovistory does not predefine the specific data model but rather consumes it via an API from the community-driven ontology management environment OntoMe (see elaborations above), allowing data semantification and community-led interoperability.

Second, the Geovistory ecosystem is connected dynamically to the information systems of producers of authority records (such as IdRef, GND) and other data repositories (such as Wikidata) in view of interconnecting bibliographic information systems and scale up to a large knowledge graph. In this area, a data exchange pipeline has been built with the French Agence Bibliographique de l'Enseignement Supérieur (ABES) and further cooperations are currently being set up.

Third, long-term data storage solutions for projects that concluded active research is important. This can be done in the Zenodo repository. Ideally, cooperation with DaSCH, OLOS and Huma-Num would be established allowing dynamic updates, with first exchanges with DaSCH underway.

About the Academic Education & Careers project

The project *Academic Education & Careers* (https://www.geovistory.org/project/1483135) is a collaborative project enriched by the Geovistory community. This makes this project an ideal showcase of the potential of Geovistory, both in terms of the collaboration philosophy and in terms of data management.

The project *Academic Education & Careers* uses the Geovistory infrastructure and has been initiated at the vDHd (German Digital Humanities Conference) in 2021, as part

of a collaborative experiment. In this experiment, the first dataset, the directory of scholars at the University of Kiel in Germany, has been added. It was further enriched with data from the Siprojuris project on French law professors (1804-1950) (http://siprojuris.symogih.org/). Siprojuris has been constituted by the career files of law teachers, mainly held by the Archives Nationales and the Centre des Archives Contemporaines de Fontainebleau, as well as civil status records and printed sources, primarily obituaries published in the local, national, or specialist press.

Since then, the project *Academic Education & Careers* has become an open and collaborative **community project** on the history of science and universities. It documents information about professors from different universities and their academic careers, ranging from their graduation to their professorship. The project aims to deepen the understanding of the individual careers of academic professors, enabling researchers to place the intellectual output of these scholars in a broader perspective, comparing social, religious, and political dimensions. By providing bibliographic information to better link the academic world with the trajectories of its actors, the project Academic Education & Careers sheds light on the mechanisms behind the production of academic knowledge.

This project is a collaborative endeavor. In other words, it is open and available to all users of Geovistory and has two purposes:

- To allow users to discover Geovistory Public as well as the(?) Geovistory Toolbox, explore the various entities, sources, and texts, and understand how the knowledge graph is structured in the data management tools. Generic pre-written queries are also available to showcase the analysis of temporal and geographical data;
- To encourage users to contribute to the project, either by adding new entities, enriching existing information, or adding new corpora to the knowledge graph.

About the workshop

The workshop is planned as a tutorial based on the community project “Academic Education & Careers” and is open to all participants with an interest in learning about Open Research Data practices within the Geovistory environment. Participants should bring their own laptops. The workshop has two objectives:

- Introduce participants to the principles of FAIR and Open Research Data management, so that participants can understand the usefulness of following those standards, not only for the publication of data and its potential reuse but also for how it can help the researcher throughout the research cycle.
- Give participants a concrete hands-on experience of the Geovistory Research environment (Geovistory Toolbox, Project Webpage, and SPARQL-endpoint) from data production, to publication, to analysis and reuse.

For this, the workshop is structured as follows:

- Introduction to Geovistory and the *Academic Education & Careers* community project: context & history of the project and of the Geovistory platform, experiences & results of the first phase of the project in the first workshop vDHd21 and how the current workshop will build upon it.
- Presentation on the standards and principles of FAIR data, and the general Open Data Research practices, and how it can affect Humanities research data and how researchers can benefit from it.
- Introduction on ontologies for the humanities, with the ICOM standard CIDOC-CRM (ISO 21127:2014), as well as the SDHSS extensions for the Humanities. The goal is to help the participants understand the underlying principles and standards that structure Geovistory and apprehend the semantics behind the data model with the help of examples of the community project. This understanding will be strenghen during the hands-on part of the workshop.
- Hands-on insight into the Geovistory environment, and especially the toolbox, by using sample data from the *Academic Education & Careers* community project, and following the work process of data ingestion, structuration, enrichment, and linking with useful resources (from the Geovistory shared knowledge graph but also external resources, such as the GND, Getty vocabularies, etc.).
- Hands-on introduction on the possibilities of data analysis within the Toolbox, but also on the public project knowledge graph via SPARQL endpoint.
- Open discussion to reflect on the workshop based on the impressions gained and the participant’s own experiences, and on what Geovistory can bring to the participant’s research.

Bibliographie

Swiss National Science Foundation, "Open Research Data." Swiss National Science Foundation. Accessed July 19, 2023. https://www.snf.ch/en/dMILj9t4LNk8NwyR/topic/open-research-data

European Commission, "Open Data, Software and Code Guidelines." Open Research Europe. Accessed July 19, 2023. https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#standardsandfair

ANR, "La science ouverte." Agence nationale de la recherche. Accessed July 19, 2023. https://anr.fr/fr/lanr/engagements/la-science-ouverte/

UNESCO, "UNESCO Recommendation on Open Science." UNESCO. Accessed July 19, 2023. https://www.unesco.org/en/open-science/about?hub=686

Uncovering the Forgotten Bits: Perspektiven von Retrocomputing und Emulation für die DH

Roeder, Torsten
torsten.roeder@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland
ORCID: 0000-0001-7043-7820

Leitgeb, Johannes
johannes.leitgeb@stud-mail.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland
ORCID: 0009-0006-2058-9133

Marenec, Madlin
madlin@mrsmuseum.de
Julius-Maximilians-Universität Würzburg, Deutschland
ORCID: 0009-0003-7434-827X

Shtohryn, Tomash
tomash.shtohryn@stud-mail.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland
ORCID: 0009-0000-4597-603X

Herbst, Yannik
yannik.herbst@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland
ORCID: 0000-0002-6547-9599

Die prekäre Lage des digitalen Kulturerbes in der Gegenwart

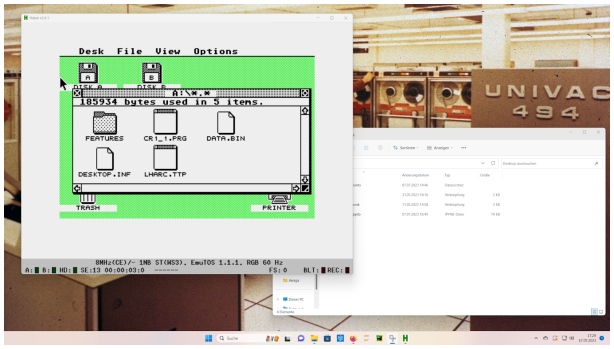


Abb. 1: Emulation eines Atari-ST-Systems im Emulator Hatari auf einem PC mit Windows 11.

Ein großer Teil des kulturellen Erbes der Vergangenheit, Gegenwart und der Zukunft ist digital: Zu den relevanten Überlieferungsträgern zählen insbesondere seit der Etablierung von Heimcomputern in den 1980er Jahren in großer Menge elektronische Medienformate sowie die unmittelbar dazugehörige Hard- und Software. Die UNESCO hat früh den Stellenwert dieses besonderen Kulturerbes erkannt: Im Oktober 2003 deklarierte sie mit ihrer *Charter on the Preservation of the Digital Heritage* erstmals Grundsätze zum Umgang mit digitalem kulturellem Erbe auf internationaler Ebene, um es präventiv vor Verlust zu schützen und für künftige Generationen zu bewahren (vgl. UNESCO, 2003).

Digitales Kulturerbe wird jedoch bislang von kulturellen Institutionen und Gedächtnisorganisationen im Hinblick auf ihren originären Auftrag zu sammeln, zu erforschen, zu bewahren und auszustellen nicht ausreichend mitgedacht. Diese Aufgabe kann nicht allein von öffentlichen Institutionen und Organisationen bewältigt werden; vielmehr handelt es sich um eine weitreichende Herausforderung für alle Akteur*innen im digitalen Raum. Ohne Strategien der Erhaltung und etablierte Praktiken steht ein beträchtlicher Teil des frühen digitalen kulturellen Erbes aktuell vor der Gefahr, für die Nachwelt verloren zu gehen (vgl. Lusenet, 2007).

Die in der Charta formulierten präventiven Handlungsabsichten sind jedoch nicht trivial umzusetzen, da die überlieferten Daten häufig nur im Rahmen ihrer ursprünglichen technischen Infrastruktur aktiv erhalten werden können. Hardware wiederum ist von großer Flüchtigkeit und kann nur mit großem Aufwand funktionsfähig gehalten werden. Zur Abbildung dieser hat sich daher die Methode der Emulation als Zugang zu älteren Dateiformaten in deren ursprünglichem Kontext etabliert: Auf dem Host-System wird ein Großteil der internen Strukturen sowie der Benutzeroberfläche des emulierten Systems virtualisiert (vgl. Abb. 1). Vor dem Hintergrund methodologischer Strömungen im Bereich des Retrocomputings, namentlich der ‚Codephilologie' (vgl. Höltgen, 2022), steht die Ar-

beit an historischer Software in der neuralgischen Schnittstelle zwischen Informatik und Geisteswissenschaften. Die Erforschung digitaler Objekte in ihrem historischen wie technischen Kontext erfordert eine hybride Arbeitsweise, ein gewisses Maß an Kreativität sowie ein Grundverständnis informationstechnischer Prozesse, was den direkten Anschluss an die Digital Humanities ermöglicht.

Themen des Workshops

Der geplante Workshop findet vor dem Hintergrund eines Forschungsprojektes statt, welches sich seit Anfang 2023 am Zentrum für Philologie und Digitalität der Julius-Maximilians-Universität Würzburg mit der Erschließung und Erhaltung sogenannter Diskettenmagazine (Diskmags) auseinandersetzt. Bei diesen handelt es sich um periodische Publikationen auf Diskette mit Inhalten der Heimcomputer-, Computerspiele- und Demoszene (vgl. Roeder, 2022). Um diese zu erschließen, kommt im Laufe des Projekts eine Vielzahl an Emulatoren von Heimcomputern zum Einsatz, welche mit Rücksicht auf die wissenschaftliche Arbeit jedoch eine besondere Reflektion erfahren: Als Vermittlungsinstanz müssen Emulatoren als „Interpretation zweiter Ordnung“ (Pias, 2017, 384) gelten und nehmen einen „Umschreibprozess“ (Höltgen, 2022, 106) am informatischen Artefakt vor. In vielen Aspekten können die Emulatoren kein vollständiges Surrogat für die originalen Systeme sein; die Betrachtung der digitalen Objekte im Emulator ist immer defizitär. Auch diese Reflektion und Perspektivierung der Emulatoren als eigenes System mit den entstehenden Trade-Off-Effekten ist jedoch noch nicht ausreichend vorangetrieben worden, sodass die Emulatoren nicht nur als Werkzeug, sondern auch als Forschungsobjekt von Relevanz für die DH sind.

Erkenntnisinteressen

In einer Vielfalt an Disziplinen ist die Erhaltung historischer Software und Dateien inzwischen von besonderer Relevanz, darunter in den Game Studies und in der Erforschung digitaler Literatur genauso wie in der Archivarbeit sowie in der Museumspraxis. Emulatoren können dabei die Zugänglichkeit zu digitalen Objekten stärken – sowohl für die Forschenden als auch für die öffentliche Präsentation. Als heuristisches Modell für das Originalsystem ermöglichen sie einen Zugriff, der das digitale Artefakt in das Licht unmittelbarer Interagierbarkeit stellt. So erlauben Emulatoren in Museen wie dem Computerspielemuseum in Berlin die Präsentation von Spielen, ohne die ebenso erhaltenswerte historische Hardware einer übermäßigen Abnutzung zu unterziehen. (vgl. Lange, 2012) Auch in der Erforschung digitaler Literatur spielen Emulatoren eine wichtige Rolle, können sie doch die spezifischen Episteme in der Entstehung der Texte dokumentieren. In der Erschließung digitaler Nachlässe ermöglichen Emulatoren den Zugriff auf die oftmals proprietären Dateiformate älterer Textverarbeitungssysteme. Ferner werden Emulatoren in den Kontext von Langzeitarchivierung digitaler Objekte eingeordnet (vgl. Loebel, 2014), wo sie den Erhalt von Software unabhängig von der zugehörigen Hardware ermöglichen.

Die angesprochenen Felder sind Teil verschiedener zentraler Diskurse in den DH, die den Erhalt historischer Daten als Gemeinsamkeit haben. Im vorgeschlagenen Workshop ist es diese Abhängigkeit, die als Motivation der Stärkung einer Schnittstelle zwischen Retrocomputing und DH-Forschung gelten muss. Darüber hinaus begründet die überaus hohe Heterogenität der Datei- und Systemstrukturen den Hands-on-Ansatz als unmittelbare Notwendigkeit in der Vermittlung der Problemstellen computerhistorischen Arbeitens.

Der Hands-on-Ansatz des Arbeitens mit Emulatoren als Kerngebiet des Retrocomputings soll daher im Workshop durch theoretische Perspektiven und Ausblicke ergänzt werden. Dabei kann auf die Sensibilität für die Datenquellen im Retrocomputing und für den Erhalt von historischen Datenträgern hingewiesen werden. Als Beispiel lassen sich hier die im oben erwähnten Projekt untersuchten Diskmags heranziehen: In vielen Fällen wurden die Disketten nur dezentral von privaten Sammler*innen gesammelt, archiviert und online zugänglich gemacht. Damit sind sie prototypisch für viele Ressourcen im Retrocomputing, die zusätzliche Reflektion im Arbeitsprozess erfordern. Ferner stellen sich auch Fragen zum Verhältnis zwischen Emulation, digitaler Edition und originaler Hardware: Für die Erschließung historischer digitaler Artefakte muss eine Abwägung zwischen deren verschiedenen Möglichkeiten und Bedingungen getroffen werden. Aus diesen Feldern speist sich die Diskussion im Seminar, nicht zuletzt jedoch mit dem übergeordneten Ziel, Anwendungs- und Entwicklungspotenziale des Retrocomputings in den DH ausgehend von den fachlichen Hintergründen der Teilnehmenden zu evaluieren.

Ablauf des Workshops

Das Ziel des Workshops ist es daher zunächst, den Teilnehmenden das theoretische und praktische Wissen über die Grundlagen und Prinzipien von Retrocomputing zu vermitteln. Die Teilnehmenden müssen dabei kein spezifisches Wissen zum Thema besitzen. Im Workshop werden sie erfahren, wie die ausgewählten Systeme funktionieren, wie man mit einem Emulator arbeitet und wie die Daten aus den Disketten in virtueller Form (*Disk Images*) extrahiert und verarbeitet werden. Über das praktische Wissen aus dem Hands-on-Verfahren hinaus werden mit der Diskussion des Einsatzes von Emulatoren für die Erhaltung digitaler Objekte potenzielle Forschungsdesiderate in den DH offengelegt.

Um eine angenehme und produktive Arbeitsatmosphäre während des Workshops gewährleisten zu können, ist die Teilnehmerzahl auf 30 Personen begrenzt. Da es sich um ein Hands-on-Format handelt, sollen die Teilnehmenden ei-

gene Laptops mitbringen. Die erwähnten Emulatoren sind auf den üblichen Betriebssystemen lauffähig. Zudem ist seitens der Referent*innen wünschenswert, dass der Workshop-Raum über einen stabilen Internetzugang verfügt und mit einem Beamer sowie einer Tafel oder einem Flipchart und außerdem mit Steckdosen in erreichbarer Nähe der Arbeitsplätze ausgestattet ist. Sinnvoll wäre weiterhin die Möglichkeit, Gruppentische für jeweils 4 bis 6 Personen zu bilden.

Der Workshop ist in Form eines ganztägigen Emulator-Hackathons mit zwei Sessions je vier Stunden angelegt.

Die erste Hälfte des Workshops startet mit einer kurzen Vorstellungsrunde (0,5h), in der sich die Teilnehmenden über ihre Interessen und Erfahrungen mit Retrocomputing austauschen können. Anschließend folgt der erste theoretische Teil, welcher den Teilnehmenden die Grundlagen der Emulation näher bringt und Kenntnisse über gängige Emulatoren sowie deren Anwendung vermittelt. Dabei nimmt insbesondere das Verhältnis der Emulatoren zur originalen Hardware einen großen Teil ein; geplant ist dafür auch, historische Geräte vorzuführen (1h). Nach einer 15-minütigen Pause folgt der erste praktische Teil (2h), in welchem sich die Teilnehmenden gemeinsam mit den Referent*innen mit dem VICE-Emulator für das System *Commodore 64* auseinandersetzen. Hierbei wird das Programmieren mit der BASIC-Sprache umrissen; die Teilnehmenden können sich zudem an simplen Textverarbeitungsaufgaben ausprobieren sowie ausgewählte Diskmags und deren Inhalt entdecken.

Die zweite Hälfte des Workshops dient dazu, die in der ersten Session erlernten Fähigkeiten und Kenntnisse zu reflektieren und zu erweitern. Auf Basis dessen wird zu Beginn in einem zweiten theoretischen Teil der Trade-Off-Effekt zwischen Hardware und Emulation mit den Teilnehmer*innen diskutiert sowie sich mit der Frage auseinandergesetzt, welche Vorteile das Arbeiten mit Emulatoren für die Forschung im wissenschaftlichen Bereich bieten kann (1h). Diesem schließt sich ein zweiter praktischer Teil (2h) an, in welchem die Teilnehmer*innen sich in mehreren Kleingruppen zusammenschließen, um so partizipativ verschiedene andere Systeme (z.B. Atari, Amiga, Apple II etc.) und ihre Emulatoren zu testen und auszuprobieren. Diese aktive Teilhabe soll das bis zu diesem Zeitpunkt Erlernte festigen, um so einen Erkenntnisgewinn für die Teilnehmenden zu gewährleisten.

Dem Workshop schließt sich gegen Ende hin eine Abschlussdiskussion (1h) an, in welcher die Teilnehmer*innen gemeinsam mit den Referent*innen auf den Verlauf des Workshops zurückblicken, die Inhalte kritisch reflektieren sowie aktiv miteinander über Emulatoren und das Potenzial ihres Einsatzes im Kontext der wissenschaftlichen Arbeit innerhalb der Digital Humanities diskutieren.

Vor dem Workshop werden wir eine kurze Umfrage zu besonderen Interessen und Kenntnissen der Teilnehmer*innen erstellen, um die zweite Hälfte des Workshops in der inhaltlichen Ausrichtung daraufhin vorzubereiten und anzupassen.

Kontaktdaten und Forschungsinteressen

Dr. phil. Torsten Roeder

Kontakt: torsten.roeder@uni-wuerzburg.de

Wissenschaftlicher Mitarbeiter am Zentrum für Philologie und Digitalität an der Universität Würzburg, zuständig für Digitale Editionen, aktuelle Schwerpunkte Hybridität und frühes digitales Kulturerbe.

Yannik Werner Herbst, B.A.

Kontakt: yannik.herbst@uni-wuerzburg.de

Ist Student der Digital Humanities im Master und arbeitet als wissenschaftlicher Mitarbeiter im Zentrum für Philologie und Digitalität an der Universität Würzburg. Seine Forschungsinteressen umfassen Digitale Editionen mit Schwerpunkt auf Synoptischen Ansichten, Interface Beschreibungen, Data Mining, Retrocomputing etc.

Johannes Leitgeb, B.A.

Kontakt: johannes.leitgeb@stud-mail.uni-wuerzburg.de

Ist Student der Digital Humanities und der Germanistik im Master und arbeitet als wissenschaftliche Hilfskraft im Zentrum für Philologie und Digitalität an der Universität Würzburg. Seine Forschungsinteressen umfassen Retrocomputing, Computational Literary Studies sowie die Darstellung von Informatik und Physik in der deutschsprachigen Literatur.

Madlin Marenec, B. A.

Kontakt: madlin@mrsmuseum.de

Studiert Digital Humanities sowie Museum Studies im Master und arbeitet als wissenschaftliche Hilfskraft im Zentrum für Philologie und Digitalität an der Universität Würzburg. Ihre Forschungsinteressen umfassen die Digital Literacy im Museumsbereich, die Computational Museology sowie digitale Sammlungsforschung.

Tomash Shtohryn, B.A.

Kontakt: tomash.shtohryn@stud-mail.uni-wuerzburg.de

Studiert Digital Humanities im Master und arbeitet als wissenschaftliche Hilfskraft am Zentrum für Philologie und Digitalität an der Universität Würzburg. Seine Forschungsinteressen umfassen die Anwendung von Data Science-Methoden im Bereich der Digitalen Editionen und Natural Language Processing.

Bibliographie

Höltgen, Stefan. 2022. *Open History. Archäologie des Retrocomputings.* Berlin: Kulturverlag Kadmos.

Lange, Andreas. 2012. "Pacman im Archiv. Computerspiele als digitales Kulturgut. " *Zeithistorische Forschungen/Studies in Contemporary History* 9(2). 10.14765/zzf.dok-1587 (zugegriffen: 05. Dezember 2023).

Loebel, Jens-Martin. 2014. *Lost in Translation. Leistungsfähigkeit, Einsatz und Grenzen von Emulatoren bei der Langzeitbewahrung digitaler multimedialer Objekte am Beispiel von Computerspielen.* Glückstadt: Verlag Werner Hülsbusch.

Lusenet, Yola de. 2007. "Tending the garden or harvesting the fields: digital preservation and the UNESCO Charter on the Preservation of the Digital Heritage. " *Library trends* 56(1): 165-182. 10.1353/lib.2007.0053 (zugegriffen: 19. Juli 2023).

Pias, Claus. 2017. "Medienphilologie und ihre Grenzen. " In *Medienphilologie: Konturen eines Paradigmas*, hg. von Friedrich Balke und Rupert Gaderer, 365-385. Göttingen: Wallstein.

Roeder, Torsten. 2022. "Rescuing Diskmags: Towards Scholarly [Re]Digitisation of an Early Born-Digital Heritage. " *magazén* 3(1): 139-158. 10.30687/mag/2724-3923/2022/05/006 (zugegriffen: 18. Juli 2023).

UNESCO. 2003. "Charter on the Preservation of the Digital Heritage. " CL/3865. https://unesdoc.unesco.org/ark:/48223/pf0000179529 (zugegriffen: 18. Juli 2023).

Vernetzte Forschungsdaten - wer kennt wen im Mittelalter?

Pultar, Yannick

yannick.pultar@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0009-0002-0819-958X

Abel, Christina

Christina.Abel@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland; Universität des Saarlandes, Deutschland
ORCID: 0009-0001-5858-7698

Weber, Matthias

Matthias.Weber@rub.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland; Ruhr-Universität-Bochum
ORCID: 0000-0003-2198-8989

Kasper, Dominik

Dominik.Kasper@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0000-0002-6587-381X

Kuczera, Andreas

andreas.kuczera@mni.thm.de
Technische Hochschule Mittelhessen, Deutschland
ORCID: 0000-0003-1020-507X

Abstract

Der Workshop bietet eine Einführung in die Arbeit mit Orts- und Personendaten sowie Graphdatenbanken am Beispiel eines mediävistischen Datenkorpus'. Die Teilnehmerinnen und Teilnehmer lernen, wie sie Forschungsdaten nachnutzen, in eine Graphdatenbank integrieren und dort die Beziehungen von Entitäten analysieren können sowie wie sie die Aggregation von Normdaten und weiterer Forschungsdaten für die Verfolgung wissenschaftlicher Fragestellungen nutzen können. Durch hands-on-sessions und die Vorstellung von konkreten Anwendungsbeispielen wird das Thema niedrigschwellig, anschaulich und praxisnah vermittelt. Der Workshop richtet sich an Forschende mit Interesse an digitaler Geschichtswissenschaft. Vorkenntnisse im Bereich Graphentechnologie bzw. -datenbanken sind nicht notwendig.

Einführung

Im Rahmen der Digitalisierungsbemühungen der letzten drei Jahrzehnte sind eine Reihe relevanter Quellen für die Geschichtswissenschaften als Forschungsdaten nach den FAIR-Prinzipien frei und nachnutzbar zur Verfügung gestellt worden, auch wenn sie bisher selten Teil des Semantic Web sind. Angesichts verschieden tief und nach unterschiedlichen Standards strukturierter Datenbestände aus vielfältigen Forschungstraditionen stehen Historiker*innen in der Praxis aber oft vor besonderen Herausforderungen, wenn es um die konkrete Nachnutzung der heterogenen Daten geht, mit denen ein sinnvoller Umgang gefunden werden muss, um Auswertungsperspektiven zu eröffnen. Eine davon ist der Umgang mit Orts- und Personendaten. Die Entitäten können durch Normdaten eindeutig identifiziert und formal beschrieben werden, sodass Informationen über diese aggregiert und sie über verschiedene Ressourcen hin-

weg eindeutig adressierbar sind. Für die Entitäten, die in der Mediävistik erforscht werden, stehen etwa bisher verhältnismäßig wenig Normdaten zur Verfügung. GND und Wikidata als zwei zentrale Normdatenbestände, verzeichnen etwa nur 37.000 beziehungsweise 13.000 Personendatensätze mit Belegen für der Zeit zwischen 500-1500 mit einem Schwerpunkt auf das 15. Jahrhundert.

Graphdatenbanken können durch ihre Modellierung als Entitäten- bzw. Relationennetzwerk gut geeignet sein, einen Zugang zu solchen Personen- und Ortsdaten schaffen und diese zu analysieren: Sie bieten etwa die Möglichkeit komplexer Abfragen bei hoher Performanz und der Visualisierung des Entitätennetzwerks oder von Teilnetzwerken als Graph.

Ziel des Workshops

Der vorgeschlagene Workshop bietet eine Einführung im Umgang mit der Nachnutzung von Forschungsdaten, die historische Quellen mit personen- und ortsbezogenen Entitäten verbinden, mittels der Nutzung von Graphdatenbanken. Inhaltlich dient ein Datensatz der Regesta Imperii, als einer für den Bereich der Mediävistik zentralen Bereitsteller digitaler Forschungsressourcen, als Beispiel. Als Graphdatenbanksystem wird Neo4J vorgestellt. Der Workshop richtet sich an Forschende mit Interesse an digitaler Geschichtswissenschaft. Ohne Vorkenntnisse im Bereich der Graphentechnologien. Durch *hands-on-sessions* und die Vorstellung von konkreten Anwendungsbeispielen wird im Laufe des Workshops ein vollständiger Workflow zur Bearbeitung einer Forschungsfrage mit frei verfügbaren Daten praxisnah durchgespielt und das Thema so anschaulich und niedrigschwellig vermittelt:

- Einführung in einen Beispieldatensatz
- Import der Daten in eine eigene Neo4J-Datenbank als Arbeitsdatenbank
- Aggregation und Anreicherung der Forschungsdaten mit anderen Datenbeständen am Beispiel von Normdaten
- Möglichkeiten der Analyse der Entitäten und ihrer Beziehungen mithilfe der Abfragesprache Cypher

Die Daten der Regesta Imperii

Die Regesten der Regesta Imperii bieten deutschsprachige Zusammenfassungen mittelalterlicher Urkunden und narrativer Texte, in denen nicht nur die Inhalte der Quelle, sondern auch sämtliche in den Quellen genannten Orts- und Personennamen aufgenommen werden. Mit ihren rund 200.000 Regestendatensätzen für den Zeitraum von 700 bis 1519 steht damit ein enormer Schatz an Personen- und Ortsdaten zur Verfügung, der zum größten Teil durch Register mit einer ähnlich großen Zahl an Einträgen erschlossen ist. Register sind ein zentraler Schlüssel für die Nutzung von Forschungsdaten Zum einen identifizieren, erschließen und strukturieren sie die in den historischen Quellen vorkommenden Entitäten und zum anderen weisen sie diese Entitäten konkreten Schriftstücken zu, geben ihnen damit einen räumlichen und zeitlichen Kontext.

Der Datenbestand der Regesta Imperii, bereitgestellt im CSV-Format sowie in CEI- und TEI-XML-Datensätzen unter http://www.regesta-imperii.de/daten , kann mit einer zweihundertjährigen Erhebungs- und einer zwanzigjährigen Digitalisierungsgeschichte in nuce für die heutige historisch ausgerichtete Forschungsdatenlandschaft stehen. Die durch die Kombination aus Regesten- und Entitätendatensätzen entstehenden vernetzten Forschungsdaten lassen sich sehr gut mit Hilfe von Graphdatenbanken modellieren und analysieren. So wird ein Wechsel des Fokus von der Quelle, bzw. dem Regest, hin zu den in den Quellen genannten Entitäten ermöglicht.

Die hands-on-sessions des Workshops wollen diese Möglichkeiten an einem Ausschnitt zu einem Herrscher ausloten: dem Dante-Kaiser Heinrich VII. (1308-1313), dessen Regesten aktuell an der Saarbrücker Arbeitsstelle des Akademienprojekts erarbeitet werden. Die kurze Regierungszeit dieses Herrschers ist außergewöhnlich gut durch Urkunden, Verwaltungsschriften und Chroniken dokumentiert und bietet damit einen sehr dichten Bestand an Personen-und Ortsdaten ohne größere chronologische Lücken. Aus diesem Bestand wird den Teilnehmenden ein Datensatz von Regest- und Registerdaten im CSV-Format bereitgestellt, der von Mitarbeitenden des Projekts in domänenspezifische Fragen eingeordnet wird.

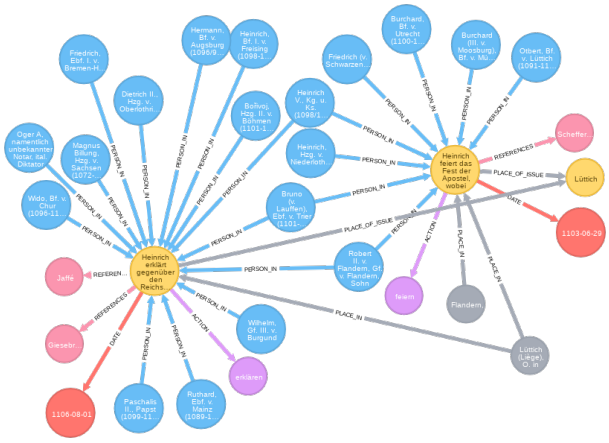


Abbildung 1: In der Abbildung ist beispielhaft das Regest RI III,2,3 Nr. 1487 als Modell im Graphen abgebildet (neben anderen Regesten und den verknüpften Entitäten).

Ablaufplan

Der Workshop soll an zwei Tagen für jeweils vier Stunden stattfinden. Der erste Tag teilt sich in zwei Abschnitte. Zunächst findet eine Einführung in den Beispieldatensatz und das Projekt Regesta Imperii statt. Dabei wird mit den

Teilnehmenden das Datenmodell der Regesten und der damit verbundenen personen- und ortsbezogenen Datensätze erarbeitet und dessen Eignung für die Verfolgung verschiedener Fragestellungen diskutiert. Der zweite Abschnitt ist der Vorstellung und Erprobung der Arbeit mit der Graphdatenbank Neo4j gewidmet. Dabei werden grundlegende Konzepte der Graphentechnologie mittels einer *hands-on-session* vermittelt, bei der die Teilnehmenden jeweils eigene Neo4j-Datenbanken erstellen, mit denen sie im Laufe des Workshops weiterarbeiten werden. Sie lernen die Funktionsweise und das Layout der grafischen Oberfläche von Neo4j kennen und werden in die Lage versetzt, die zur Verfügung gestellten Datensätze in die Graphdatenbank eigenständig zu importieren. Eine Neo4j-Sandbox-Umgebung wird den Teilnehmenden vorab zur Verfügung gestellt.

Am zweiten Tag folgen, aufbauend auf den am ersten Tag erworbenen Kenntnissen, weitere, fortgeschrittenere Übungen zu Auswertungsperspektiven mittels der Graphdatenbank und deren Abfragesprache Cypher und anderer niedrigschwelliger Technologien, mit Schwerpunkt auf der Auswertung der enthaltenen Personen- und Ortsdaten. Dabei werden begleitend Vor- und Nachteile graphbasierter Erschließung von Forschungsdaten diskutiert. An diesem Tag werden v. a. Fragen zur Operationalisierung und Formalisierung von Fragestellungen und der weiteren Anreicherung der Daten mit nicht im Beispieldatensatz enthaltenen Informationen am Beispiel von Normdaten behandelt und in hands-on-sessions praktisch erprobt. Zunächst wird in verfügbare Normdaten zur Mediävistik (GND, Wikidata, Germania Sacra-Personendatenbank) eingeführt und anschließend die orts- und personenbezogenen Entitäten mit den Normdaten aus der Wikidata-Datenbank angereichert, für deren SPARQL-Endpoint ein niedrigschwelliger Query Service zur Verfügung steht. Dabei soll auch der Umgang mit Forschungsfragen reflektiert werden: Es wird diskutiert, welche Forschungsfragen an den Datenbestand unter Berücksichtigung der Datenmodellierung und des Informationsgehalts gestellt werden können und Schritte zu deren Operationalisierung und Formalisierung an die Forschungsdaten erarbeitet. Dabei stehen Fragen zu den Beziehungen der im Korpus erschlossenen Personenentitäten im Mittelpunkt.

Der letzte Teil des Workshops ist für die Frage der Übertragbarkeit des Erlernten, die Anwendung von Graphentechnologien und den Umgang mit Orts- und Personendaten, auf eigene Projekte der Teilnehmenden reserviert. Die Teilnehmenden können hierfür eigenes Datenmaterial und Projektideen mitbringen.

Benötigte Ausstattung

Beamer, ausreichend Steckdosen und WLAN

Teilnehmende: Laptops, Installation von Neo4j Desktop

Bereitstellung durch Workshop-Veranstaltende: kollaborativ nutzbare Markdown-Umgebung für gemeinsame Notizen, Beispieldatensätze in verschiedenen Formaten

Neo4j-Sandbox-Umgebung für die Datenbanken der Teilnehmenden werden im Vorhinein gemeinsam mit einer Installationsdokumentation bereitgestellt, um einen reibungslosen Einstieg auf einer gemeinsamen Basis zu gewährleisten

Teilnehmerzahl

5-20

Beitragende

Yannick Pultar (https://orcid.org/0009-0002-0819-958X) ist Arbeitsstellenleiter der RI Online an der Akademie der Wissenschaften und der Literatur | Mainz. Seine Forschungsschwerpunkte liegen auf der digitalen Erschließung und Modellierung historischer Quellen sowie auf der Analyse der Entitätennetzwerken in der Schriftlichkeit der römisch-deutschen Herrscher des 14. Jahrhunderts.

Christina Abel (https://orcid.org/0009-0001-5858-769) ist Arbeitsstellenleiterin des RI-Teilprojekts zu Kaiser Heinrich VII., das an der Universität des Saarlandes (Saarbrücken) und an der Akademie der Wissenschaften und der Literatur | Mainz angesiedelt ist. Im Rahmen des Projekts arbeitet sie an einer prosopographischen Aufarbeitung der verschiedenen Personengruppen und -netzwerke an Heinrichs Hof und erforscht deren Einfluss auf politische und administrative Entscheidungen.

Matthias Weber (https://orcid.org/0000-0003-2198-8989) ist Juniorprofessor für die Geschichte des Hochmittelalters und digitale Prosopographie an der Ruhr-Universität Bochum und der Akademie der Wissenschaften und der Literatur | Mainz. Seine Forschungsschwerpunkte liegen in der Zeit des salischen Jahrhunderts (1024-1125), der Mentalitätsgeschichte zum Tod sowie in der Historiographiegeschichte.

Dominik Kasper (https://orcid.org/0000-0002-6587-381X) ist DevOps Engineer (Entwickler und Systemadministrator) an der Akademie der Wissenschaften und der Literatur | Mainz, u. a. im Projekt Regesta Imperii. Seine Interessenschwerpunkte liegen in den Bereichen Research Software Engineering, der Theorie und Praxis digitaler Methoden in den Geistes- und Kulturwissenschaften, insbesondere von digitalen Editionen und Sammlungen.

Andreas Kuczera (https://orcid.org/0000-0003-1020-507X) ist Professor für anwendungsbezogene digitale Methodik in den Geistes- und Sozialwissenschaften an der Technischen Hochschule Mittelhessen in G ießen. Seine Forschungsschwerpunkte liegen in der Erforschung anwendungsbezogener Methoden in den digitalen Geistes- und Sozialwissenschaften.

Bibliographie

Bornhofen, Stefan und Marten Düring . 2020. “Exploring dynamic multilayer graphs for digital humanities.” In *Applied Network Science* 5. https://doi.org/10.1007/s41109-020-00295-x (zugegriffen: 18.07.2023).

Hitzler, Pascal, Markus Krötzsch, Sebastian Rudolph und York Sure . 2008. Semantic Web. Grundlagen. Berlin/ Heidelberg: Springer.

Kuczera, Andreas . 2019. “Die ‚Regesta Imperii‘ im digitalen Zeitalter. Das Regest als Netzwerk von Entitäten.” In *Das Mittelalter* 24: 157–172.

Kuczera, Andreas . 2018. Regestenmodellierung im Graphen. In *Graphentechnologien in den digitalen Geisteswissenschaften* . https://kuczera.github.io/ Graphentechnologien/20_Regestenmodellierung-im-Graphen.html (zugegriffen : 18.07.2023).

Opitz, Juri . 2020. “Automatic Creation of a Large-Scale Tempo-Spatial and Semantic Medieval European Information System.” In *Proceedings of the Workshop on Computational Humanities Research (CHR)* : 397–419. http://ceur-ws.org/Vol-2723/long12.pdf (zugegriffen : 18.07.2023).

Schulz, Julian . 2017. "Review of 'Regesta Imperii Online'." In *RIDE* 6. http://doi.org/10.18716/ride.a.6.5 (zugegriffen: 18.07.2023).

Stadler, Peter. 2012. “Normdateien in der Edition.” In *Editio* 26/1: 174–183. https://doi.org/10.1515/ editio-2012-0013 (zugegriffen: 18.07.2023).

Videoanalyse mit der Plattform TIB-AV-A. Grundlagen, Schnittstellen, Zukunftsperspektiven

Baresch, Ariadne
baresch@uni-trier.de
Universität Trier, Deutschland

Diecke, Josephine
diecke@staff.uni-marburg.de
Philipps-Universität Marburg, Deutschland

Ewerth, Ralph
ralph.ewerth@tib.eu
TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften Hannover, Deutschland
ORCID: 0000-0003-0918-6297

Howanitz, Gernot
gernot.howanitz@uibk.ac.at
Universität Innsbruck, Österreich

Müller-Budack, Eric
eric.mueller@tib.eu
TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften Hannover, Deutschland
ORCID: 0000-0002-6802-1241

Radisch, Erik
radisch@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0002-0089-9082

Springstein, Matthias
matthias.springstein@tib.eu
TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften Hannover, Deutschland
ORCID: 0000-0002-6509-8534

Workshop-Konzept

Während sich die Digital Humanities traditionell stark auf die Analyse von textuellen Daten fokussiert haben, hat das Interesse an der Film- und Videoanalyse innerhalb der Community in den letzten Jahren stark zugenommen (Burghardt et al. 2020, Sittel 2017). Videos sind typischerweise multimodal (Bewegtbild, ggf. Ton und überlagerter Text) und Analysen und Studien auf Basis manueller Annotationen nehmen in der Regel sehr viel Zeit in Anspruch. Deshalb sind Informatik-Methoden der Mustererkennung zur (semi-)automatischen Auswertung der verschiedenen Modalitäten von hoher Bedeutung, da sie Forscher:innen aus den Digital Humanities Informationen über die interne Dynamik und Verhältnismäßigkeit einzelner Videos oder sogar ganzer Korpora liefern können. Allerdings schreiten die Entwicklungen insbesondere im Bereich der Künstlichen Intelligenz und dem Teilgebiet des maschinellen Lernens sehr schnell voran und die Implementierung aktueller Methoden erfordert ein hohes Maß an technischem Verständnis und Rechenressourcen. Aus diesem Grund haben Forscher:innen aus den Digital Humanities oft keinen oder nur sehr eingeschränkten Zugriff auf aktuelle KI-Methoden zur automatischen Videoanalyse. Zwar existieren bereits verschiedene Plattformen für die Videoanalyse, diese ermöglichen jedoch entweder ausschließlich manuelle Annotationen (z.B. ANVIL [1] , Kipp 2014; Cinemetrics [2] ; ELAN [3] , Wittenburg et al. 2016) oder enthalten nur wenige ausgewählte, dafür aber speziell optimierte Methoden für die automatische Inhaltsanalyse (z.B. Videana , Ewerth et al. 2009; VIAN [4] , Halter et al. 2019, VIAN-DH [5]). Zudem ba-

sieren diese lokalen Software-Lösungen in der Regel nicht auf aktuellen KI-Ansätzen bzw. erfordern entsprechende Hardware seitens der Nutzer:innen. Die Bereitstellung von leicht zu bedienenden (Web)-Plattformen, die Forschende aus vielen Disziplinen der Digital Humanities Zugang zu State-of-the-Art-Ansätzen aus dem Bereichen der Mustererkennung und Multimedia Retrieval bieten ist daher von größter Bedeutung, um praktische Anwendungen im Bereich der Medienforschung zu ermöglichen.

Dieses desideratum adressiert die webbasierte Plattform “TIB AV-Analytics” (TIB-AV-A, https://service.tib.eu/tibava), die einen niederschwelligen Zugang zur systematischen Film- und Videoanalyse anbietet. Ziel des Workshops ist, Teilnehmer:innen aus den Digital Humanities sowie weiteren Interessierten TIB-AV-A vorzustellen. Zu diesem Zweck kooperieren die Entwickler:innen von TIB-AV-A mit der DHd AG Film & Video. Es werden keine inhaltlichen Vorkenntnisse von den Teilnehmenden erwartet. Die Videoanalyse-Platform TIB-AV-A wird einem von der Deutschen Forschungsgemeinschaft geförderten Projekt (Laufzeit: 01.01.2021–31.12.2023) an der Technischen Informationsbibliothek (TIB) in Hannover in Kooperation mit Filmwissenschaftler:innen der Johannes Gutenberg-Universität Mainz (JGU) entwickelt. Im Rahmen des Projekts wurde eine Anforderungsanalyse erstellt. Im Gegensatz zu den oben genannten Plattformen nutzt TIB-AV-A moderne Webtechnologien, um Nutzer:innen eine reaktionsschnelle und interaktive Weboberfläche zur Verfügung zu stellen, die manuelle Annotationen und Zugang zu leistungsstarken Deep-Learning-Verfahren bietet, ohne dass fortgeschrittene technische Kenntnisse oder spezifische Hardwareanforderungen erforderlich sind. Details der Plattform werden überblicksartig von Springstein et al. (2023) beschrieben. Ein Screenshot der Plattform ist in Abb. 1 dargestellt.

Abb. 1: Die Benutzungsschnittstelle von TIB-AV-A am Beispiel einer Nachrichtensendung. TIB-AV-A enthält einen Videoplayer (a), eine Übersicht der erkannten Schnitte (b1) und Annotationen (b2), Zeitleisten (c), die kategoriale (z.B. Zeitleiste "Newscaster”) und numerische Werte (z.B. Zeitleiste "FIFA WC Trophy [CLIP]") anzeigen, und eine Navigationsleiste (d). Zeitleisten mit numerischen Werten zeigen z.B. die Wahrscheinlichkeit an, mit der ein Konzept in einem Videosegment abgebildet ist. Die Art der Visualisierung (Liniendiagramm, Farbdiagramm) und die Farbe (hier: von weiß unwahrscheinlich bis rot wahrscheinlich) kann frei gewählt werden.

Anders als bisherige Plattformen integriert TIB-AV-A für viele relevante Aufgaben der computergestützten Videoanalyse neueste Ansätze aus den Bereichen des maschinellen Sehens, der Audioanalyse und der natürlichen Sprachverarbeitung. Neben einer grundlegenden Einführung zur Bedienung der Plattform werden den Teilnehmenden in einer Einführung im ersten Teil des Workshops folgende Ansätze (bis zum Workshop werden noch weitere folgen) im Detail vorgestellt:

- Bild-/Videoanalyse:
  - Grundlegende Bildmerkmale: dominante Farbe(n) und Helligkeit
  - Erkennung von Einstellungswechseln
  - Schnittrate (Redfern, 2022), d. h. die Häufigkeit von Einstellungswechseln
  - Klassifikation der Einstellungsgröße zur Unterscheidung zwischen “Extreme Close-Ups”, “Close-Ups”, “Medium Shots”, “Full Shots” und “Long Shots”
  - Klassifikation von allgemeinen Orten bzw. Umgebungen (z. B. Schloss, Markt, Restaurant usw.)
  - Personenerkennung auf der Grundlage eines Beispielbildes
  - Erkennung von Gesichtsausdrücken (z. B. wütend, glücklich)
  - Zero-Shot-Bildklassifikation für beliebige visuelle Konzepte auf der Grundlage von Textbeschreibungen (z.B. "An image showing the FIFA WC Tropy" , siehe in Abb. 1c die Zeitleiste “ FIFA WC Trophy [CLIP]”)
  - Zero-Shot-Aktionserkennung auf der Grundlage von Textbeschreibungen (z.B. "A video showing people celebrating” , siehe in Abb. 1c die Zeitleiste “ Celebrating [X-CLIP]”)DH-Default
- Audio- und Sprachanalyse:
  - Grundlegende Audiomerkmale: Amplitudenkurve (Wellenform), Lautstärke (Root Mean Square) und das Frequenzspektrum
  - Spracherkennung zur automatischen Transkription von Sprache in Videos

Ein besonderer Fokus wird auf die Möglichkeiten aktueller Vision-Language-Modelle (Radford et al. 2021; Ma et al. 2022) und deren Kombination mit Large-Language-Modellen (Alayrac et al. 2022; Li et al. 2023) gelegt, die es Nutzenden erlaubt, in Videos nach beliebigen Konzepten, Objekten und Aktionen zu suchen. Auf Basis dieser Einführung sollen die Teilnehmenden TIB-AV-A für eigene Problemstellungen und Videos anwenden. Im zweiten Teil des Workshops werden die Import- und Exportmöglichkeiten vorgestellt, für welche TIB-AV-A zur Gewährleistung der Interoperabilität geeignete Schnittstellen und Import-Export-Funktionen zu bestehenden Plattformen wie beispielsweise ELAN und VIAN anbietet. Zum Abschluss werden von den Teilnehmenden die Ergebnisse und Erfahrungen mit TIB-AV-A vorgestellt und diskutiert.

Nach dem Workshop können die Teilnehmer:innen mithilfe von TIB-AV-A Problemstellungen der computergestützten Videoanalyse selbstständig bearbeiten. Zudem setzt der Workshop Impulse für neue Projekte, die visuelle Medien quantitativ erfassen und gleichzeitig die vorge-

stellten Methoden einer kritischen Evaluation unterziehen möchten. Auf Basis des Feedbacks der Teilnehmer:innen können neue Anregungen und Anforderungen gesammelt werden, die zur Weiterentwicklung von TIB-AV-A genutzt werden können. Darüber hinaus wird in Zusammenarbeit mit der AG Film & Video aktiv am Aufbau einer TIB-AV-A-Community gearbeitet, die Forschende und Entwickler:innen weiter vernetzt.

Programm

Vor dem Workshop

Die Vernetzung der Teilnehmerinnen und Teilnehmer geschieht über GitHub [6] und Discord [7] zur Identifizierung von gemeinsamen Forschungsinteressen und eventuellen Bildung von Gruppen. Zudem erhalten die Nutzer:innen ein Video-Tutorial zu TIB-AV-A (https://service.tib.eu/tibava) und werden gebeten, bereits im Vorfeld eigene Filmausschnitte (ca. 20 Minuten) zur Analyse hochzuladen, um beim Workshop selbst möglichst wenig Zeit zu verlieren.

Workshop-Programm

Mo., 26.02.2024, 14:00-14:30 *Kick-Off und Kennenlernrunde, Abfragen der Erwartungen*

Mo., 26.02.2024, 14:30-15:30 *Allgemeine Einführung in TIB-AV-A und filmwissenschaftliche Bezüge*

Zum Auftakt des Workshops wird eine allgemeine kurz gehaltene Einführung zu TIB-AV-A (siehe Screenshot in Abb. 1) gegeben, um ein Verständnis der Funktionsweise zur manuellen Annotation, automatischen Analyse und Visualisierung von Ergebnissen zu entwickeln. Den Teilnehmenden werden alle Funktionalitäten und deren Anwendungsmöglichkeiten an konkreten Beispielen gezeigt. Ein besonderer Fokus wird dabei auf die Erstellung von textuellen Beschreibungen (sogenannter “Prompts”) zur Klassifikation beliebiger Konzepte mithilfe von multimodalen Transformer-Modellen (z.B. Radford et al. 2021) gelegt. Hierbei wird eine filmwissenschaftliche Rahmung und Einordnung gegeben, welche Bezug auf einschlägige Debatten zum Thema nimmt, die in den letzten Jahren vermehrt in der Filmwissenschaft und innerhalb der AG Film & Video geführt wurden (Flückiger 2011; Melgar Estrada et al. 2017; Sittel 2017; Vonderau 2017; Burghardt et al. 2020) sowie die Motivation und Anwendungsmöglichkeiten der verschiedenen Analysemöglichkeiten aufgezeigt.

Mo., 26.02.2024, 15:30-16:00 *Kaffeepause*

Mo., 26.02.2024, 16:00-17:30 *TIB-AV-A Hands-on mit eigenen Videodaten*

Nach der allgemeinen Einführung wird den Teilnehmenden die Möglichkeit geboten, eigene Videos auf Basis individueller Use Cases und Fragestellungen mithilfe von TIB-AV-A zu analysieren. Die Entwickler der Plattform geben während dieses Prozesses Feedback und stehen für Nachfragen zur Verfügung. Hierbei ist das Ziel außerdem, weitere konkrete Anforderungen der Nutzenden an ein Videoanalyseportal zu erfassen.

Di., 27.02.2024, 9:30 - 10:30 *Einführung in Import- und Exportmöglichkeiten*

Im zweiten Teil des Workshops werden den Teilnehmenden die Import- und Exportmöglichkeiten von TIB-AV-A vorgestellt. Der Import der Daten erlaubt die weiterführende Nutzung und Analyse von Ergebnissen, die mithilfe anderer Tools oder Videoanalyseplattformen (z.B. ELAN, VIAN, VIAN-DH) erstellt worden sind. Mithilfe des Exports können die Ergebnisse durch andere Programme visualisiert und weiterverarbeitet werden. Nach der Vorstellung können die Teilnehmenden die Ergebnisse aus dem ersten Teil des Workshops mithilfe der Import- und Exportmöglichkeiten weiter verfeinern.

Di., 27.02.2024, 10:30-11:00 *Kaffeepause*

Di., 27.02.2024, 11:30-12:00 *Vorstellung der Ergebnisse*

Im letzten Teil des Workshops werden zunächst die Ergebnisse und Erfahrungen der Teilnehmenden vorgestellt. Im Anschluss erfolgt eine Diskussionsrunde, um Feedback und Anforderungen für TIB-AV-A, aber auch Videoanalyseplattformen für die Digital Humanities im Allgemeinen zu sammeln.

Di., 27.02.2024, 12:00-13:00 *Schlussrunde, Workshop-Evaluation*

In der Schlussrunde werden noch einmal die wesentlichen Ergebnisse zusammengefasst und Anschlussmöglichkeiten für Folgeprojekte, Community-Building etc. diskutiert.

Nach dem Workshop

Dieser Workshop soll Nutzer:innen aus den Digital Humanities Zugriff auf aktuelle KI-Ansätze zur systematischen Film- und Videoanalyse für verschiedene Anwendungen bieten. Eine Nachbereitung und weitere Vernetzung über GitHub und Discord ist ausdrücklich erwünscht, um eine weitere Begleitung der Projekte zu garantieren.

Zusätzliche Angaben

Benötigte technische Ausstattung: Beamer, WLAN-Zugang, ausreichend Steckdosen für die Laptops der Teilnehmer:innen

Zahl der möglichen Teilnehmer: 30

Forschungsinteressen

Ariadne Baresch (*baresch@uni-trier.de*): Computergestützte Film- und Bildanalyse, Erproben digitaler Methoden zur vergleichenden Analyse von Adaptionen (Film, Comic, Roman).

Josephine Diecke (*josephine.diecke@uzh.ch*): Einsatz und Reflexion digitaler Methoden und deren explorativer Potenziale in der filmwissenschaftlichen Forschung und

Lehre, insbesondere die computergestützte Filmanalyse mit manuellen und (semi)automatischen Annotationstools wie ELAN, VIAN und DVT.

Ralph Ewerth (*ralph.ewerth@tib.eu*): Erforschung und Entwicklung von Methoden zur Videoanalyse für Digital Humanities, Forschungsinfrastrukturen für Digital Humanities, interdisziplinäre Forschung zu Film-/Videoanalyse und Kunstgeschichte; Forschungsinteressen Informatik: Multimedia Retrieval, Computer Vision, Multimodal Computing, Visual Analytics.

Gernot Howanitz (*gernot.howanitz@uibk.ac.at*): Neue Medien in Osteuropa, insbesondere YouTube-Clips über politische Proteste, Deep Learning im geisteswissenschaftlichen Kontext

Eric Müller-Budack (*eric.mueller@tib.eu*): Erforschung und Entwicklung von unimodalen (Bild, Audio und Text) und multimodalen KI-Methoden zur Nachrichten-, Film- und Videoanalyse; Forschungsinteressen Informatik: Multimedia Retrieval, Computer Vision, Multimodal Computing

Erik Radisch (*radisch@saw-leipzig.de*): Deep Learning Algorithmen für visuelle Medien, Distant Viewing, Forschungsdatenmanagement, Entwicklung von Forschungsumgebungen für Bilddaten.

Matthias Springstein (matthias.springstein@tib.eu): Entwicklung und Implementierung von Methoden zur Videoanalyse für Digital Humanities, Computer Vision für Digital Art History.

Fußnoten

1. http://anvil-software.de
2. http://www.cinemetrics.lv
3. https://archive.mpi.nl/tla/elan
4. https://www.vian.app
5. https://www.liri.uzh.ch/en/projects/VIAN-DH.html
6. https://github.com/TIBHannover/TIBAVA-DHd24
7. https://discord.com/invite/EXMgGpXBDD

Bibliographie

Alayrac, Jean-Baptiste / Jeff Donahue / Pauline Luc / Antoine Miech / Iain Barr / Yana Hasson / Karel Lenc / Arthur Mensch / Katie Millican / Malcolm Reynolds / Roman Ring / Eliza Rutherford / Serkan Cabi / Tengda Han / Zhitao Gong / Sina Samangooei / Marianne Monteiro / Jacob Menick / Sebastian Borgeaud / Andrew Brock / Aida Nematzadeh / Sahand Sharifzadeh / Mikolaj Binkowski / Ricardo Barreira / Oriol Vinyals / Andrew Zisserman / Karen Simonyan (2022): "Flamingo: A Visual Language Model for Few-Shot Learning", in: *NeurIPS* 2022, 23716–23736. https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf [letzter Zugriff: 18. Juni 2023].

Burghardt, Manuel / Adelheid Heftberger / Johannes Pause / Niels Oliver Walkowski / Matthias Zeppelzauer (2020): "Film and Video Analysis in the Digital Humanities – An Interdisciplinary Dialog", in: *Digital Humanities Quarterly* 14(4). http://www.digitalhumanities.org/dhq/vol/14/4/000532/000532.html [letzter Zugriff: 18. Juni 2023].

Ewerth, Ralph / Markus Mühling / Thilo Stadelmann / Julinda Gllavata / Manfred Grauer / Bernd Freisleben (2009): "Videana: A Software Toolkit for Scientific Film Studies", in: Ross. Michael / Manfred Grauer / Bernd Freisleben (eds.): *Digital Tools in Media Studies: Analysis and Research. An Overview*. Bielefeld: Transcript, 101–116.

Flückiger, Barbara (2011): "Die Vermessung ästhetischer Erscheinungen", in: *Zeitschrift für Medienwissenschaft* 3 (2): 44–60. http://dx.doi.org/10.25969/mediarep/2606 [letzter Zugriff: 6. Dezember 2023].

Halter, Gaudenz / Rafael Ballester-Ripoll / Barbara Flueckiger / Renato Pajarola (2019): "VIAN: A Visual Annotation Tool for Film Analysis", in: *Computer Graphics Forum* 38(3), 119–129. https://doi.org/10.1111/cgf.13676 [letzter Zugriff: 18. Juni 2023].

Kipp, Michael (2014): "ANVIL: A Universal Video Research Tool", in: Durand, Jacques / Ulrike Gut / Gjert Kristoffersen (eds.) *The Oxford Handbook of Corpus Phonology* . Oxford: Oxford University Press, 420–436. https://doi.org/10.1093/oxfordhb/9780199571932.013.024 [letzter Zugriff: 18. Juni 2023].

Li, Junnan / Dongxu Li / Silvio Savarese / Steven Hoi (2023): "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", in: arXiv preprint, arXiv:2301.12597.

Ma, Yiwei / Guohai Xu / Xiaoshuai Sun / Ming Yan / Ji Zhang / Rongrong Ji (2022). "X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval", in: *ACM International Conference on Multimedia* , ACM MM 2022, 638–647. https://dl.acm.org/doi/abs/10.1145/3503161.3547910 [letzter Zugriff: 6. Dezember 2023].

Melgar Estrada, Liliana / Eva Hielscher / Marijn Koolen / Christian Gosvig Olesen / Julia Noordegraaf / Jaap Blom (2017): "Film Analysis as Annotation. Exploring Current Tools", in: *The Moving Image: The Journal of the Association of Moving Image Archivists*, 17 (2), 40–70. https://doi.org/10.5749/movingimage.17.2.0040 [letzter Zugriff: 6. Dezember 2023].

Radford, Alec / Jong Wook Kim / Chris Hallacy / Aditya Ramesh / Gabriel Goh / Sandhini Agarwal / Girish Sastry / Amanda Askell / Pamela Mishkin / Jack Clark / Gretchen Krueger / Ilya Sutskever (2021): "Learning Transferable Visual Models From Natural Language Supervision", in: *International Conference on Machine Learning* , ICML 2021, 8748–8763. https://

doi.org/10.48550/arXiv.2103.00020 [letzter Zugriff: 18. Juni 2023].

Sittel, Julian (2017): “Digital Humanities in der Filmwissenschaft”, in: *MEDIENwissenschaft: Rezensionen | Reviews* 34(4), 472-489. https://doi.org/10.17192/ep2017.4.7636 [letzter Zugriff: 18. Juni 2023].

Springstein, Matthias / Stamatakis, Markos / Margret Plank, / Sittel, Julian / Mauer, Roman / Bulgakowa, Oksana / Ewerth, Ralph / Müller-Budack, Eric (forthcoming): “TIB AV-Analytics: A Web-based Platform for Scholarly Video Analysis and Film Studies”, in: *International ACM SIGIR Conference on Research and Development in Information Retrieval* , SIGIR 2023 Demo Track.

Vonderau, Patrick (2017): “Quantitative Werkzeuge”, in: Hagener, Malte / Pantenburg, Volker (eds.) Handbuch Filmanalyse. Wiesbaden: Springer Reference Geisteswissenschaften. Springer Fachmedien Wiesbaden, 1–15. https://doi.org/10.1007/978-3-658-13352-8_28-1 [letzter Zugriff: 6. Dezember 2023].

Wittenburg, Peter / Brugman, Hennie / Russel, Albert / Klassmann, Alex / Sloetjes, Han (2006): “ELAN: a Professional Framework for Multimodality Research”, in *International Conference on Language Resources and Evaluation* , LREC 2006, 1556-1559. http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf [letzter Zugriff: 18. Juni 2023].

Panels

Bedeutung in Zeiten großer Sprachmodelle

Gengnagel, Tessa
tessa.gengnagel@uni-koeln.de
Cologne Center for eHumanities (CCeH), Universität zu Köln, Deutschland
ORCID: 0000-0001-8820-5112

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg
ORCID: 0000-0001-6944-6113

Kleymann, Rabea
rabea.kleymann@phil.tu-chemnitz.de
Technische Universität Chemnitz, Deutschland
ORCID: 0000-0003-3856-2685

Schröter, Julian
j.schroeter@lmu.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-0168-2608

Zinsmeister, Heike
heike.zinsmeister@uni-hamburg.de
Universität Hamburg, Deutschland
ORCID: 0009-0006-0505-7606

Die Performanz künstlicher Intelligenz ist, nicht zuletzt durch die großen Sprachmodelle (LLMs) in den letzten Jahren rasant angestiegen. Das hat zu einer intensiven Diskussion um die Definition anthropologisch relevanter Konzepte geführt; so wurde etwa die Diskussion des Begriffs „Intelligenz“ zu immer genaueren Bestimmungen genötigt. Die Arbeit in den Digital Humanities (DH) ist ganz direkt von diesen Entwicklungen betroffen, weil zentrale Konzepte der Geistes- und Kulturwissenschaften an der Schnittstelle zur KI affiziert sind. So wird der Bedeutungsbegriff aktuell für die Beschreibung von intelligenten Systemen mobilisiert, der zusammen mit den Begriffen „Interpretation“ und „Verstehen“ repräsentativ für ein hermeneutisches Selbstverständnis in den DH steht (vgl. Fickers 2020). [1] Mehr noch, über den Bedeutungsbegriff wird das disziplinäre Profil der DH verhandelt: „The search for truth and meaning remains a primary goal of scholarship and research, but it is articulated within a reflexive framework of discovery“ (Smithies 2017, 161). Dabei wird die Explikation von impliziten Bedeutungen unter Berücksichtigung historischer und kultureller Kontingenzen oftmals als eine zentrale Aufgabe der DH verstanden (vgl. Berry, Fagerjord 2017). Prominent rezipiert wurde in den DH vor allem die Phrase „how do we get from numbers to meaning“ (Heuser, Le Khac 2012, 46). Liu spricht in diesem Zusammenhang vom „meaning problem“ (2013, 411) der DH. Daran schließt sich eine fortwährende Diskussion über bedeutungstragende Einheiten bei statistischen Verfahren an (vgl. Gavin 2020).

Insbesondere die Entwicklung großer Sprachmodelle hat zuletzt eine Auseinandersetzung mit menschlicher und maschineller Sinnbildung provoziert (Kirschenbaum 2023), weswegen wir uns auf sprachliche Bedeutung, also die Bedeutung von Worten, Sätzen und Texten, konzentrieren werden. [2] Ziel unseres Panels ist es, danach zu fragen, wie Philosophie, Linguistik und andere Wissenschaften die Repräsentation, Konstitution und Konstruktion von Wort-, Satz- und Textbedeutung beschreiben. Durch Theorieimpulse wollen wir den DH eine angereicherte Beschreibungssprache zur Verfügung stellen, die es erlaubt, Differenzen zwischen den Bedeutungsprozeduren von Maschinen und Menschen genauer zu erfassen und somit auch zu bestimmen, wo scheinbar gleiche oder ähnliche Phänomene doch so unterschiedlich sind, dass die Applikation der KI fraglich und problematisch ist. Dazu ist es notwendig, die genauen technischen Grundlagen und das in den Operationen der Sprachmodelle vorliegende Wissen über Bedeutung auszuführen. Nicht zuletzt soll rekonstruiert werden, welche Formen der Bedeutungsanalyse und welche Prozeduren des Bedeutungsretrievals in Arbeiten aus dem Feld der Digital Humanities durchgeführt werden. [3]

Das Panel wird daher vier Perspektiven zusammenführen:

1. Sprachmodelle und Bedeutung: Funktionsweisen (Jannidis). Eine genauere Rekonstruktion der Bedeutungskonstitution in großen Sprachmodellen (z.B. BERT, GPT, LaMDA, OCRA etc.) soll der Frage nachgehen, welche Aspekte von Bedeutung hierbei präsent sind. Eine Antwort können insbesondere die Forschungen zu den Fehlleistungen der Modelle und etwa zu Fragen der Abstraktionsfähigkeit bieten. Einen guten Überblick über diese Diskussion geben Chang & Bergen 2023: Viele der Fehler sind auf Über- oder Untergeneralisierung textlicher Eigenschaften zurückzuführen. Sie erklären z.B. warum LLMs stärker auf sequentiellen Input als auf logisches und numerisches Kalkül reagieren. Erzeugen die LLMs also Bedeutung nur als situative Reaktion auf Input auf der Wort- und Satzebene? Oder ist die fehlerhafte Generalisierung der Trainingsdaten, also die “Lernerfahrung” des LLMs Indikator dafür, das “Bedeutung” hier als Kontextualisierung von Wörtern und Sätzen in Texten und zwischen Texten verstanden werden sollte? Neben solchen Analysen aus einer “Verhaltensforschung” der LLMs lassen mathematische Analysen der Modelle die “Bedeutung” in individuellen Knoten im neuronalen Netz, der Wahl der Parameter oder der Gestaltung des Attention heads suchen. Liegt die Bedeutungskonstitution der LLMs also eher in einer - vielleicht noch unverstandenen - Mathematik der Prozeduren, in denen Wörter, Sätze und Texte aufeinander bezogen werden?

Nicht zuletzt prüfen die Evaluationsinstrumente der Computerlinguistik und Informatik immer detaillierte Aspekte der semantische Verarbeitungsleistung von großen Sprachmodellen ab (z.B. Chang et al. 2023), wodurch ein informatives Beschreibungsinventar von Bedeutung entsteht.

2. Bedeutung und Sprache: Sprachphilosophie(Schröter). Die Frage, ob und in welchem Sinn von LLM generierte Texte Bedeutung haben, wurde jüngst in einem speziellen Forum des *Critical Inquiry* (CI) mit dem Titel »Again Theory: A Forum on Language, Meaning, and Intent in the Time of Stochastic Parrots« (Kirschenbaum 2023) als die zentrale literaturtheoretische Problemlage identifiziert, und zwar als ein Problem des Verhältnisses von Intention und Bedeutung. Im Zuge dieser Diskussion haben sich drei Positionen herauskristallisiert: (a) die intentionalistische Position, wonach Bedeutung in einer Sprecher:innenabsicht gegründet sein müsste, so dass KI-generierte Texte keine Bedeutung haben könnten (Bender et al. 2021, Knapp/ Michaels und Siraganian in Kirschenbaum 2023), (b) die Position, dass die Sprachmodelle anti-intentionalistische und poststrukturalistische Sprachtheorien bestätigen (Underwood in Kirschenbaum 2023), und (c) die Position, dass man es mit interpretationsbedüftigen Texten zu tun habe und deshalb von Schattierungen der Bedeutung zu sprechen sei (Bajohr in Kirschenbaum 2023). Ziel der hier vorgestellten Perspektive ist es, diese Diskussion auf den aktuellen Reflexionsstand zum Intentionalismus zu heben, um dann exemplarisch einzelne der im Licht der jüngsten technologischen Entwicklungen interessant gewordenen Probleme der Bedeutungskonstitution zu diskutieren. Dazu gehört die Unterscheidung zwischen Bedeutungszuschreibung im Kontext simulierter Kommunikationssituationen (wie etwa bei ChatGPT) mit einem Anschluss an Modellierungen kommunikativer Intentionen (Jannidis 2007) und Bedeutungszuschreibungen in Kontexten KI-generierter Spracherzeugung, die Kommunikationssimulationen überspringen oder ausblenden – wie etwa Bajohrs literarisches KI-Experiment »Die Zukunft der Gegenwart« (Bajohr 2023).

3. Bedeutung und Wissen: Modellierung (Gengnagel). Wie kann Bedeutung ohne ein ‘Wissen *um* etwas’ gegeben sein? Für die DH relevant erscheint nicht zuletzt das Verhältnis von dem Sinngehalt einer Aussage und deren Weltbezug. In Hinblick auf LLMs ist dieses Problem, reduziert auf Faktizität, vielfach als *Hallucination*-Phänomens besprochen worden (Maynez et al. 2020, Lin et al. 2022, OpenAI 2023). Abhilfe sollen “society of minds” Multiagent-Ansätze (Du et al. 2023) oder auch das Training auf strukturierten Triplen aus Knowledge Graphen wie Wikidata schaffen (Moiseev et al. 2022). Im Rahmen von DH-Modelltheorien drückt sich der Grundkonflikt in der Frage der Abbildbarkeit aus: McCartys Aussage einer “fundamental dependence of any computing system on an explicit, delimited conception of the world or ‘model’ of it” (McCarty 2005, S. 21) auf der einen und N. Katherine Hayles’ Feststellung, “[that] there are large gaps in the knowledge LLMs display, for they have no models of the world, only of language” (Hayles 2023, Kommentar) auf der anderen Seite. Gerade in den DH als Teil der Geisteswissenschaften, die auf ein Verständnis ganzheitlicher Deutungen (d.h. der Kontextualisierung und des Abgleichs) zielen, gilt es daher, einen möglichen Unterschied zwischen maschineller und menschlicher Sinnbildung unter Berücksichtigung eines vorhandenen Weltbezugs zu diskutieren.

4. Bedeutung und Interpretation: Textrepräsentationen (Kleymann) . Interpretation kann als ein Verfahren der Bedeutungszuschreibung verstanden werden (vgl. Lenk 2011; Jacke 2023). In einem iterativen und regelgeleiteten Prozess werden unter anderem Kontexte mit textuellen Entitäten verknüpft. In den Computational Literary Studies werden für solche Bedeutungszuschreibungen auch statistische und datenbasierte Ansätze eingesetzt. Aktuell lässt sich jedoch beobachten, wie der Wandel der Sprachmodelle von Bag-of-words-Repräsentationen und probabilistischen N-Grammen zu vektorbasierten Embedding-Repräsentationen scheinbar die Grenzen zwischen maschinellen und menschlichen Bedeutungszuschreibungen verschiebt (vgl. Biemann et al. 2022, 212). Während bspw. N-Gramme nur relativ lokale Abhängigkeiten modellieren, erfassen Embeddings deutlich größere Kontexte und können kohärente Texte generieren, indem sie die distributionelle Ähnlichkeit von Wörtern berechnen (vgl. Harris 1954). Bender et al. (2021, 615) betonen, dass Embeddings jedoch keinen Zugang zu Bedeutung haben. Dennoch werden solche Embeddings nicht nur für Verfahren der Bedeutungszuschreibung oder -explikation eingesetzt, wie z. B. für Zusammenfassungen, Vergleiche, Ermittlungen von Ähnlichkeit usw. Vielmehr imitieren auch KI-Chatbots dialogische Verstehensprozesse. Wie können wir als Forschende und Lehrende in den DH mit diesen Textrepräsentationen umgehen? Inwiefern stellen Embeddings geisteswissenschaftliche Bedeutungsverfahren auf die Probe?

5. Bedeutung und mentales Lexikon: Linguistik (Zinsmeister). Lappin (2023) argumentiert gegen Bender et al. (2021), dass LLMs nicht nur "stochastische Papageien" sind, die auf konkrete Prompts flüssig klingende Antworten aus gelernten Trainingsdaten synthetisieren. Probing-Experimente zeigen, dass LLMs in der Lage sind, komplexe semantische Zusammenhänge in Daten zu erkennen und Schlussfolgerungen zu ziehen, die z.B. auf hierarchischen Ober- und Unterbegriffen oder kausalen Zusammenhängen basieren. Eine interessante Frage ist dabei, inwieweit die Bedeutungsrepräsentation von LLMs als Modell für das menschliche, mentale Lexikon dienen kann bzw. umgekehrt, inwie-

weit Beschreibungskonzepte aus dem mentalen Lexikon auch im Kontext von LLMs Anwendung finden. Herausfordernd für eine solche Übertragung ist, dass es deutliche Unterschiede zwischen dem menschlichen Spracherwerb und dem Training von LLMs gibt, da LLMs z.B. wesentlich mehr Trainingsdaten benötigen als Kinder und der Lernprozess nicht interaktiv ist (außer ggf. dem nachgelagerten Schritt des sog. Reinforcement Learning), vgl. Lappin (2023), Abschnitt 3.1. Ein weiterer Aspekt ist die Tokenisierung, d.h. die sprachliche Segmentierung in Einheiten, für die eine Bedeutungsrepräsentation aufgebaut wird. Die derzeit leistungsfähigsten LLMs lernen Bedeutung nicht auf der Basis von Wörtern oder linguistisch motivierten Morphemen, sondern auf der Basis von häufigen Zeichenketten, die von beliebigen Teilwörtern bis zu Mehrwortfolgen reichen (sog. Byte-Pair-Encoding, Sennrich et al. 2015). Eine interessante Forschungsperspektive bietet hier das Probing auf der Grundlage konstruktionslinguistischer Form-Bedeutungs-Paare (vgl. Weissweiler et al. 2023).

Methodik und Ablauf des Panels

Nach einer gemeinsamen einführenden Einleitung werden alle Panelist:innen ihre oben skizzierten Perspektiven in 5–7-minütigen Statements erläutern. Auf diese Impulse wird eine 10-minütige Phase folgen, in der die Panelist:innen auf die Statements der anderen Diskussionsteilnehmer:innen reagieren können. Anschließend wird die Diskussion für das Publikum geöffnet, um eine engagierte Debatte zu ermöglichen. Je nach Publikumspartizipation soll so außerdem der Raum geschaffen werden, weitere relevante Aspekte einzubringen, so etwa aus dem Bereich der Leseforschung und Kognitionspsychologie.

Schlussbemerkung

Das Panel verspricht nicht nur menschliche und maschinelle Bedeutungsverfahren in den DH zu explorieren, sondern stellt auch einen ersten Versuch dar, ein geisteswissenschaftliches Vokabular für die Beschreibung und Evaluierung von intelligenten Systemen zu entwickeln. Insbesondere die Konjunktur des Bedeutungsbegriffes in den Datenwissenschaften (vgl. Donoho 2017, 746) macht eine systematische Auseinandersetzung mit der geisteswissenschaftlichen Begriffstradition erforderlich, um die Rolle der Geisteswissenschaften zukünftig zu vermessen. Vor dem Hintergrund der projektbasierten Arbeit in den DH stellt sich außerdem die Frage, wie sich diese in ihren Aufgaben und Zielen durch die Fortschritte in der generativen KI sowohl unmittelbar als auch langfristig verändern wird. Hierzu wird das Panel unter Einbeziehung des Konferenzthemas “Quo vadis?” wichtige Impulse in einer Zeit des Umbruchs liefern.

Fußnoten

1. Der Bedeutungsbegriff hat schon sehr breite Diskussionen angeregt, wie sie z.B. den beiden Workshops des Santa Fe Institutes zur künstlichen Intelligenz und der „Barrier of Meaning“, 2018 und 2023 dokumentiert sind. Vgl. auch Mitchell 2018,2019. https://www.santafe.edu/news-center/news/workshop-asks-will-ai-ever-crash-barrier-meaning und https://santafe.edu/events/ai-and-barrier-meaning-2.
2. Allgemein gilt für jede Disziplin in den Geisteswissenschaften, dass sie sich mit „Bedeutung“auseinandersetzt, wie Erwin Panofsky es beispielsweise grundlegend in der Kunstgeschichte getan hat (Panofsky 1955). Jüngste Entwicklungen in der Digital Art History werden u.a. in Hinblick auf eine “meaning trap“thematisiert (vgl. Offert 2023).
3. Wir danken Georg Vogeler, der wesentliche Beiträge zum Zustandekommen des Panels geleistet hat.

Bibliographie

Bajohr, Hannes. 2023. „Die Zukunft der Gegenwart“. In *metamorphosen*, herausgegeben von Michael Watzka und Moritz Müller-Schwefe, 31:8–15. Berlin: Verbrecher-Verlag.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major und Shmargaret Shmitchell. 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*, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Bergen, Benjamin. 2019. „Chapter 1: Embodiment“. In *Cognitive Linguistics - Foundations of Language*, herausgegeben von Ewa Dąbrowska und Dagmar Divjak, 11–35. Berlin, Boston: De Gruyter Mouton. https://doi.org/doi:10.1515/9783110626476-002.

Berry, David M., und Anders Fagerjord. 2017. *Digital Humanities: Knowledge and Critique in a Digital Age*. Oxford: Polity Press. https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=4875256.

Biemann, Chris, Gerhard Heyer, und Uwe Quasthoff. 2022. *Wissensrohstoff Text: eine Einführung in das Text Mining*. 2., Wesentlich überarbeitete Auflage. Lehrbuch. Wiesbaden [Heidelberg]: Springer Vieweg. https://doi.org/10.1007/978-3-658-35969-0.

Chang, Tyler A., und Benjamin K. Bergen. 2023. „Language Model Behavior: A Comprehensive Survey“. arXiv. http://arxiv.org/abs/2303.11504.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, u. a. 2023. „A Survey on Evaluation of Large Language Models“.

Donoho, David. 2017. „50 Years of Data Science“. *Journal of Computational and Graphical Statistics* 26 (4): 745–66. https://doi.org/10.1080/10618600.2017.1384734.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, und Igor Mordatch. 2023. „Improving Factuality and Reasoning in Language Models through Multiagent Debate“.

Fickers, Andreas. 2020. „Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?“ *Zeithistorische Forschungen – Studies in Contemporary History* 17 (1): 157–68. https://doi.org/10.14765/ZZF.DOK-1765.

Gavin, Michael. 2020. „Is there a text in my data? (Part 1): on counting words“. *Journal of Cultural Analytics*. https://doi.org/10.22148/001c.11830.

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, u. a. 2022. „Shared Computational Principles for Language Processing in Humans and Deep Language Models“. *Nature Neuroscience* 25 (3): 369–80. https://doi.org/10.1038/s41593-022-01026-4.

Harris, Zellig S. 1954. „Distributional Structure“. *WORD* 10 (2–3): 146–62. https://doi.org/10.1080/00437956.1954.11659520.

Hayles, N. Katherine. 2023. „Afterword: Learning to Read AI Texts“. *Critical Inquiry* Again Theory: A Forum on Language, Meaning, and Intent in the Time of Stochastic Parrots. https://critinq.wordpress.com/2023/06/30/afterword-learning-to-read-ai-texts/.

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*, Pamphlet, , Nr. 4. https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

Jacke, Janina. 2023. „Interpretation“. *Working Paper 2 der Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/WP_2023_006.

Jannidis, Fotis. 2007. „Zur kommunikativen Intention“. In *Im Rücken der Kulturen*, herausgegeben von Karl Eibl, 185–204. Poetogenesis. Paderborn: Mentis.

Kirschenbaum, Matthew. 2023. „Again Theory: A Forum on Language, Meaning, and Intent in the Time of Stochastic Parrots“. *Critical Inquiry. In the Moment*. https://critinq.wordpress.com/.

Lappin, Shalom. 2023. "Assessing the Strengths and Weaknesses of Large Language Models". *Journal of Logic, Language and Information*. (doi:10.1007/s10849-023-09409-x)

Lenk, Hans. 2011. „Deutung (Interpretation)“. In *Neues Handbuch philosophischer Grundbegriffe*, herausgegeben von Petra Kolmer und Armin G. Wildfeuer, 3:509–21. Freiburg im Breisgau: Karl Alber.

Lin, Stephanie, Jacob Hilton, und Owain Evans. 2022. „TruthfulQA: Measuring How Models Mimic Human Falsehoods“. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–52. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229.

Liu, Alan. 2013. „The Meaning of the Digital Humanities“. *PMLA* 128 (2): 409–23. https://doi.org/10.1632/pmla.2013.128.2.409.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, und Ryan McDonald. 2020. „On Faithfulness and Factuality in Abstractive Summarization“. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–19. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173.

McCarty, Willard. 2005. *Humanities Computing*. Basingstoke: Palgrave Macmillan. https://link.springer.com/book/9781403935045.

Mitchell, Melanie. 2018. „Opinion \textbar Artificial Intelligence Hits the Barrier of Meaning“. *The New York Times*, November. https://www.nytimes.com/2018/11/05/opinion/artificial-intelligence-machine-learning.html.

———. 2019. „Artificial Intelligence Hits the Barrier of Meaning“. *Information* 10 (2): 51. https://doi.org/10.3390/info10020051.

———. 2020. „On Crashing the Barrier of Meaning in Artificial Intelligence“. *AI Magazine* 41 (2): 86–92. https://doi.org/10.1609/aimag.v41i2.5259.

Moiseev, Fedor, Zhe Dong, Enrique Alfonseca, und Martin Jaggi. 2022. „SKILL: Structured Knowledge Infusion for Large Language Models“. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1581–88. Seattle, United States: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.113.

Offert, Fabian. 2023. „The Meaning Trap“. In *ChatGPT Und Andere»Quatschmaschinen«- Gespräche Mit Künstlicher Intelligenz*, herausgegeben von Anna Tuschling, Andreas Sudmann, und Bernhard J. Dotzler. Bielefeld: transcript Verlag.

OpenAI. 2023. „GPT-4 Technical Report“.

Panofsky, Erwin. 1955. *Meaning in the Visual Arts: Papers in and on Art History*. Anchor Books. Garden City, N.Y.: Doubleday.

Rahimi, Sadeq. 2019. „Extended Mind, Embedded AI, and “the Barrier of Meaning”“. In *IAW 2019 Interpretable AI for Well-Being: Understanding Cognitive Bias and Social Embeddedness*. CEUR Workshop Proceedings 2448. https://ceur-ws.org/Vol-2448/SSS19_Paper_Upload_221.pdf.

Rota, Gian-Carlo. 1985. „The Barrier of Meaning“. *Letters in Mathematical Physics* 10 (2): 97–99. https://doi.org/10.1007/BF00398144.

Sennrich, Rico, Barry Haddow, und Alexandra Birch. 2016. "Neural Machine Translation of Rare Words with Subword Units". In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), herausgegeben von Katrin Erk und Noah A. Smith, 1715–1725. Berlin, Germany:

Association for Computational Linguistics. (doi:10.18653/v1/P16-1162)

Smithies, James. 2017. *The Digital Humanities and the Digital Modern*. London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-49944-8.

Trott, Sean, Cameron Jones, Tyler Chang, James Michaelov, und Benjamin Bergen. 2023. „Do Large Language Models Know What Humans Know?“ *Cognitive Science* 47 (7): e13309. https://doi.org/10.1111/cogs.13309.

Weber, Arne M. 2017. „Klassische Kognitionswissenschaft“. In *Die körperliche Konstitution von Kognition*, herausgegeben von Arne M. Weber, 17–55. Wiesbaden: Springer Fachmedien. https://doi.org/10.1007/978-3-658-17219-0_2.

Weissweiler, Leonie, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, und Hinrich Schütze. 2023. "Construction grammar provides unique insight into Neural Language Models". In *Proceedings of the First International Workshop on Construction Grammars and NLP* (CxGs+NLP, GURT/SyntaxFest 2023), herausgegeben von Claire Bonial und Harish Tayyar Madabushi, 85–95. Washington, D.C.: Association for Computational Linguistics. (https://aclanthology.org/2023.cxgsnlp-1.10)

Xu, Qihui, Yingying Peng, Minghua Wu, Feng Xiao, Martin Chodorow, und Ping Li. 2023. „Does Conceptual Representation Require Embodiment? Insights From Large Language Models“.

Der Modelle Tugend 3.0 – Digitale 3D-Rekonstruktion als Forschungsraum und Transfermedium

Kuroczyński, Piotr
piotr.kuroczynski@hs-mainz.de
Hochschule Mainz, Deutschland
ORCID: 0000-0001-9847-8368

Münster, Sander
sander.muenster@uni-jena.de
Friedrich-Schiller-Universität Jena, Deutschland
ORCID: 0000-0001-9344-912X

Blümel, Ina
ina.bluemel@hs-hannover.de
Hochschule Hannover, Deutschland
ORCID: 0000-0002-3075-7640

Hoppe, Stephan
email@stephan-hoppe.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0002-8444-624X

Grellert, Marc
grellert@dg.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland
ORCID: 0000-0002-5587-902X

Die objektorientierte historische Forschung profitiert von den Entwicklungen in den Digital Humanities. Die Verwendung digitaler 3D-Modelle hat sich etabliert und traditionellen Darstellungsformen in den Forschungskontexten, wie analogen Modellen und zweidimensionalen Abbildungen und Texten, nicht nur ergänzt, sondern einen neuen Forschungs- und Vermittlungsraum geöffnet (Münster, 2022). Seit den frühen 1980er Jahren sind quellenbasierte, hypothetische 3D-Rekonstruktionen zunehmend zu Forschungsinstrumenten und unverzichtbaren Darstellungsmitteln geworden. Sie bieten neue Untersuchungsmethoden und ermöglichen neue Erkenntnisse für die objektbezogene Forschung (Messemer, 2020).

Die Forschungsergebnisse und die in diesem Zusammenhang stehenden digitalen 3D-Modelle sind jedoch aufgrund der unterschiedlichen Arbeits- und Modellierungsmethoden sowie vielfältiger Softwarelösungen oft nicht über das einzelne Projekt hinaus anwendbar bzw. wiederverwendbar. Darüber hinaus wird ein 3D-Modell in vielen (den meisten) Fällen nicht als wissenschaftliches Erkenntnisinstrument, sondern als reines Visualisierungswerkzeug gesehen, welches Ergebnisse in Form von Bildern veranschaulicht. Hinzu kommen die Zurückhaltung bei der Weitergabe der digitalen 3D-Modelle an sich und fehlende Standards in der Dokumentation und Veröffentlichung der 3D-Datensätze (Kuroczyński, 2018). Im Resultat werden Forschungsobjekte immer wieder aufs Neue virtuell rekonstruiert, ohne sich dabei auf vorausgegangene Projekte und früher erstellte 3D-Modelle beziehen zu können.

In diesem Zusammenhang sind in den letzten Jahren zahlreiche Initiativen und Forschungsprojekte entstanden, deren gemeinsames Ziel es ist, die verschiedenen von der Fachcommunity festgestellten Herausforderungen und Desiderate zu systematisieren und zu rationalisieren. Das war auch der Impuls für die Gründung der Arbeitsgruppe Digitale 3D-Rekonstruktion, die sich zum ersten Mal im Rahmen der 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (25.-28.02.2014, Universität Passau) traf. Das Panel Pecha Kucha – Virtuelle Rekonstruktion versammelte 2014 Kolleginnen und Kollegen, die sich dem Thema aus dem Blickwinkel der Architektur, Archäologie, Bau- und Kunstgeschichte sowie Computergraphik und Informatik verschrieben haben. Die Gründungsmitglieder der Arbeitsgruppe nutzten die Gelegenheit in Passau, um eine Plattform für einen engeren Austausch und eine feste Etablierung der digitalen 3D-Rekonstruktion des Kulturerbes

innerhalb der Digital Humanities einzurichten. Vorrangiges Ziel der Arbeitsgruppe war es, die Akteure im deutschsprachigen Raum zusammenzubringen, um sich den Fragen der Begriffsklärung und der Arbeitsmethodik sowie der Dokumentation, der Publikation und der Langzeitarchivierung von digitalen 3D-Rekonstruktionsmodellen zu widmen.

Die Arbeitsgruppe möchte die diesjährige Tagung zum Anlass nehmen, zehn Jahre ihres Wirkens und der Entwicklung der digitalen 3D-Rekonstruktion in der objektorientierten Forschung aus unterschiedlichen Perspektiven kritisch zu beleuchten. Die Teilnehmerinnen und Teilnehmer des Panels gehören zu den Gründungsmitgliedern der Arbeitsgruppe und stehen epistemologisch für unterschiedliche Schwerpunkte innerhalb der Auseinandersetzung mit dem Kulturerbe mittels digitalen 3D-Modellen innerhalb der Digital Humanities. Ganz dem diesjährigen Motto der Tagung in Passau verpflichtet, möchte das Panel die Frage nach den bisher zurückgelegten und für die Zukunft erkennbaren Wegen aufstellen. Dabei soll sowohl das Erreichte, als auch der heutige Stand und der Ausblick in kurzen Impulsvorträgen zur Diskussion und Reflexion anregen.

Ein wichtiges Ergebnis der erfolgreichen Zusammenarbeit der international vernetzten Arbeitsgruppe stellt die Veröffentlichung des Handbuchs zur digitalen 3D-Rekonstruktion historischer Architektur dar, welches aus dem DFG-Netzwerk Digitale 3D-Rekonstruktionen als Werkzeuge der architekturgeschichtlichen Forschung hervorgegangen ist (Münster et al., 2024). Aus der Arbeitsgruppe heraus konnten darüber hinaus zwei Infrastrukturprojekte auf dem Weg gebracht werden, die ebenfalls von der DFG gefördert und vorgestellt werden sollen. IDOVIR zur webbasierten Dokumentation des Rekonstruktionsprozesses (https://idovir.com/) und DFG 3D-Viewer zur nachhaltigen Publikation der 3D-Datensätze im Web (https://dfg-viewer.de/dfg-3d-viewer). Hinzu kommt die erfolgreiche Einrichtung einer begutachteten Buchreihe Computing in Art und Architecture bei arthistoricum.net in Kooperation mit dem Schwester-Arbeitskreis Digitale Kunstgeschichte (Kuroczyński et al., 2018 und 2019), deren inhaltliche Aus-einandersetzung mit der digitalen 3D-Rekonstruktion im Panel aufgenommen wird. Die einzelnen aufeinander abge-stimmten Beiträge möchten die wesentlichen Aspekte, Po-tenziale und Herausforderungen in einen breiteren Zusam-menhang bringen und beleuchten.

Das erste Statement von Piotr Kuroczyński stellt die Entwicklung und Bewertung einer anwendbaren Methodik für die hypothetische historische 3D-Rekonstruktion auf der Grundlage eines gemeinsamen theoretischen Ansatzes vor. Im Fokus des Impulses steht die Frage nach dem kleinsten gemeinsamen Nenner und dem Mehrwert einer Standardisierung für die Nachhaltigkeit und Wiederverwendbarkeit der Arbeitsergebnisse einer 3D-Rekonstruktion (Kuroczyński et al., 2023). Zur Sprache kommt dabei die Heterogenität hinsichtlich der Anforderungen an eine 3D-Rekonstruktion, die Vielfältigkeit der Datenformate und der Modellierungsmethoden sowie die Praxis bei der Doku-mentation und Veröffentlichung der Forschungsdaten.

Mögliche zur Diskussion und Reflexion gestellte Fragen: Inwieweit ist eine Standardisierung und Normierung einer digitale 3D-Rekonstruktion vorstellbar und umsetzbar, um als wissenschaftliche Forschungsmethode anerkannt zu werden? Wer ist die Community und worauf kann (muss) sie sich einigen, um Praxisregeln einer guten Forschung gerecht zu werden?

Das zweite Statement von Sander Münster nimmt die Frage nach der Methodik, der Dokumentation und Veröffentlichung auf und beleuchtet diese unter den derzeitigen Entwicklungen im Bereich von Computer Vision und der künstlichen Intelligenz. Dabei werden Bezüge zu den Forschungsprojekten auf nationaler und europäischer Ebene hergestellt. Hier wird u.a. die (semi-)automatische 3D-Rekonstruktion von ganzen Städten unter zu Hilfenahme von den digitalisierten, georeferenzierten Quellen in den Landes- und Universitätsbibliotheken und Archiven vorgestellt.

Mögliche zur Diskussion und Reflexion gestellte Fragen: Lassen sich 3D-Rekonstruktionen automatisiert erstellen und wo liegen epistemische, methodische und technologische Grenzen? Welche Trends und Perspektiven ergeben sich dafür auf europäischer Ebene?

Das dritte Statement von Ina Blümel nähert sich dem Thema aus der Perspektive von Open Science, Open GLAM und digitalen Forschungsinfrastrukturen, unter anderem der NFDI. Dabei wird der Blick auf die Erstellung und kontinuierliche Anreicherung des digitalen Kulturerbes mit den Anforderungen von Forschenden, GLAM-Expertinnen und -Experten und verschiedenen anderen Zielgruppen gelenkt. Von Bedeutung ist dabei auch die Darstellung der Kooperation zwischen den Produzierenden und Nutzenden von digitalen 3D-Materialien und die daraus abgeleiteten Anforderungen für nachhaltige Dienste.

Mögliche zur Diskussion und Reflexion gestellte Fragen: Wie lassen sich Werkzeuge für die digitale 3D-Rekonstruktion in enger Abstimmung mit Forschenden und ihren Bedarfen (weiter-)entwickeln? Wie bedingen sich Community-Building und FOSS Softwareentwicklung? Welche Trends und Perspektiven ergeben sich auf nationaler Ebene (NFDI) und darüber hinaus?

Das vierte Statement von Stephan Hoppe eröffnet die Perspektive des tatsächlichen Einsatzes der 3D-Rekonstruktion in der kunsthistorischen Forschung und Lehre. Dabei wird ausgehend von abgeschlossenen und laufenden Forschungsprojekten in der Kunstgeschichte der Mehrwert und die wissenschaftliche Relevanz von 3D-Modellen hinterfragt. Ein besonderes Augenmerk liegt dabei auf der Visualisierung und Simulation von historischen Zuständen verlorengegangener Kunst- und Architekturdenkmäler. Die Frage nach der Teilhabe, (Wieder-)Verwendbarkeit und (eigener) Steuerung der 3D-Rekonstruktion seitens der Kunsthistorikerinnen und Kunsthistoriker steht hier im Mittelpunkt.

Mögliche zur Diskussion und Reflexion gestellte Fragen: Wird die wissenschaftliche 3D-Rekonstruktion von der Kunstwissenschaft angenommen? Welche Publikationskanäle spielen hier eine Rolle? Gibt es disziplinäre Diffe-

renzmarker im Einsatz virtueller Rekonstruktionsmodelle, beispielsweise in Hinblick auf die Nachbarwissenschaft Archäologie? Wie wird die Wissenschaftskommunikation dadurch unterstützt?

Das fünfte Statement von Marc Grellert zeigt zum einen die Perspektive des breiten Einsatzes in der Vermittlung vom kulturellen Erbe in Museen und Dokumentationsfilmen. Damit einhergehend werden die Potenziale und Herausforderungen am Beispiel vergangener und laufender Projekte näher betrachtet. Zum anderen wird die Frage nach der Dokumentation der Entscheidungsprozesse bei digitalen 3D-Rekonstruktionen beleuchtet und der Umgang mit Unsicherheiten thematisiert. Gelingt es nicht das Wissen, das in den Rekonstruktionen eingebettet ist, nachhaltig zu sichern droht dessen Verlust und ohne öffentlich zugängliche Nachvollziebarkeit fehlt ein Pfeiler guter wissenschaftlicher Praxis.

Mögliche zur Diskussion und Reflexion gestellte Fragen: Wie umgehen mit dem Spannungsfeld von immer realistisch werdenden Rekonstruktionen bei gleichbleibender heterogener Wissensbasis? Welche ästhetischen Darstellungsformen existieren im Kontext der Abkehr von Eindeutigkeiten und Unsicherheiten? Wie kann es gelingen, die Dokumentation von Entscheidungsprozessen weiter zu etablieren? Welche Darlegungsformen der Plausibilität sind hilfreich?

Bibliographie

Kuroczyński, Piotr. 2018. “Neuer Forschungsraum für die Kunstgeschichte: Virtuelle Forschungsumgebungen für digitale 3D-Rekonstruktionen.“ In Kuroczyński, P., Bell, P. und Dieckmann, L. (Hg.): Computing Art Reader: Einführung in die digitale Kunstgeschichte, Heidelberg: arthistoricum.net-ART-Books, S. 160–181. https://doi.org/10.11588/arthistoricum.413.c5821

Kuroczyński, Piotr, Peter Bell und Lisa Dieckmann (Hg.). 2018. “Computing Art Reader: Einführung in die digitale Kunstgeschichte.“ In Heidelberg: arthistoricum.net-ART-Books (Computing in Art and Architecture, Band 1). https://doi.org/10.11588/arthistoricum.413

Kuroczyński, Piotr, Mieke Pfarr-Harfst und Sander Münster (Hg.). 2019. “Der Modelle Tugend 2.0: Digitale 3D-Rekonstruktion als virtueller Raum der architekturhistorischen Forschung.“ In Heidelberg: arthistoricum.net-ART-Books (Computing in Art and Architecture, Band 2). https://doi.org/10.11588/arthistoricum.515

Kuroczyński, Piotr, Fabrizio Ivan Apollonio, Igor Piotr Bajena und Irene Cazzaro. 2023. “Scientific Reference Model – Defining standards, methodology and implementation of serious 3d models in archaeology, art and architecture history.” In International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M 2–2023 29th CIPA Symposium “Documenting, Understanding, Preserving Cultural Heritage: Humanities and Digital Technologies for Shaping the Future”, 25–30 June 2023, Florence, Italy. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-895-2023

Messemer, Heike. 2020. “Digitale 3D-Modelle historischer Architektur: Entwicklung, Potentiale und Analyse eines neuen Bildmediums aus kunsthistorischer Perspektive.“ In Heidelberg: arthistoricum.net-ART-Books, https://doi.org/10.11588/arthistoricum.516

Münster, Sander. 2022. “Digital 3D Technologies for Humanities Research and Education: An Overview.” In Appl. Sci. 2022, 12(5), 2426; https://doi.org/10.3390/app12052426

Münster, Sander, Fabrizio Ivan Apollonio, Ina Bluemel, Federico Fallavollita, Riccardo Foschi, Marc Grellert, Marinos Ioannides, Peter Heinrich Jahn, Richard Kurdiovsky, Piotr Kuroczyński, Jan-Eric Lutteroth, Heike Messemer und Georg Schelbert. 2024. “Handbook of Digital 3D Reconstruction of Historical Architecture.” In Springer, Synthesis Lectures on Engineers, Technology, & Society 28. https://link.springer.com/book/9783031433627

DH – Cui bono? Zielgruppenerschließung für Digital Humanities und Cultural Heritage

Mayer, Simon
simon.mayer@onb.ac.at
Österreichische Nationalbibliothek, Österreich

Janjuš, Olja
olja.janjus@lmu.de
Ludwigs-Maximilian-Universität München, Deutschland
ORCID: 0009-0004-1498-495X

Ďurčo, Matej
matej.durco@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), Österreich
ORCID: 0000-0002-5274-8278

Hammer, Sophie
sophie.hammer@onb.ac.at
Österreichische Nationalbibliothek, Österreich

Windhager, Florian
florian.windhager@donau-uni.ac.at
Universität für Weiterbildung Krems, Österreich
ORCID: 0000-0002-5170-2243

Problemstellung

Das Arbeiten in den Bereichen Digital Humanities (DH) und Cultural Heritage (CH) bringt im Vergleich zu den klassischen Geisteswissenschaften verschiedene Herausforderungen mit sich: neue Daten, Methoden, Disseminationsmöglichkeiten, sowie auch neue Nutzer:innengruppen. Durch die beliebige Kombination dieser Aspekte werden sowohl an die DH als auch an die CH[1] sehr spezifische Anforderungen aus dem jeweiligen Forschungskontext gestellt, welche einzelne Vorhaben und Projekte individuell umsetzen.

Es werden unterschiedliche Zugänge erarbeitet, die sich im Normalfall an einer gezielten Gruppe von Nutzer:innen orientieren (vgl. Könneker 2017 und Ziegler/Fischer 2020). Gerade wenn digitale Publikationsmöglichkeiten gewählt werden, sollte der Veröffentlichung eine umfangreiche Zielgruppenanalyse vorangegangen sein, was aber in Realität – sofern nicht integraler Bestandteil der Forschungsfrage – oft vernachlässigt wird. Ebenfalls wird nur vereinzelt der Austausch zwischen den unterschiedlichen, heterogenen Zielgruppen (z.B. Laien und Expert:innen) forciert, da die Reichweite des jeweiligen Forschungsprojekts meist auf sehr spezifische (Forschungs-)Communities beschränkt ist. Stattdessen werden mit großem Entwicklungsaufwand individuelle Lösungen für einzelne Forschungsgebiete implementiert, die aber selten über den sogenannten Tellerrand hinausschauen. Um die Wirksamkeit von Forschungsprojekten in den DH und CH zu maximieren, ist zudem entscheidend, die FAIR-Prinzipien (go-fairund Wilkinson u. a. 2016) zu berücksichtigen, Openness zu fördern (UNESCO 2021 und Goodrey u. a. 2022) und die Voraussetzungen für eine nachhaltige Nachnutzung zu schaffen, was oft aufgrund der Rahmenbedingungen von Projekten und Vorhaben nicht ausreichend beachtet werden kann.

Der Frage, ob und wie die Zielpublika aus unterschiedlichen Teilaspekten der digitalen Geisteswissenschaften am besten in den DH-Diskurs integriert werden können, soll anhand ausgewählter Aspekte in einer Diskussion nachgegangen werden: In diesem Sinne sollen die in Folge angeführten Fragen dazu dienen, ansatz-übergreifende Herausforderungen mit den Forscher:innen des Panels zu diskutieren. Die am Panel teilnehmenden Personen decken mit der Expertise aus ihren jeweiligen Projekten und Forschungsvorhaben ein breites Feld an Zugängen innerhalb der DH & CH ab, wie etwa Digitale Editionen (aus den Geschichts- und Musikwissenschaften), Visualisierung (vgl. Windhager u.a. 2019), aber auch partizipative Zugänge (z.B. Hackathons, s. Briscoe/Mulligan 2014 und Mischke u.a. 2022) sind vertreten, welche Künstler:innen, Musiker:innen, Forscher:innen oder interessierte Laien direkt einbinden, indem ihnen die Möglichkeit zur Mitarbeit durch Bearbeitung oder Interpretation ausgewählten Materials eröffnet wird.

Leitfragen und Diskussionspunkte

Die diversen, transdisziplinären Perspektiven der Panelist:innen sollen genutzt werden, um gemeinsam über zentrale Fragen im Bereich der Öffentlichkeitseinbindung im Kontext des DH-Diskurses zu diskutieren. Hierbei abgedeckte Themenkomplexe reichen von der Adressierung und Auswahl des Zielpublikums, über die Wahl geeigneter zielgruppenorientierter Medien, nötige Anpassungen und damit verbundene Einschränkungen, bis hin zu Fragen zur Zugänglichkeit, Nachhaltigkeit und Archivierbarkeit der erzeugten Ergebnisse.

Die Teilnehmer:innen stellen sich zunächst anhand kurzer Statements vor. Im Anschluss folgt eine offene Diskussion, die sich an den folgenden Leitfragen orientiert:

Zielgruppenauswahl:

- Welche Faktoren beeinflussen die Auswahl der Zielgruppe, insbesondere im Hinblick auf das zugrunde liegende Material, begleitende Fragestellungen und verfügbare Ressourcen?
- Wie können Kooperationsmöglichkeiten mit Interessierten genutzt werden, um verschiedene Aspekte im Zusammenhang mit den Projekten abzudecken?

Einbindungsstrategien:

- Welche Erwartungen und Ziele stehen im Mittelpunkt der Einbindung von Zielgruppen?
- Wie sollte die Rolle des Zielpublikums im Gesamtprozess definiert werden (aktive gestaltende Teilnahme versus Informationskonsum)?
- Welche Kanäle eignen sich am besten zur Erreichung unterschiedlicher Nutzergruppen?
- Welche Limitationen und Herausforderungen ergeben sich bei der Anwendung von DH-Ansätzen, insbesondere im Hinblick auf Barrierefreiheit und Inklusion von Personen mit geringer Digitalkompetenz?

Zielgruppengerichtete Aufbereitung von Daten:

- Ab welchem Punkt kann die zielgruppengerichtete Aufbereitung von Daten durch die vereinfachte Darstellung als Verfälschung des Gesamtbilds betrachtet werden?
- Wie kann eine ausgewogene Balance zwischen der Bereitstellung uninterpretierter Materialien[2] für die Wissenschaft und der Bedürfnisse von Nicht-Experten gefunden werden?

- Wann und in welcher Form sollten zusätzliche, detailliertere Informationen für die Forschung bereitgestellt werden?

Materialität der Medien und Transformationen:

- Welche Vor- und Nachteile ergeben sich aus der Nutzung von Originalmaterial im Vergleich zu durch in DH-Vorhaben interpretierten Materialien hinsichtlich Nachhaltigkeit, Zugänglichkeit und Archivierbarkeit?
- Wie können Transformationen von Medien im DH-Vorhaben optimiert werden, um eine effiziente Nutzung und Bewahrung zu gewährleisten?

Erwartungen an das Panel / Perspektiven

Die Erwartungen an das Panel lassen sich in zwei Bereiche aufteilen, zum einen die unmittelbaren Erwartungen an die Diskussion und Präsentation während der DHd2024 und zum anderen die erwarteten Effekte für die beteiligten Projektgruppen.

Das vorgeschlagene Panel selbst soll als eine sichtbare Schnittstelle für eine Diskussion des Selbstverständnisses der DH dienen. Forschende sollen aktiv und gezielt mit unterschiedlichen Zielgruppen zusammengebracht werden, um so einen Austausch und daraus resultierende Synergieeffekte zwischen fachlich verschiedenen Gruppen zu nutzen, was in gängigen Projektplänen in der Regel nicht vorgesehen ist.

Die Diskussion soll eine Sensibilisierung für die Wichtigkeit der Zielpublikums-Frage hervorrufen. Dies soll dazu anregen, die Analyse und Auswahl des Zielpublikums hinsichtlich zukünftiger Vorhaben transparenter zu gestalten. Ebenso könnte dies bei bereits laufenden Projekten während der verbleibenden Projektlaufzeit berücksichtigt werden, beispielsweise durch einen evaluativen Prozess. Durch die offene Diskussion soll eine Bewertung und/oder Bestätigung des Ansatzes zur Dissemination der jeweiligen Vorhaben geschehen, jedoch sollen auch mögliche Änderungen von Strategien nicht ausgeschlossen werden. Insbesondere in Bezug auf zum Beispiel Visualisierungen ist eine Schärfung des Profils der einzelnen Vorhaben erwünscht.

Der angestoßene Diskurs soll sowohl die Panelist:innen als auch das Publikum in Hinblick auf ein breiteres Verständnis bezüglich der Zielgruppen-Überschneidungen einzelner Vorhaben und Projekte aufklären. Dies soll dazu beitragen, dass die DH-Forschungsvorhaben von den von Anderen gewählten Ansätzen lernen und ihre Zielgruppen dementsprechend auf die größte Schnittmenge – und möglicherweise darüber hinaus – erweitern können.

Vorstellung der einzelnen Vorhaben

Als Teilnehmer:innen am Panel wurden Wissenschaftler:innen aus unterschiedlichen Vorhaben und Projekten eingeladen, die im Folgenden jeweils kurz beschrieben werden. Durch die ausgewählten Projekte wird speziell auf die Vielfältigkeit der DH und des CH-Bereiches und deren Zielpublikum eingegangen.

Bibliotheca Eugeniana Digital

Das Projekt Bibliotheca Eugeniana Digital setzt sich die digitale Rekonstruktion und visuelle Darstellung der Büchersammlung Prinz Eugens von Savoyen zum Ziel. 1738 wurde die Sammlung Teil der habsburgischen Hofbibliothek, und stellt damit heute einen wichtigen Teil des Grundbestandes der Österreichischen Nationalbibliothek (ÖNB) dar.

Die Sammlung wurde wenig später im Mitteloval des Prunksaals aufgestellt, wo sie sich großteils bis heute befindet. Im Laufe der Jahrhunderte unterlag die Hofbibliothek jedoch ständigem Wandel, sodass die bis heute verbreitete Angabe, im Mitteloval des Prunksaals könne die Sammlung Prinz Eugens bewundert werden, obsolet ist.

Erst die unlängst erfolgte Digitalisierung des manuell nicht bewältigbaren Korpus an Druckwerken, zusammen mit Werkzeugen und Methoden aus den Digital Humanities und den Data Sciences, ermöglichen eine systematische Aufarbeitung der Sammlung, die in Folge durch verschiedene visuelle Explorationsmöglichkeiten unterschiedliche Zielpublika ansprechen soll.

E-LAUTE

E-LAUTE - Electronic Linked Annotated Unified Tablature Edition erstellt eine elektronische Notenedition von nicht mehr benutzten, aber nicht seltenen deutschen Lautentabulaturen. Diese spezielle Notation wurde vor allem im deutschsprachigen Raum im 15. und 16. Jahrhundert verwendet. Heute sind die Aufbewahrungsorte der in der deutschen Lautentabulatur geschrieben Manuskripte über ganz Mitteleuropa verstreut und der Bestand nur punktuell erschlossen und erforscht. Das Projekt E-LAUTE schafft daher durch die Edition einen Zugang sowohl für Wissenschaftler:innen und professionelle Musiker:innen als auch für einen breiten Kreis von Interessierten. Mit der “open knowledge platform” wird eine neuartige Form der Musikedition entwickelt, in der Musikwissenschaft, Musikpraxis, Musikinformatik und Literaturwissenschaft ineinandergreifen und traditionelle Editionsmethoden in disziplinärer und interdisziplinärer Forschung vernetzen. Computergestützte Technologien wie Enkodierung, Verlinkung, Erkennung (Optical Musical Recognition) und automatische Transkription werden eingesetzt.

ExploreSalon 2023

Der einwöchige Cultural Hackathon " ExploreSalon: Unveil Hidden Stories from the Past" erkundete in einem kollaborativen Setting digitalisierte Erinnerungen (Briefe, Tagebücher, etc.) in Form von kuratierten biografischen, räumlichen und zeitlichen Datensätzen.

Das ExploreSalon-Konzept zielt darauf ab, kreative Köpfe mit unterschiedlichen Fähigkeiten und Fachkenntnissen aus den Bereichen DH, CH sowie GLAM zusammenzubringen, um innovative Wege des datenbasierten Storytellings zu entdecken. Im Vordergrund stehen Kreativität als Form des generativen Wandels und der Aufbau einer Community of Practice durch ein Raum-Geben für neue Ideen und gemeinsame Ziele. Kommunikation auf Augenhöhe und partizipative Führung werden anstelle spezifischer Methoden propagiert.

Offene Plenarsitzungen und umfangreiche Unterstützung durch das Betreuungsteam ermöglichen kollaborativen Austausch; der zweckgerichtete Ansatz und eine Loslösung von starren (Zeit-)Rahmen fördern die eigenverantwortliche Steuerung des Arbeitsprozesses.

Open Digital Libraries for Creative Users

Das Projekt "Open Digital Libraries for Creative Users" ist ein aus Mitteln des EU-Programmes "Creative Europe" kofinanziertes Partnerprojekt der Österreichischen Nationalbibliothek mit der Niederländischen und der Estnischen Nationalbibliothek. Ziel ist es, die kreative Nutzung digitaler Bibliotheksbestände zu fördern. Im Rahmen des Projekts wurden eine Reihe von "artistic experiments" durchgeführt – Experimente, die verschiedene Formen der Zusammenarbeit zwischen Künstler:innen/Kunststudierenden und der Bibliothek erprobten und gleichzeitig eine Wissensbasis aufbauten, die mit anderen Bibliotheken geteilt werden kann. Künstler:innen waren eingeladen, sich kreativ und kritisch mit ausgewählten digitalen Bibliotheksbeständen auseinanderzusetzen und web-basierte Arbeiten zu entwickeln. Die Ergebnisse der drei Programme sind im virtuellen Kunstraum der ÖNB Labs (Artspace) zu sehen. Ein wesentliches Anliegen der Projektpartner ist die Vermittlung der im Projekt gewonnenen Erfahrungen innerhalb der eigenen Institutionen und in Netzwerken der GLAM-Community.

Fußnoten

1. Im Folgenden betrachten wir nicht nur Problemstellungen aus den DH, denn der enge Zusammenhang zu CH erlaubt es den Problemaufriss in diese Richtung zu erweitern.
2. Mit "uninterpretierten Materialien" bezeichnen wir solche, die im Prozess weder durch eine bestimmte Fragestellung sortiert noch annotiert wurden.

Bibliographie

Bibliotheca Eugeniana Digital". Online. https://labs.onb.ac.at/bed/ (zugegriffen: 19. Juli 2023)

Briscoe, Gerard und Catherine Mulligan. 2014. "Digital Innovation: The Hackathon Phenomenon." In *Creativeworks London Working Paper* No.6. http://www.creativeworkslondon.org.uk/wp-content/uploads/2013/11/Digital-Innovation-The-Hackathon-Phenomenon1.pdf (zugegriffen: 19. Juli 2023)

Coding Da Vinci". Online. https://codingdavinci.de/ (zugegriffen: 19. Juli 2023)

Coding Dürer - International Hackathon for Art History". Online. https://codingdurer.de/(zugegriffen: 19. Juli 2023)

Corrigan, Chris. 2016. "The chaordic stepping stones. A planning tool for designing participatory processes." https://www.chriscorrigan.com/parkinglot/wp-content/uploads/2015/01/Chaordic-Stepping-Stones-1.pdf (zugegriffen: 19. Juli 2023)

E-LAUTE - Electronic Linked Annotated Unified Tablature Edition". Online. https://e-laute.info/de/ (zugegriffen: 19. Juli 2023)

ExploreSalon 2023". Online. https://clariah.at/exploresalon2023 (zugegriffen: 19. Juli 2023)

FAIR-Prinzipien". Online. https://www.go-fair.org/fair-principles/(zugegriffen: 6. Dezember 2023)

Goodey, Gregory, Mark Hahnel, Yuanchun Zhou, Lulu Jiang, Ishwar Chandramouliswaran u.a. (2022). "The State of Open Data 2022". In *Digital Science. Report.* https://doi.org/10.6084/m9.figshare.21276984.v5

Hammer, Sophie und Martin Krickl. (in Druck). "Kunst in/aus Bibliotheken – Kreative Nutzung von digitalen Bibliotheken." In *Bibliothek Forschung und Praxis.*

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.

Krickl, Martin, Simon Mayer und Emanuel Zangger. 2022. "Mit Machine Learning auf der Suche nach Provenienzen – ein Use Case der Bildklassifikation an der Österreichischen Nationalbibliothek." In *Bibliothek Forschung und Praxis*, 46(1), 227-238. https://doi.org/10.1515/bfp-2021-0090

Mauthe, Gabriele. 2010. "Die Bibliotheca Eugeniana im europäischen Zeitvergleich." In *Prinz Eugen. Feldherr, Philosoph und Kunstfreund. Katalogbuch zur Ausstellung im Belvedere, hg. von Agnes Husslein-Arco und Marie-Louise von Plessen,* 190-220. Wien: Hirmer Verlag.

Mischke, Dennis, Peer Trilcke und Henny Sluyter-Gäthje. 2022. "Hackathons als kollektiv-kreative Bildungsereignisse. Ein Konzept zur Gestaltung offener Lehrveranstaltungen in den Digital Humanities." In *DHd*

2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2022). Potsdam.

ONB Labs Artspace". Online. https://labs.onb.ac.at/artspace/(zugegriffen: 6. Dezember 2023)

Open Digital Libraries for Creative Users". Online. https://open-digital-libraries.nl/ (zugegriffen: 19. Juli 2023)

OpenGLAM". Online. https://www.openglam.at/ (zugegriffen: 19. Juli 2023)

Simon, Tobias, Sven Pagel, und Harald F. O. von Korflesch. 2020. "Influencing Factors for Acceptance of Digital Tools in the Humanities". In *Proceedings of Mensch Und Computer 2020*, 17–27. Magdeburg, Germany: ACM. https://doi.org/10.1145/3404983.3405524

UNESCO Recommendation on Open Science". Online. https://doi.org/10.54677/MNMH8546 (zugegriffen: 6. Dezember 2023)

Wilkinson, M., M. Dumontier, I. Aalbersberg u.a . 2016. "The FAIR Guiding Principles for scientific data management and stewardship". In *Sci Data* **3**, 160018. https://doi.org/10.1038/sdata.2016.18

Windhager, Florian u . a. 2019. "Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges". In *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 6, pp. 2311-2330, 1 June 2019, doi: 10.1109/TVCG.2018.2830759.

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*.

Quo Vadis kulturwissenschaftliche Digital Humanities?

Dietzsch, Ina
ina.dietzsch@staff.uni-marburg.de
Universität Marburg, Deutschland

Franken, Lina
lina.franken@uni-vechta.de
Universität Vechta, Deutschland
ORCID: 0000-0002-2587-4068

Imeri, Sabine
sabine.imeri.1@ub.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID: 0000-0002-8844-4014

Kinder-Kurlanda, Katharina
katharina.kinder-kurlanda@aau.at
Universität Klagenfurt, Österreich
ORCID: 0000-0002-7749-645X

Sørensen, Estrid
estrid.sorensen@rub.de
Ruhr-Universität Bochum, Deutschland
ORCID: 0000-0002-3131-9415

Vepřek, Libuše Hannah
libuse.veprek@uni-tuebingen.de
Universität Tübingen, Deutschland
ORCID: 0000-0002-0672-6633

Kulturwissenschaftliche und insbesondere empirisch-kulturwissenschaftliche Perspektiven sind in den deutschsprachigen Digital Humanities (DH) auch zehn Jahre nach der ersten DHd-Konferenz noch eine Ausnahme. Der Fachzusammenhang Empirische Kulturwissenschaft (EKW, je nach Standort auch als Europäische Ethnologie, vergleichende Kulturwissenschaft oder Kulturanthropologie bezeichnet) basiert auf einem weiten Kulturbegriff, das heißt auf einem umfassenden Verständnis von Kultur als Gesamtheit der Lebensweisen und Praktiken, die den Alltag von Menschen bilden (Heimerdinger/Tauschek 2020; Hinrichs/Röthl/Seifert 2021). Diesem folgend, untersucht die EKW Alltage in Vergangenheit und Gegenwart, oft mit ethnografischen Methoden und verbindet dabei historische und qualitativ-sozialwissenschaftliche Zugänge (Bischoff/Oehme-Jüngling/Leimgruber 2014; Hess/Moser/Schwertl 2013). Mit ihrem Fokus auf menschlichem Handeln und dessen Bedeutungen in der Gegenwart, gleichwohl aber auch auf einer historischen wie begrifflichen Verortung als Kulturwissenschaft, versteht sich der Fachzusammenhang aktuell nur in Teilen den DH zugehörig und öffnet sich mit seinen spezifischen Bedürfnissen und Ansätzen eher zögerlich hin zu den entsprechenden Communities und Perspektiven. Das Panel öffnet sich auch hin zu einem übergreifenden Verständnis von Kulturwissenschaften, wie es in vielen Bereichen der Geisteswissenschaften vertreten wird.

Obwohl die ethnografisch arbeitenden Disziplinen sich in einem Dazwischen von DH und Computational Social Sciences (CSS) befinden und beiden Bereichen zugerechnet werden können (Franken 2022b), bringen sie ihre spezifischen Bedürfnisse bisher noch relativ wenig in diese Zusammenhänge ein. Insbesondere da die eigenen Materialen bzw. Forschungsdaten in der Regel erst im Zuge der (Feld-)Forschung erhoben und damit generiert werden, sind die gegenwartsbezogenen Korpora anders gelagert als viele DH-Korpora. Da in den CSS wiederum quantitative Datensets überwiegen, die aus sogenannten Big Data entstehen, sind die ethnografischen Erhebungen bisher wenig praktikabel mit den entsprechenden Zugängen umsetzbar, zumal mit multimodalen Datensätzen.

Sowohl die DH als auch die EKW können dabei von einer engeren Zusammenarbeit profitieren. Digital Humanities werden hier verstanden als Transformations- und Dachdisziplin im Sinne eines "Big Tent" (Svensson 2016). Wo die (empirischen) Kulturwissenschaften ihren Platz innerhalb dieses "großen Zeltes" mit pluralen Stimmen haben, gilt es noch auszuloten – bisher stehen sie eher am Rande oder sogar außerhalb. In empirisch-kulturwissenschaftlicher Forschung werden aktuell eher etablierte Verfahren wie Topic Modeling oder relationale Datenbanken als die Ansätze angesehen, die wahrgenommen und umgesetzt werden können, etwa um für empirisch-kulturwissenschaftliche Fragestellungen relevantes Material zu identifizieren und zu interpretieren (Egger et al. 2023; Koch & Franken 2020). Möglichkeiten, wie sie etwa mit generativer KI und Machine Learning bestehen, sind hingegen bisher kaum als Methode der EKW diskutiert worden, u.a. weil insbesondere ethnografische Forschungsdaten sensibel sind, und bisher in Frage steht, ob überhaupt und wenn ja, in welcher Form, sie in entsprechende Large Language Models eingebunden werden können oder sollten. In diesem Sinne haben ethische Fragestellungen eine hohe Relevanz, obwohl die Gründe für diese Zögerlichkeit noch stärker zu reflektieren sind.

Gleichzeitig kann es für die DH besonders produktiv sein, in Analyseprozesse oder in Fragen der Dokumentation von Schritten im Forschungsprozess (kritische) Reflexionsprozesse einzubinden, wie sie in empirisch-kulturwissenschaftlicher Forschung ohnehin etablierte Praxis sind. "Daten" ist hier ein Begriff mit Implikationen und Setzungen, der (produktive) Spannungsfelder erzeugt. Auch die (veränderte) Relevanz von Algorithmen ist dabei nicht Mittel zum Zweck, sondern Teil des Erkenntnisinteresses. Methodische Reflexion und Kombination unterschiedlicher Ansätze in und aus den empirischen Kulturwissenschaften können in den DH bestehende Ansätze aber auch weiterentwickeln, weil sie Komplexitäten spezifisch hin zu Praktiken und Alltagen denken und Phänomene situiert in ihren sozio-historischen und -kulturellen Kontexten untersuchen. Gleichzeitig werden insbesondere mit Blick auf das in empirisch-kulturwissenschaftlichen Forschungen verwendete heterogene Quellenmaterial – etwa Text, Bilder, Objekte, Videos, Materialien aus dem Web oder eigene Feldnotizen – die Anwendung einzelner Verfahren ebenso wie Anforderungen an die dafür benötigten Infrastrukturen komplexer. Besonders relevant ist dabei die Abbildung von Ambiguitäten in um DH-Verfahren erweiterten Forschungsprozessen (vgl. Franken 2022a). Zudem ist bisher weitgehend offen, wie Kontexte der Datenproduktion, Kritik und mögliche ethische Vorbehalte abgebildet werden können, wenn Metadaten hierfür nicht ausreichen.

Gerade die in größeren Forschungsprojekten der EKW entstehenden Forschungsdaten sowie neue Perspektiven auf bestehendes Material und historische Quellen sind für DH-Ansätze vielversprechend, allerdings noch kaum zusammengedacht. In den letzten etwa 20 Jahren hat sich eine digitale Anthropologie stärker als Weiterentwicklung von ethnografisch-feldforschenden Zugängen zu Praxen etabliert, die in Online- und Offline-Kontexten verwoben sind (u.a. Horst & Miller 2020; Pink et al. 2016; Koch 2017; Fleischhack 2019). Insbesondere Aktivitäten auf Social Media Plattformen werden mit diesen Zugängen in ihren Kontexten beforscht, aber auch sich zu großen Teilen im Internet entfaltende Phänomene wie etwa digitale Protestkulturen. In der Regel geschieht dies jedoch mit manuellen und nicht automatisierten Ansätzen (vgl. Franken 2023: 196-200). Die methodischen Weiterentwicklungen lösen sich jedoch zunehmend von diesen Dimensionen bzw. ergänzen sie (vgl. die Beiträge in Klausner/Eckhardt 2023) und arbeiten etwa mit DH-affinen Umsetzungen zur Analyse von Chatverläufen (Vepřek 2023) oder Interviewtranskripten (Franken et al. 2023). Als stärkerer experimentell-ethnografischer Zugang zu Daten wird zudem mit Data Walks gearbeitet: Diese ermöglichen eine kritische Auseinandersetzung mit Daten und ihrer Produktion sowie Erfahrbarkeit und kombinieren diese mit klassisch ethnografischen Erhebungsverfahren des Beobachtens und der Bewegung im Feld sowie experimenteller Daten- und Wissensproduktion etwa in Form von Visualisierungen (Amelang/Klausner/Sørensen/Straube 2023). Zudem bestehen explizit ethnografisch-anthropologische Entwicklungen von Infrastrukturen für Forschungsprojekte wie die Platform for Experimental Collaborative Ethnography (PECE, Fortun et al. 2014), welche standardkonformes Forschungsdatenmanagement und kollaborative Analysen der betreffenden Daten mit einer multimodalen Präsentation von Forschungsergebnissen kombiniert. Bisher sind auch diese Infrastrukturen nur in Ansätzen mit DH-Verfahren zusammengedacht, wenngleich das Potenzial offensichtlich ist. Auch zum Forschungsdatenmanagement liegen differenzierte fachspezifische Positionen vor (DGEKW 2023). Sowohl Methoden, digital verfügbare Daten als auch zugehörige Infrastrukturen sind jedoch noch deutlich ausbaubar.

Für die DH können zudem technikanthropologische Ansätze (Beck/Niewöhner/Sørensen 2012) fruchtbar sein, etwa mit Blick auf epistemologische Weiterentwicklungen der DH im Zuge der Anwendung und Entwicklung von Daten und Computercode. Sie ermöglichen eine Erweiterung und Fundierung der bestehenden Ansätze zur Theoretisierung von DH (etwa bei Burghardt 2023; Geiger 2023; Geiger/Pfeiffer 2020) hin zu Perspektiven darauf, dass Daten und Methoden nie neutral sind und Forschung stets positioniert ist. Die sich in den DH und darüber hinaus verändernden Forschungsprozesse können durch diese Perspektiven anders verstanden und kontextualisiert werden. Zentral sind dabei Überlegungen zu Computercode als Teil von Assemblages (Carlson et al. 2021; Amelang/Bauer 2019; Vepřek/Thanner/Franken 2023), zu Daten in ihren Verwobenheiten (Kinder-Kurlanda 2020), zu Infrastrukturen als emergenten, sozialen wie materiellen Phänomenen (Niewöhner 2015) oder zu mehr-als-digitalen Praxen (Klausner 2022). Bisher sind diese Perspektiven noch wenig mit den Debatten in den DH verbunden. Das gilt ähnlich mit Blick auf die Reflexion der Historizität von Standards, Regelwerken und Forschungsmethoden (Moeller et al. 2022), disziplinä-

rer “data ideologies” (Poirer et al. 2020) sowie allgemeiner ethischer Fragestellungen.

Das Panel diskutiert erstmals grundlegend, wie empirisch- kulturwissenschaftliche DH ausgestaltet ist und welche Richtungen hier künftig eingeschlagen werden können. Dabei werden konkrete methodische Entwicklungen ebenso thematisiert wie konzeptionell-theoretische Überlegungen. Insbesondere folgende Perspektiven werden sowohl unter den Teilnehmenden als auch mit dem Publikum diskutiert:

- Was sind relevante Bereiche der DH in den empirisch-kulturwissenschaftlichen Disziplinen?
- Welche Rolle spielen die DH aktuell in den empirischen Kulturwissenschaften? Wie könnte sich dies in näherer Zukunft ändern und welche Schritte sind dafür zielführend?
- Was hindert die empirischen Kulturwissenschaften an einer stärkeren Einbindung in die DH?
- Was sind die neuen Perspektiven, die empirisch-kulturwissenschaftliche Forschung in die DH einbringen?
- Wo können die empirischen Kulturwissenschaften Perspektiven der DH übernehmen, welche gilt es gemeinsam weiterzuentwickeln?

Ziel ist es, sowohl methodologische Potentiale zu identifizieren als auch theoretisch-konzeptionelle Perspektiven herauszuarbeiten, die kulturwissenschaftliche DH weiter ausdifferenzieren. Das Panel startet mit kurzen Inputs der beteiligten Personen, gefolgt von einer Plenumsdiskussion, in die das Publikum aktiv eingebunden wird.

Das Panel setzt sich aus folgenden Personen zusammen:

- Prof. Dr. Ina Dietzsch (Europäische Ethnologie / Kulturwissenschaft, Marburg) geht auf die Frage ein, wie mit ethnografischen Plattformen infrastrukturiert werden kann und betrachtet dies am Beispiel der Platform for Experimental Collaborative Ethnography (PECE). Darüber hinaus zeigt sie auf, wie feministische Technikanthropologie in den DH fruchtbar gemacht werden kann.
- Dr. Sabine Imeri (Fachinformationsdienst Sozial- und Kulturanthropologie, Berlin) stellt die besondere Bedeutung von Forschungsdaten in der empirischen Kulturwissenschaft dar und geht darauf ein, inwiefern dies spezifische Infrastrukturen notwendig macht.
- Prof. Dr. Katharina Kinder-Kurlanda (Humanwissenschaft des Digitalen, Klagenfurt) betrachtet Social Media und andere digitale Materialien als Forschungsdaten für empirisch-kulturwissenschaftliche Fragestellungen.
- Prof. Dr. Estrid Sørensen (Cultural Psychology and Anthropology of Knowledge, Bochum) stellt die Methode der Data Walks vor und verbindet diese mit einer Problematisierung von Visualisierungen.
- Dr. des. Libuše Hannah Vepřek (Empirische Kulturwissenschaft, Tübingen) zeigt auf, wie Computercode ethnografisch bearbeitet werden kann und wie Mensch-Maschine-Relationen in soziotechnischen Assemblages beforscht werden können.

Organisation und Moderation: Prof. Dr. Lina Franken (Digital Humanities in den Kulturwissenschaften, Vechta).

Bibliographie

Amelang, Katrin, Bauer, Susanne. 2019. “Following the Algorithm. How Epidemiological Risk-Scores do Accountability.” *Social Studies of Science 49/*4, 476–502.

Amelang, Katrin, Martina Klausner, Estrid Sørensen und Till Straube. 2023. “Daten erfahren und situieren. Data Walking als explorative Methode ethnografischer Forschung.” In: *KA Notizen 85*. https://doi.org/10.21248/ka-notizen.85.20.

Beck, Stefan, Jörg Niewöhner und Estrid Sørensen (Hg.). 2012. *Science and Technology Studies. Eine sozialanthropologische Einführung.* Bielefeld.

Bischoff, Christine, Karoline Oehme-Jüngling, Walter Leimgruber (Hg.). 2014. *Methoden der Kulturanthropologie.* Bern.

Burghardt, Manuel. 2023. *Kritische Überlegungen zum Algorithmizitätsbegriff*. https://dhtheorien.hypotheses.org/1316.

Carlson, Rebecca, Ruth Dorothea Eggel, Lina Franken, Sarah Thanner und Libuše Hannah Vepřek. 2021. “Approaching Code as Process. Prototyping Ethnographic Methodologies.” In: *Kuckuck. Notizen zu Alltagskultur und Volkskunde*, 13–17.

Deutsche Gesellschaft für Empirische Kulturwissenschaft (DGEKW). 2023. *Positionspapier zur Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten.* https://dgekw.de/wp-content/uploads/2023/11/DGEKW_Positionspapier-FDM_2023.pdf.

Egger, Nils, Lina Franken, Dennis Möbus und Florian Schmid. 2023. “Oral History auf dem Weg zu Big Data: menschliche und maschinelle Annotation lebensgeschichtlicher Interviews im Vergleich.” In: *Digital Humanities im deutschsprachigen Raum (DHd) 2023. Book of Abstracts.* Trier. https://zenodo.org/record/7715317.

Fleischhack, Julia. 2019. “Ethnografisch (um)denken. Zu den Besonderheiten und Herausforderungen von Digitaler und Virtueller Ethnografie.” In: *Forschungsdesign 4.0. Datengenerierung und Wissenstransfer in interdisziplinärer Perspektive*. hg. von Jens Klingner und Merve Lühr. Dresden, 94–106.

Fortun, Kim, Mike Fortun, Erik Bigras, Tahereh Saheb, Brandon Costelloe-Kuehn, Jerome Crowder, Daniel Price und Alison Kenner. 2014. “Experimental Ethnography Online.” In: *Cultural Studies 28/4*, 632–642. DOI:10.1080/09502386.2014.888923.

Franken, Lina. 2022a. “Digitale Daten und Methoden als Erweiterung qualitativer Forschungsprozesse. Herausforderungen und Potenziale aus den Digital Humanities und Computational Social Sciences.”

In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 22 (2022)*. doi.org/10.17169/fqs-22.2.3818.

Franken, Lina. 2022b. "Erweiterungen der Digital Humanities durch kulturwissenschaftliche Perspektiven." In: *Digital Humanities im deutschsprachigen Raum (DHd) 2022. Book of Abstracts*. https://zenodo.org/record/6327985.

Franken, Lina. 2023. *Digitale Methoden für qualitative Forschung. Computationelle Daten und Verfahren*. Münster/New York.

Franken, Lina, Nils Egger, Luis Fischer, Katharina Lillich, Florian Schmid, Florian. 2023. "Nachnutzung von Forschungsdaten für qualitative Forschungen. Text Mining als Ansatz zur Exploration transkribierter Interviews". In: *KA Notizen* 85. https://doi.org/10.21248/ka-notizen.85.16.

Geiger, Jonathan. 2023. *„Algorithmizität" – Begriffsanalyse und Plädoyer*. https://dhtheorien.hypotheses.org/1400.

Geiger, Jonathan und Jasmin Pfeiffer. 2020. "Spielplätze der Theoriebildung in den Digital Humanities." In: *Digital Humanities im deutschsprachigen Raum (DHd) 2020. Book of Abstracts.* Paderborn, https://zenodo.org/record/3666690.

Heimerdinger, Timo, Markus Tauschek (Hg.). 2020. *Kulturtheoretisch argumentieren. Ein Arbeitsbuch*. Münster u.a.

Hess, Sabine, Johannes Moser, Maria Schwertl (Hg.). 2013. *Europäisch-ethnologisches Forschen. Neue Methoden und Konzepte*. Berlin.

Hinrichs, Peter, Martina Röthl, Manfred Seifert (Hg.). 2021. *Theoretische Reflexionen. Perspektiven der Europäischen Ethnologie*. Berlin.

Horst, Heather A. und Daniel Miller (Hg.). 2020. *Digital Anthropology*. London.

Kinder-Kurlanda, Katharina E. 2020. "Big Social Media Data als epistemologische Herausforderung für die Soziologie." In: Soziologie des Digitalen - Digitale Soziologie? hg. von Sabine Maasen und Jan-Hendrik Passoth. Baden-Baden, 109–133.

Klausner, Martina. 2022. "Eine „mehr-als-digitale Anthropologie". Ethnografien der Partizipation und öffentlichen Verwaltung". In: *Zeitschrift für Empirische Kulturwissenschaft* 118, 5–24.

Klausner, Martina und Dennis Eckhardt (Hg.). 2023. *Methoden in digitalen Feldern. Fallbeispiele zu ethnografischen Methoden in digitalen Feldern und methodologische Reflexionen* (KA-Notizen). Frankfurt a.M.

Koch, Gertraud. 2017. "Ethnografie digitaler Infrastrukturen." In: *Digitalisierung. Theorien und Konzepte für die empirische Kulturforschung*. hg. von Gertraud Koch. Konstanz/München, 107–126.

Koch, Gertraud und Lina Franken. 2020. "Filtern als digitales Verfahren in der wissenssoziologischen Diskursanalyse. Potentiale und Herausforderungen der Automatisierung im Kontext der Grounded Theory." In: *Soziale Medien. Interdisziplinäre Zugänge zur Onlinekommunikation*. hb. von Peter Klimczak, Christer Petersen und Samuel Breidenbach. Wiesbaden, 121–138.

Moeller, Katrin, Sibylle Söring, Sabine Imeri, Marina Lemaire, Nils Reichert. 2022. "Ethisch - transparent - offen - Die CARE-Prinzipien und ihre Implikationen für geisteswissenschaftliche FDM-Services." *Digital Humanities im deutschsprachigen Raum (DHd) 2022. Book of Abstracts.* Potsdam. https://doi.org/10.5281/zenodo.6328105.

Niewöhner, Jörg. 2015. "Infrastructures of Society, Anthropology of." In: International Encyclopedia of the Social & Behavioral Sciences. hg. von James D. Wright. Amsterdam u.a., 119–125. https://doi.org/10.1016/B978-0-08-097086-8.12201-9.

Pink, Sarah, Heather A. Horst, John Postill, Larissa Hjorth, Tania Lewis und Jo Tacchi. 2016. *Digital Ethnography. Principles and Practice*. Los Angeles, CA u.a.

Poirier, Lindsay, Kim Fortun, Brandon Costelloe-Kuehn und Mike Fortun. 2020. "Metadata, Digital Infrastructure, and the Data Ideologies of Cultural Anthropology." In: *Anthropological Data in the Digital Age*. hg. von Jerome W. Crowder, Mike Fortun, Rachel Besara und Lindsay Poirier. 209-237. https://doi.org/10.1007/978-3-030-24925-0_10

Svensson, Patrik. 2016. *Big Digital Humanities*. Michigan.

Vepřek, Libuše Hannah. 2023. "Ein Gefühl für die Daten entwickeln. Eine ethnografische Annäherung an große Textdaten am Beispiel digitaler Chats." In: *KA Notizen* 85. https://doi.org/10.21248/ka-notizen.85.12.

Vepřek, Libuše Hannah, Sarah Thanner und Lina Franken. 2023. "Computercode in seinen Dimensionen ethnografisch begegnen." In: *KA Notizen* 85. https://doi.org/10.21248/ka-notizen.85.13.

Still alive?! - Vom Umgang mit lebenden Systemen in den Digital Humanities

Helling, Patrick

patrick.helling@uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland
ORCID: 0000-0003-4043-165X

Rau, Felix
f.rau@uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland
ORCID: 0000-0003-4167-0601

Schildkamp, Philip
philip.schildkamp@uni-koeln.de
Cologne Center for eHumanities (CCeH), Universität zu Köln, Deutschland
ORCID: 0000-0003-0209-2837

Dieckmann, Lisa
lisa.dieckmann@uni-koeln.de
Universität zu Köln, NFDI4Culture, Deutschland
ORCID: 0000-0002-1708-7371

Puhl, Johanna
johanna.puhl@dlr.de
Deutschen Zentrum für Luft- und Raumfahrt (DLR), Köln, Deutschland

Henny-Krahmer, Ulrike
ulrike.henny-krahmer@uni-rostock.de
Universität Rostock, Deutschland
ORCID: 0000-0003-2852-065X

Einleitung

Neben (digitalen) Forschungsdaten, die es durch Maßnahmen des Forschungsdatenmanagements (FDM) im Sinne der FAIR-Prinzipien (Wilkinson et al. 2016) gilt möglichst auffindbar, zugänglich, interoperabel und nachnutzbar zu gestalten, stellen sog. *lebenden Systeme*[1] als zentrale Instrumente digitaler Ergebnissicherung und Träger wissenschaftlicher Erkenntnis einen gleichwertigen Output digitaler Forschungsvorhaben dar. Als Forschungsergebnisse, die den Erkenntnisprozess ermöglichen oder unterstützen, müssen solche Präsentationsplattformen und interaktive Visualisierungen, Recherche-Datenbanken, Kollaborationssysteme, Websites und dynamische Forschungsanwendungen ebenso wie Forschungsdaten durch Dritte nachgenutzt werden können.

Die Gewährleistung ihrer langfristigen Verfügbarkeit gehört daher gleichermaßen zu einer umfassenden Nachhaltigkeitsstrategie und ist nicht nur durch Datenzentren und technische Infrastrukturen sicherzustellen, deren Aufgabe es ist, Forschende bei der langfristigen Bereitstellung *lebender Systeme* organisatorisch und technisch zu unterstützen. Ebenso sind Drittmittelgeber in der Verantwortung, Entwicklungsvorhaben im Rahmen von Projektanträgen kritisch zu evaluieren. Und nicht zuletzt müssen auch Forschende über den gesamten Projektverlauf den nachhaltigen Betrieb zu entwickelnder, *lebender Systeme* berücksichtigen und planen.

Entsprechend stellt der Umgang mit *lebenden Systemen* eine zentrale Herausforderung im Forschungsdatenmanagement und darüber hinaus dar, bei der unterschiedliche Stakeholder zusammenarbeiten müssen.

Wenn nun bei der diesjährigen DHd-Jahreskonferenz mit dem Call for Papers unter dem Titel “DH Quo Vadis” angeregt wird darüber nachzudenken, wo die Digital Humanities mittlerweile stehen, so muss aus der Perspektive des Forschungsdatenmanagements nicht nur über die Nachhaltigkeit von digitalen Daten, sondern insbesondere auch von *lebenden Systemen* als zentrale Ergebnisse digitaler, geisteswissenschaftlicher Forschung reflektiert werden.

Der Umgang mit *lebenden Systemen* in der Praxis

Zum nachhaltigen Betrieb *lebender Systeme* existieren bereits unterschiedliche Strategien: So kann **(1)** die Bereitstellung ausreichender finanzieller und personeller Ressourcen zur Kuratierung und Betreuung *lebender Systeme* (Smithies et al. 2019) als eine solche gesehen werden, wobei sich hier bei einer wachsenden Zahl zu kuratierender *lebender Systeme* die Frage der Skalierbarkeit stellt. **(2)** Die Beschränkung von bei der Entwicklung *lebender Systeme* eingesetzten Technologie-Stacks (Arneil et al. 2019) zur Verringerung des Kurationsaufwands, ein weiterer Ansatz, steht dem grundsätzlichen Selbstverständnis vieler Forschenden entgegen, müsste vermutlich aus den jeweiligen Fachdisziplinen selbst heraus entstehen und in Form von de-facto Standards realisiert werden. Auch **(3)** die Virtualisierung oder Kapselung *lebender Systeme* (Smithies et al. 2019) stellt eine mögliche Strategie dar, bedarf allerdings geeigneter Infrastruktur und Kompetenzen in ausführenden Einrichtungen. Zuletzt **(4)** existiert noch die Vorgehensweise, *lebende Systeme* in einem spezifischen Zustand in einer möglichst statischen Form auf Infrastrukturen abzulegen (Arneil et al. 2019) und so den Kurationsaufwand möglichst gering zu halten, womit allerdings häufig Einschränkungen der Funktionalität einhergehen.

Trotz aller bestehender Strategien scheint ein breit aufgestelltes und standardisiertes Serviceangebot (sowie die notwendige Awareness) für den Umgang mit *lebenden Systemen* in den Digital Humanities allerdings noch nicht zu existieren, wie auch eine im Vorfeld dieses Panels durchgeführte quantitative, anonymisierte Kurzumfrage mit geisteswissenschaftlichen Datenzentren verdeutlicht.[2]

Umfrage mit geisteswissenschaftlichen Datenzentren zum Umgang mit *lebenden Systemen*

Insgesamt wurden 16 geisteswissenschaftliche Datenzentren im deutschsprachigen Raum dazu eingeladen, an einer Onlineumfrage mit elf Fragen zu Angeboten und Diensten zum Umgang mit *lebenden Systemen* sowie deren Nutzungsbedingungen teilzunehmen. Die Umfrage wurde über eine LimeSurvey-Instanz,[3] die durch die Universität zu Köln[4] bereitgestellt wird, realisiert und lief insgesamt zwölf Tage. Lediglich vier Vertreter*innen von geisteswissenschaftlichen Datenzentren haben an der Umfrage teilgenommen (N=4).

Drei von vier Teilnehmenden gaben an, dass sie *lebende Systeme* über eine Förderphase hinaus übernehmen und betreuen, i.d.R. jedoch nur dann, wenn sie selbst an den Forschungsprojekten beteiligt waren, sowie nur unter spezifischen Bedingungen. Diese drei Zentren vermerkten, dass sie *lebende Systeme* als bereitgestellte Software mit (möglichst) allen Funktionalitäten übernehmen, ein Zentrum gab zusätzlich auch an, *lebende Systeme* in statischer, also toter Form, zu übernehmen, wenngleich dies bedeutet, dass einige der lebenden Funktionalitäten nicht erhalten bleiben. Lediglich ein Datenzentrum verfügt dabei über einen standardisierten Workflow. In diesem Zusammenhang würde eine Webarchivierung *lebender Systeme* bereits bei der Projektplanung mitgedacht und schließlich am Ende einer Förderphase umgesetzt.

Bezüglich der spezifischen Bedingungen, unter denen *lebende Systeme* übernommen werden, gaben alle drei Zentren spezifische Technologie-Stacks an (jeweils eine Nennung):

- X-Technologien
- LAMP-Stack
- serverbasierte Systeme (kein Desktop)
- festgelegte Laufzeitumgebungen (Java, Python)

Darüber hinaus vermerkten zwei Zentren, dass die Software auf einer bestimmten, technischen Infrastruktur lauffähig sein muss (jeweils eine Nennung):

- Linux-Webserver
- Containerisiert (Docker)

Zuletzt machte mit der Aussage, ein *lebendes System* solange zu betreuen, bis Updates nicht mehr ohne zusätzliche Arbeit eingespielt werden können, lediglich ein Datenzentrum eine Aussage über den Zeitraum, für den es die Betreuung eines *lebenden Systems* garantiert.

Das Datenzentrum, das angab keine Betreuung *lebender Systeme* anzubieten notierte, dass der resultierende Arbeitsaufwand aus solchen Übernahmeprozessen nicht leistbar sei. Allerdings würde das Zentrum Forschungsdaten aus *lebenden Systemen* in eine eigene Infrastruktur übernehmen, über die diese dann in generischer Form langfristig zugänglich blieben.

Zusammenfassend lässt sich festhalten, dass es zwar bereits einige geisteswissenschaftliche Datenzentren gibt, die die Übernahme und Betreuung von *lebenden Systemen* unter bestimmten Bedingungen anbieten. Allerdings scheint es, dass die Zentren lediglich *lebende Systeme* aus Projekten übernehmen, an denen sie auch (von Beginn an) beteiligt sind. Darüber hinaus deutet der vergleichsweise geringe Rücklauf im Rahmen der Umfrage darauf hin, dass es keine flächendeckenden Angebote für die Betreuung von *lebenden Systemen* gibt. Auch scheinen die Bedingungen der existierenden Services noch recht unterschiedlich zu sein. Sie sind nachvollziehbarerweise von den jeweiligen technischen und strukturellen Gegebenheiten der einzelnen Zentren abhängig und lassen entsprechend keine de-facto Standards in der Betreuung von *lebenden Systemen*, sondern vielmehr einen gewissen Pragmatismus erkennen.

Ziele des Panels

Bestehende Strategien im Umgang mit *lebenden Systemen* stehen im Spannungsfeld zwischen Skalierbarkeit, Verlust von Funktionalitäten und effektiver Nachnutzbarkeit der Ressourcen. Es fehlt weiterhin ein Ansatz, der sowohl finanzielle und personelle Anforderungen niedrig hält, gleichzeitig das Bestehen substanzieller Funktionalitäten pauschal gewährleisten kann und alle Stakeholder – Forschende, Datenzentren und Drittmittelgeber – zusammenbringt. Darüber hinaus bleibt abzuwarten, ob national oder international ausgerichtete Infrastrukturen wie bspw. die Nationale Forschungsdateninfrastruktur (NFDI)[5] oder die European Open Science Cloud (EOSC)[6] entscheidende Dienste für die langfristige Verfügbarkeit von *lebenden Systemen* community-weit bereitstellen können.

Unser Panel soll die Möglichkeit geben einerseits das Spannungsfeld zwischen Skalierbarkeit und Funktionalitätsbewahrung beim Umgang mit *lebenden Systemen* aus verschiedenen Stakeholder-Perspektiven zu diskutieren, andererseits soll es ein Forum sein, in dem gemeinsam mit der Community über Möglichkeiten der Orchestrierung der verschiedenen Stakeholder für eine nachhaltigere Entwicklung von und einen langfristigen Umgang mit *lebenden Systemen* diskutiert wird. Konkret sollen folgende Fragen im Panel diskutiert werden:

- Welche Ansätze (und Bedarfe) gibt es, um *lebende Systeme* in den Digital Humanities nachhaltig zu realisieren und langfristig bereitzustellen?
- (Wie) können wir gemeinsame Standards und Best Practices sowohl in der Entwicklung als auch im Dauerbetrieb *lebender Systeme* finden und etablieren?
- Welche Maßnahmen braucht es, um die Herausforderungen im Umgang mit *lebenden Systemen* durch alle Stakeholder – Forschende, Datenzentren und Drittmittelgeber – gemeinsam zu adressieren?

Panelist*innen

Perspektive des Datenzentrums

Philip Schildkamp war von 2018 bis 2021 Teil des DFG-LIS-Forschungsprojekts *SustainLife – Erhalt lebender, digitaler Systeme für die Geisteswissenschaften*[7] am Data Center for the Humanities (DCH)[8] und arbeitet seitdem im Bereich der IT-Infrastruktur des Cologne Center for eHumanities (CCeH).[9] Als Systemadministrator mit einschlägigen Erfahrungen hinsichtlich des Erhalts *lebender Systeme* einerseits und dem theoretischen Hintergrundwissen um diverse Nachhaltigkeitsstrategien andererseits, bietet seine Perspektive einen pragmatischen Einblick in das Tagesgeschäft eines Datenzentrums, was den Umgang mit Forschungsapplikationen, deren Planung, Entwicklung und langfristigen Betrieb anbelangt.

Perspektive der NFDI

Lisa Dieckmann ist Geschäftsführerin des prometheus-Bildarchivs[10] an der Universität zu Köln und Co-Spokesperson von NFDI4Culture[11] für die Task Area "Research Tools und Data Services",[12] in der es u.a. auch um die nachhaltige Softwareentwicklung geht, bei welcher auch der Erhalt von *lebenden Systemen* einen entscheidenden Aspekt darstellt. Sie wird die im Kontext von NFDI diskutierten Strategien für die langfristige Zurverfügungstellung von *lebenden Systemen* und ihre langjährige Erfahrung in der Verstetigung von Forschungsprojekten und Services in die Diskussion einbringen.

Perspektive der Drittmittelgeber

Johanna Puhl arbeitet beim Projektträger im Deutschen Zentrum für Luft- und Raumfahrt[13] und betreut dort im Auftrag des BMBF[14] Projekte aus dem Bereich Digitalisierung von Kulturerbe, Projekte aus dem sozialwissenschaftlichen Umfeld sowie einige geistes- und sozialwissenschaftliche Infrastrukturen auf europäischer Ebene. Vorher arbeitete sie sieben Jahre als wissenschaftliche Mitarbeiterin am Lehrstuhl Historisch-kulturwissenschaftliche Informationsverarbeitung (heute Institut für Digital Humanities)[15] an der Universität zu Köln in Projekten, die u.a. die Themen digitale Langzeitarchivierung und wissenschaftliche Infrastrukturen beleuchteten. In vielen der vom BMBF geförderten Projekten werden *lebende Systeme* entwickelt, deren Nachhaltigkeit zuweilen Thema ist.

Perspektive der Forschung

Ulrike Henny-Krahmer ist seit 2021 Akademie-Juniorprofessorin für Digital Humanities an der Universität Rostock.[16] Sie hat an der Universität Würzburg[17] promoviert und zuvor vier Jahre am Cologne Center for eHumanities (CCeH) in digitalen Editions- und Archivprojekten mitgearbeitet. Aktuell ist sie PI bzw. Co-PI in drei Forschungsprojekten (Pessoa digital,[18] DEMel,[19] CANSpiN[20]) aus dem Bereich der Philologien und arbeitet im Akademievorhaben Uwe Johnson-Werkausgabe[21] mit. In drei der Vorhaben sind *lebende Systeme* involviert, im vierten Forschungssoftware zur Textanalyse. Aus ihrer Perspektive ist neben technischen Faktoren vor allem auch der organisatorische Aspekt für den Umgang mit und Erhalt von *lebenden Systemen* entscheidend.

Die Panelist*innen haben jeweils 10 Minuten Zeit für ein Eröffnungsstatement, in dem sie sich zu den im Panel adressierten Fragen positionieren. Im Anschluss wird es eine 20 minütige, moderierte Diskussionsrunde zwischen den Panelist*innen geben. In den letzten 30 Minuten des Panels wird die Diskussion für das gesamte Publikum geöffnet.

Fußnoten

1. Im Vergleich zu Forschungsdaten sind *lebende Systeme* dynamische bzw. regelmäßig zu kuratierende Umgebungen, Zugangs- oder Repräsentationsschichten zu Forschungsdaten. Ihre Funktion ist i.d.R. die Zugänglichkeit und Nachnutzung von Forschungsdaten in einer spezifischen Form zu ermöglichen. Im Gegensatz zu *lebenden Systemen* können auch *tote Systeme* solche Funktionen erfüllen. Allerdings bedarf es zu ihrem Betrieb / zu ihrer Bereitstellung keine regelmäßige Kuration und Betreuung, da sie bspw. statisch sind und es keine Abhängigkeiten von sich verändernden Technologie-Stacks gibt.
2. https://doi.org/10.5281/zenodo.8160752 (letzter Zugriff: 30. November 2023).
3. https://rrzk.uni-koeln.de/internetzugang-web/bausteine-fuer-webseiten/online-umfragen/limesurvey (letzter Zugriff: 17. Juli 2023).
4. https://www.uni-koeln.de/ (letzter Zugriff: 17. Juli 2023).
5. https://www.nfdi.de (letzter Zugriff: 17. Juli 2023).
6. https://eosc-portal.eu/about/eosc (letzter Zugriff: 30. November 2023).
7. https://dch.phil-fak.uni-koeln.de/forschung/sustainlife (letzter Zugriff: 17. Juli 2023).
8. https://dch.phil-fak.uni-koeln.de/ (letzter Zugriff: 17. Juli 2023).
9. https://cceh.uni-koeln.de/ (letzter Zugriff: 17. Juli 2023).
10. https://prometheus-bildarchiv.de/de/ (letzter Zugriff: 17. Juli 2023).
11. https://nfdi4culture.de/ (letzter Zugriff: 17. Juli 2023).
12. https://nfdi4culture.de/about-us/task-areas/task-area-3.html (letzter Zugriff: 17. Juli 2023).
13. https://www.dlr.de/ (letzter Zugriff: 17. Juli 2023).
14. https://www.bmbf.de/ (letzter Zugriff: 17. Juli 2023).
15. https://dh.phil-fak.uni-koeln.de/ (letzter Zugriff: 17. Juli 2023).
16. https://www.uni-rostock.de/ (letzter Zugriff: 17. Juli 2023).

17. https://www.uni-wuerzburg.de/ (letzter Zugriff: 17. Juli 2023).
18. https://cceh.uni-koeln.de/portfolio/pessoa-digital/ (letzter Zugriff: 17. Juli 2023).
19. https://www.romanistik.uni-rostock.de/forschung/sprachwissenschaft/demel/ (letzter Zugriff: 17. Juli 2023).
20. https://www.canspin.uni-rostock.de/ (letzter Zugriff: 17. Juli 2023).
21. http://www.uwe-johnson-werkausgabe.de/ (letzter Zugriff: 17. Juli 2023).

Bibliographie

Arneil, Stewart, Martin Holmes, und Greg Newton. 2019. „Project Endings: Early Impressions From Our Recent Survey On Project Longevity In DH“. DataverseNL. https://doi.org/10.34894/SIKOBN .

Smithies, James, Carina Westling, Anna-Maria Sichani, Pam Mellen und Arianna Ciula. 2019. “Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King’s Digital Lab.” digital humanities quarterly. http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html (letzter Zugriff: 17. Juli 2023).

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 (1): 160018. https://doi.org/10.1038/sdata.2016.18 .

The Epistemological Value of the Computational Turn in Scholarly Editing

Cugliana, Elisa
elisa.cugliana@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-6460-2954

Kuczera, Andreas
andreas.kuczera@mni.thm.de
Technische Hochschule Mittelhessen, Gießen
Deutschland
ORCID: 0000-0003-1020-507X

Grüntgens, Max
gruentgens@fotomarburg.de
Deutsches Dokumentationszentrum für Kunstgeschichte, Bildarchiv Foto Marburg, Deutschland

Ward, Aengus
a.m.m.ward@bham.ac.uk
University of Birmingham, Großbritannien
ORCID: 0000-0001-9526-0718

van Zundert, Joris J.
joris.van.zundert@huygens.knaw.nl
Huygens Institute, Amsterdam, Niederlande
ORCID: 0000-0003-3862-7602

Andrews, Tara
tara.andrews@univie.ac.at
Universität Wien, Österreich
ORCID: 0000-0001-6930-3470

Dekker, Ronald Haentjens
ronald.dekker@di.huc.knaw.nl
DHLab / Huygens Institute, Amsterdam, Niederlande
ORCID: 0000-0001-6737-7986

Summary

In the last decade a decidedly computational turn in scholarly editing is noticeable. Texts are not just encoded anymore. To store and execute scholarly decisions increasingly also program code is applied, which thereby becomes a computational expression of scholarly theory. Theorists and practitioners of such computational approaches have not just put forward scalability and replicability as motivations for computational methods, but also the promise of increased analytical power and on the horizon a changed epistemology. This panel invites some of the most visible proponents of these approaches to reflect on these promises and especially to try to deepen our understanding of the epistemological value that is tied to applying code in the domain of scholarly textual editing.

Introduction

How is the computational turn in scholarly editing fundamentally altering our understanding of the texts that we edit? There is no question that the computational turn has happened over the last decades (cf. Cugliana and Van Zundert 2022; Ward 2022; Van Hulle 2021; Andrews 2019; Buschenhenke 2019; Haentjens, Bleeker and Buitendijk 2018; Ries 2017; Barabucci and Fischer 2017; Hadersbeck et al. 2015; Oldman, Doerr, and Gradmann 2015; Rosenthal

2015; Buzzetti 2002).This panel poses, and tries to address, the question of the long-term implications of the nature of what we are now doing as digital textual scholars. Digital approaches from the recent decades in scholarly editing have produced solid groundwork in theoretical principles and concepts, practical methods and standards, as well as functional tools and processes. Although there is something of a broad consensus emerging in these areas, many of the discussions about the future of digital and computational[1] critical editing focus on the development of specific tools and methods stemming directly from this consensus and not on the long-term implications for the nature of what we do as textual scholars.

THINK, as a research group, looks at the currently dispersed and technically oriented discourse from a wider angle and an open stance. THINK seeks to highlight without prejudice the extent of the possibilities of computational textual analysis that are currently emerging, and by foregrounding to what degree these may, or already do, represent an epistemological shift in the nature of the textual knowledge and understanding we seek by scholarly editing. The discourse initiated by THINK thus questions whether the computationally created editions of the future will merely exist as a quantitatively upscaled means of the scholarly practice that exists today. Or if such computational tools will rather allow us to analyse and represent the knowledge we edit as well as the knowledge we thus produce in some dynamic reciprocal fashion. Or even if their application will constitute a fundamental reorientation of the very object of study itself, making it possible to engage with the object of study in a methodically controlled dialectic that considers both the analog and digital mode of being of the object. The panel we propose is structured in three sub-themes that serve to explore these meta-questions pertinent to the future of computational scholarly editing.

Contents of the Panel

New frontiers: Where does the current methodological frontier of digital and computational textual scholarship lie?

In the past, textual scholars have paid much attention to limits, boundaries, and differences between the book and the digital text. In many respects the frontier of digital philological methodology has moved past these questions of materiality (Kirschenbaum 2008) and feasibility (e.g. Barabucci and Fischer 2017:47-49), not implying that these questions are unimportant, but rather pointing out that the leading edge of methodological thinking seems to have progressed into an area of questioning how scholarly process, observation, interpretation, and even reasoning already is or might be calculated upon algorithmically (Cugliana and Van Zundert 2022). The assumption of these new methodologies seems to be that they also improve in some ways not just our analytics, but also our epistemological thinking, and even our knowledge. This bears new emerging challenges concerning the ways this new kind of thinking should be transmitted to its users and audiences, so to be evaluated. As a matter of fact, if the nature of our work is changing, then there is a need to deepen our study of the language of interfaces (cf. Andrews and Van Zundert 2018) so as to find the best strategies to communicate not only our results but also the very nature of our processes.

We ask the panellists to reflect on some questions in this regard. What is already or may already be calculated algorithmically in the field of philology, and where is the current “computational horizon” for textual scholarship? What do we think computation rather than digitization delivers us? In what ways does this advance our textual scholarship thinking? How do we communicate computed results to various textual scholarship audiences? Do we need new kinds of visualisations and knowledge structures that go well beyond the book metaphor?

The complexities of operationalizing scholarly theory and perspectives in data and process

In related fields such as computational linguistics and computational literary studies a discussion is going on concerning the need for theoretical concepts to be effectively connected with observed data in an unambiguous way (Pichler and Reiter 2022, Bode 2023). The same need exists in scholarly editing. In order to bridge the gap between the theory and its actual instantiations it is necessary to express the former as a discrete set of operations to be performed on the data. What are the claims in this respect of current advanced computational or digital textual scholarship? On a practical level this operationalization raises questions on how to deal with the intricacies and complexities of expressing scholarly observation: what granularity of computational description is needed and feasible, and how does a scholar practically engage with such granular complex digital data and information? The levels of indirection required to realise this approach might seem overly complex. How can we augment the granularity of data, its addressability, and tractability in order to break down our theories and create explicit connections between our observations and textual phenomena?

Panellists could reflect on the following questions. How do we make computational and digital data, information, and processes of the highest granularity tractable not just by machines but also by the human scholar? Does computational textual scholarship lead to a hairball problem squared? Are graph based structures of text and annotations a more effective means to express interpretation and annotation of text, and even scholarly reasoning about texts? Are semantic web solutions such as RDF-LOD sufficient? Are current leading-edge graph models and solutions able to embrace an open world assumption in a dialectical fashion and practically relate it to standing textual scholarly practice? Can we augment open world knowledge graphs with a techno-

logy supporting logical reasoning to push ahead computational hermeneutics?

The object of research: Are we moving into virtualized textual scholarship?

THINK emerges from a group of scholars who, for the most part, were concerned with the edition of mediaeval manuscript works. The theory and practice of such editing was always necessarily a negotiation between the material and unique instances of manuscript text, and the desire to systematise knowledge of "text" (cf., for instance, Cerquiglini 1999). Editing in print tended to prioritise the latter, in the name of eradicating error or other forms of variance. Digitization and computationalization, on the other hand, exert a strong reifying force on textual scholarship. Involved and intimate working with digital and computational methods tends to substitute the materiality of manuscript sources with the immediacy of digital representation and datafication. We can therefore state that our work is more and more centred on virtualized objects of study, which affects the very nature of our own workflows and methods (cf. Ward 2022). As a matter of fact, digital textual scholarship to an extent virtualizes different classes of knowledge about the work in the same digital space - creative links which might previously have been (only) conceptual and cognitive can now be expressed editorially. Additionally, the statements, annotations and processes of the researcher and editor can now be, must be, explicit. In this dispensation, what then is the epistemological link between the sum of assertions about a (textual) object, and that object itself, however defined? Can it be the case that the very object of study, in this case the work that is presumed to exist in the material object, is itself altered? What does this mean with regard to our objects of research, and how does this impact our epistemological processes?

Thus we ask our panellists: Is the nature of the digital object of research different with respect to the analog object? In what sense? How does our knowledge about the objects of study change when we recognise and start making assertions on digital objects? If we move into an age of simulated textual scholarship, what gains do we see for philological process and knowledge inference? What are the potential gains and pitfalls of the fact that virtualization and digital objects provide the possibility to make observations and processes of philology completely explicit and provenance completely transparent?

Invited Panellists

Prof.in Dr. Tara Andrews Universität Wien, Austria (0000-0001-6930-3470)

Ronald Haentjens Dekker DHLab / Huygens Institute, Amsterdam (0000-0001-6737-7986)

Max Grüntgens Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg, Germany

Prof. Dr. Aengus Ward University of Birmingham, United Kingdom (0000-0001-9526-0718)

Dr. Joris van Zundert Huygens Institute, Amsterdam (0000-0003-3862-7602)

Moderator

Jun.-Prof.in Dr. Elisa Cugliana Universität zu Köln, University of Cologne (0000-0002-6460-2954)

Fußnoten

1. In our work (cf. Cugliana and Van Zundert 2022) we understand the difference between digital and computational approach in that the first is primarily interested in digital encoding and representation, while the latter refers to the application of (bespoke) computer code as a tool to operationalize the very activity and processes of scholarly editing.

Bibliographie

Andrews, Tara L. 2019. "Critical Edition as Process: A Digital Model." In *16th Annual Conference of the European Society for Textual Scholarship - Book of Abstracts*, 5. Málaga: Department of English, French and German Philology of the University of Málaga. https://drive.google.com/file/d/1_Tb6xxA94lOcProU09aIgF_eHG8h42zj/view.

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, and Gerlinde Schneider, 3–33. Schriften Des Instituts Für Dokumentologie Und Editorik 12. Norderstedt: Books on Demand.

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. https://www.sidestone.com/books/advances-in-digital-scholarly-editing.

Bode, Katherine. 2023. "What's the Matter with Computational Literary Studies?" *Critical Inquiry* 49 (4): 507–29. https://doi.org/10.1086/724943.

Buschenhenke, Floor. 2019. "Track Changes: Demarcating Phases in Digital Writing Processes." In *16th Annual Conference of the European Society for Textual Scholarship - Book of Abstracts*, 7. Málaga: Department of English, French and German Philology

of the University of Málaga. https://drive.google.com/file/d/1_Tb6xxA94lOcProU09aIgF_eHG8h42zj/view.

Buzzetti, D. 2002. “Digital Representation and the Text Model.” *New Literary History* 33 (1): 61–88.

Cerquiglini, Bernard. 1999. In *Praise of the Variant: A Critical History of Philology*. Baltimore: The Johns Hopkins University Press.

Cugliana, Elisa, and Joris Van Zundert. 2022. “A Computational Turn in Digital Philology.” *Filologia Germanica / Germanic Philology* 14 (1): 43-71. http://aifg.it/rivista-14/.

Hadersbeck, Max, Alois Pichler, Florian Fink, Daniel Bruder, and Ina Arends. 2015. “Wittgensteins Nachlass: Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind.” In *DHd2015: Von Daten zu Erkenntnissen. Book of Abstracts*, 187--190. Graz: ZIM/ACDH. http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege.

Haentjens Dekker, Ronald, Elli Bleeker, Bram Buitendijk, Astrid Kulsdom, and David J. Birnbaum. 2018. “TAGML: A Markup Language of Many Dimensions.” In *Proceedings of Balisage: The Markup Conference 2018*. Balisage Series on Markup Technologies, Vol. 21. Washington D.C. https://doi.org/10.4242/BalisageVol21.HaentjensDekker01.

Kirschenbaum, Matthew. 2008. *Mechanisms: New Media and the Forensic Imagination*. Cambridge (Massachusetts), London (England): The MIT Press.

Oldman, Dominic, Martin Doerr, and Stefan Gradmann. 2015. “Zen and the Art of Linked Data.” In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 251–73. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118680605.ch18.

Pichler, Axel, and Nils Reiter. 2022. “From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities.” *Journal of Cultural Analytics* 7 (4). https://doi.org/10.22148/001c.57195.

Ries, Thorsten. 2017. “The Rationale of the Born-Digital Dossier Génétique: Digital Forensics and the Writing Process: With Examples from the Thomas Kling Archive.” *Digital Scholarship in the Humanities* 33 (2): 391–424. https://doi.org/10.1093/llc/fqx049.

Rosenthal, David S.H. 2015. “Emulation & Virtualization as Preservation Strategies.” Andrew W. Mellon Foundation. https://web.stanford.edu/group/lockss/resources/2015-10_Emulation_&_Virtualization_as_Preservation_Strategies.pdf.

Van Hulle, Dirk. 2021. “Dynamic Facsimiles: Note on the Transcription of Born-Digital Works for Genetic Criticism.” *Variants* 15–16: 231--241. https://doi.org/10.4000/variants.1450.

Ward, Aengus. 2022. “Of Digital Surrogates and Immaterial Objects: The (Digital) Future of the Iberian Manuscript in Textual Editing.” *Journal of Medieval Iberian Studies* 14 (1): 41–54. https://doi.org/10.1080/17546559.2021.2016887.

Zwischen Mehrsprachigkeit und Ressourcenlücke: Quo Vadis “Kleine Fächer” in den deutschsprachigen Digital Humanities?

Grallert, Till
till.grallert@fu-berlin.de
Humboldt-Universität Berlin, Deutschland
ORCID: 0000-0002-5739-8094

Mende, Jana-Katharina
jana-katharina.mende@germanistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0000-0001-7433-4351

Müller-Laackman, Jonas
jonas.mueller-laackman@sub.uni-hamburg.de
Staats- und Universitätsbibliothek Hamburg, Deutschland
ORCID: 0000-0003-2279-6751

Wagner, Cosima
cosima.wagner@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0003-4957-3478

Burr, Elisabeth
elisabeth.burr@uni-leipzig.de
Universität Leipzig, Deutschland
ORCID: 0000-0002-3445-150X

Elwert, Frederik
frederik.elwert@rub.de
Ruhr-Universität Bochum, Deutschland
ORCID: 0000-0001-9149-9377

Kraneiß, Natalie
n.kraneiss@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0003-3584-285X

Padieu, Damir
s2dapadi@uni-trier.de
Universität Trier, Deutschland

Tikhonov, Aleksej
aleksej.tikhonov@slavistik.uni-freiburg.de
Humboldt-Universität Berlin, Deutschland
ORCID: 00000-0002-0772-3397

Vertan, Cristina
cristina.vertan@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Sogenannte “Kleine Fächer” in den Geistes-, Sozial- und Kulturwissenschaften mit regionalen Schwerpunkten – von der Arabistik bis zur Ukrainistik – sind in besonderem Maße darauf angewiesen, Ressourcen der Digital Humanities für und in verschiedenen Sprachen zu verwenden, sei es als multilinguale Korpora, Metadaten in nichtlateinischen Schriften oder als Regelwerke für die Codierung wie z.B. TEI. Innerhalb der deutschen DHd-Community sind “Kleine Fächer”, aber auch die sogenannten “Regionalstudien”, wie auch sonst in der Wissenschaftsstruktur, noch unterrepräsentiert.

Infolgedessen gibt es einen "ressourcedness gap" (Nicholas, Bhatia 2023) bei der digitalen Präsenz dieser Disziplinen, sowie in Bezug auf Daten, Werkzeuge und Lösungen, die für die Forschung und Lehre im Bereich der DH verwendet / entwickelt werden.

Andererseits wird im Rahmen von Initiativen wie der NFDI in Deutschland von den "Kleinen Fächern" erwartet, dass sie ihre "Bedürfnisse" und “User Stories” einbringen:

> “Kleine Fächer stellten aufgrund ihrer spezifischen Gegenstände besondere Herausforderungen an die Digitalisierung. Gleichzeitig profitierten sie von der Digitalisierung in besonderer Weise. Hier wurde insbesondere die Schaffung digitaler Strukturen in der Lehre kleiner Fächer als eine große Chance gesehen. Ihre spezifischen Bedarfe sollten kleine Fächer insgesamt stärker in den Aushandlungsprozess und in politische Diskurse einbringen.” (Arbeitsstelle Kleine Fächer 2020)

Das Panel greift diese Herausforderung auf und will bezüglich der CfP-Frage nach der Entwicklung der Digital Humanities im DACH-Raum theoretische, methodische und technische Zukunftsperspektiven für eine mehrsprachige DH diskutieren und Umsetzungsstrategien vorschlagen. Dabei stehen nicht nur die bestehenden Infrastrukturen auf dem Prüfstand für die nachhaltige Einbindung “Kleiner Fächer”, es soll weiterhin aus der Perspektive dieser Disziplinen diskutiert werden, welche Bedarfe noch unzureichend oder gar nicht erfüllt sind oder werden können.

Zentral ist weiterhin die Frage, wie “Kleine Fächer” sich vernetzen können, um so gemeinsam als Akteur:in auftreten zu können. Da die betroffenen Disziplinen für sich genommen selten über eine kritische Masse an Personal, Infrastruktur und Geldern verfügen (“ressourcedness gap”), um nachhaltig und über die eigenen Fachgrenzen hinaus Änderungen zu erwirken, ist ein gemeinsames Auftreten auch im Sinne des interdisziplinären Charakters der DH notwendig. Das geschieht im Einzelnen dort, wo fachliche Zusammenarbeit auch schon vor der digitalen Wende stattgefunden hat. Für eine nachhaltige Besserung der Situation aller betroffenen Fächer reicht das aber nicht aus, da Fachinitiativen sich selten über diese Fachzusammenarbeit hinaus vernetzen (siehe Grallert et al 2023).

Abschließend muss die Berücksichtigung von Mehrsprachigkeit in der Forschungsförderung und die Ent- und Weiterentwicklung von neuen Wissenschaftsinfrastukturen thematisiert werden. Bislang gibt es noch keine Vorgaben oder Guidelines, die dafür Sorge tragen, dass bei der Entwicklung neuer Systeme oder der Aufarbeitung bestehender Infrastrukturen Mehrsprachigkeit als inhärent und jede Fachrichtung einer globalisierten Wissenschaft betreffende Notwendigkeit mitgedacht wird.

Leitfragen

Die folgenden Punkte stehen im Zentrum der Input-Referate und sollen in der Diskussion im Panel mit den Teilnehmer:innen, welche die Perspektiven verschiedener Statusgruppen von Masterstudierenden bis Professor:innen repräsentieren, vertieft werden:

- Welche Herausforderungen zeigen sich in konkreten Forschungsprojekten, die auf fehlende Ressourcen, einsprachige Standards zurückzuführen sind?
- Welche Lösungen gibt es dafür?
- Was wäre eine mehrsprachige best practices für Forschungsprojekte in Kleinen Fächern?
- Welche Forderungen gibt es aus den Kleinen Fächern an die DH-Community, um ein Gleichgewicht herzustellen?
- Welche Infrastruktur braucht eine mehrsprachige DH?
- Wie sollte ein “Pflichtenheft” für “multilingually enhanced knowledge infrastructures” (Spence 2021) aussehen, das multilinguale und -scripturale DH stärkt?

Sprachen

In Anerkennung der Vielsprachigkeit der DH im deutschsprachigen Raum und der Inklusion möglichst vieler Kolleg:innen sind die Panelsprachen Deutsch und Englisch.

Statements der Panelist:innen

Der lange Weg zur Mehrsprachigkeit: Humanities Computing / Digital Humanities Forschung, internationale Konferenzen und Fachverbände

Prof. Dr. Elisabeth Burr, Universität Leipzig

Mehrsprachigkeit war nicht von Anfang an ein zentrales Thema im Humanities Computing / in den Digital Humanities. Das lag zum einen sicher am ASCII Code und an fehlenden Ressourcen, zum anderen aber auch an der Entstehung des Humanities Computing / Computing in the Humanities im angelsächsischen und US-amerikanischen Raum und damit verbunden der Konzentration auf das Englische.

In meinem kurzen Beitrag möchte ich einen Blick zurück wagen und anhand von u. a. (veröffentlichten) Konferenzbeiträgen (cf. Weingart et al. 2020), den CLiP Seminars (cf. Fiormonte / Usher 2001, Fiormonte 2003, Nicolás / Moneglia 2005), dem ADHO Standing Committee on Multilingualism & Multi-culturalism (cf. Burr 2006) und Global Outlook::Digital Humanities (GO::DH) das „Eindringen" des Themas „Mehrsprachigkeit" in die DH nachzeichnen. Dabei werden Erfolge und Misserfolge eine Rolle spielen. Zudem soll nach dem Verhältnis von institutioneller Ebene und Bewusstwerdung der Zentralität dieses Themas sowie nach den Wissenschaftssprachen gefragt werden.

Multilingual Challenges from a Religious Studies perspective

Prof. Dr. Frederik Elwert, Ruhr Universität Bonn

From the perspective of Religious Studies, I see three challenges with regard to multilingual DH. First, in addition to a resource gap regarding non-Latin scripts, we encounter an infrastructural gap in working with the *content* of non-Western material. Controlled vocabularies for art history are heavily biased towards Western and Christian art. Similarly, there are still no good vocabularies for classifying data with regards to religious traditions.

Second, solving the problem of working with multilingual historical material in Western institutions is not necessarily the same as including non-Western (or even non-English) research traditions.

Finally, national initiatives such as the NFDI are an important step, but they also run the risk of provincialisation, which adds to these challenges. For example, the use of GND for cataloging improves interoperability, but mainly within *German* institutions. This is a particular challenge for small disciplines that don't have the critical mass in a national context and rely on international collaboration.

Multilingual requirements for software and infrastructure

Dr. Cristina Vertan, Berlin Brandenburgische Akademie der Wissenschaften

My contribution will address the challenges of digital textual scholarship in "small research fields" with a particular focus on the investigation of historical languages. This is a field lacking dramatically of ressources. Due to the specificity of such languages general digital solutions are problematic and involve in many cases a flat processing. Many researchers in area studies DH lack awareness about the restrictions already developed tools impose. These tools are often advertised as "easy to be adapted", "work for any language". However, they were only tested for English or the "rich resourced languages" incl. Chinese and Modern Arabic. As a consequence, research projects relying on "not yet multilingually adapted" tools often have to drop or adjust their research goals.

The situation can be regarded similar to the 2000s when many EU Languages were very low resourced . Maybe the actions which were taken for these languages could be followed also for strengthening software and infrastructures for "Kleine Fächer" DH.

Finally, my contribution will also tackle the impact on "Kleine Fächer" by new developments in AI with methods relying especially on large amounts of training data.

Keine 'Under-resourced languages' und 'Kleine Fächer': Perspektiven der Arabistik und Islamwissenschaft

Natalie Kraneiß, M.A. Arabistik, Islamwissenschaft, Universität Münster

Arabistik und Islamwissenschaft stehen als "Kleine Fächer" vor spezifischen inhaltlichen, multidisziplinären und sprachlichen Herausforderungen. Während ein institutionalisierter Austausch mit anderen "Kleinen Fächern" unerlässlich ist, können fachspezifische Probleme nicht allein durch Zusammenschlüsse und Konsortien gelöst werden. Stattdessen müssen sie innerhalb des Fachs in Zusammenarbeit mit verwandten Disziplinen wie Turkologie, Iranistik und Semitistik diskutiert werden. Derzeit fehlt im DACH-Raum ein Raum für diese wichtigen Diskussionen.

Der Zugang zu den Digital Humanities gestaltet sich gerade in den „Kleinen Fächern" als schwierig, da er zu großen Teilen von persönlichen Kontakten und Engagement abhängt (z. B. in der Lehre oder durch kollegiale Unterstützung). Die Verwendung von Labels wie „Kleine Fächer" oder „under-resourced languages" erschwert zusätzlich die Sichtbarkeit und Auffindbarkeit von Ressourcen zu digitalen Methoden in den entsprechenden Zielsprachen, insbesondere für Anfänger, was eine Zusammenstellung entsprechender Ressourcen nötig macht. Viele in den „großen Fächern" erprobte Tools und Methoden sind nämlich nicht ohne Probleme auf fachspezifisches Material anzuwenden, u.a. wegen mangelnder Unterstützung der Kernsprachen. Daher bedarf es institutionalisierter, universitätsübergreifender und fachgebundener Kooperationen, die allen Interessierten den Zugang ermöglichen.

KI-basierte Transliterationsmodelle: Herausforderungen und Zukunftsperspektiven am Beispiel des Ukrainischen und weiterer Sprachen

Dr. Aleksej Tikhonov, Universität Zürich, Slavistik

Die Ukraine ist ein mehr- und vielsprachiger Kulturraum *par excellence*. Im Rahmen des MultiHTR-Projekts (Multilingual Handwritten Text Recognition) haben wir ein auf maschinellem Lernen basierendes Transliterationsmodell für das handgeschriebene Ukrainische des 19.-20. Jahrhunderts veröffentlicht. MultiHTR widmet sich dabei neben dem Ukrainischen auch anderen Sprachen und Schriften. Neben dem Umgang mit Ressourcenlücken und schwer zugänglichen Textdaten ist Mehrsprachig- und -schriftlichkeit eine zentrale Herausforderung unseres Projekts., So zielen wir beispielsweise auf die Erkennung nicht-lateinischer Alphabete (Serbisch, Osmanisch, Jiddisch), unterschiedlicher Schreibrichtungen (Osmanisch, Jiddisch) und komplexer Schreibsysteme (deutsche Stenografie). Dies erfordert innovative und skalierbare Lösungen. Anhand ausgewählter Beispiele werden sowohl die Herausforderungen als auch die Besonderheiten einzelner Sprachen/Schriften mit den entsprechenden Lösungsansätzen skizziert.

Parallelentwicklungen zwischen deutscher und taiwanesischer DH

Damir Padieu, Sinologie, DH, Trier Center for Digital Humanities

Für die Digitalisierung und Bearbeitung chinesischsprachiger Dokumente sind die Standard-Tools der westlichen DH oft unzureichend oder unpassend. Die chinesischsprachige DH-Community hat deshalb eigene Werkzeuge entwickelt; ein Beispiel dafür ist das taiwanesische Markup-Format DocuXML, welches eine Alternative zum TEI-Format darstellt und in Taiwan standardmäßig benutzt wird. Hier kann weniger von einer Ressourcenlücke die Rede sein als von einer Parallelentwicklung. Statt internationaler Zusammenarbeit haben sich voneinander isolierte Systeme entwickelt. Eine Herausforderung besteht darin, über Sprachgrenzen hinweg bessere Kooperationsstrukturen aufzubauen, um Ressourcen zu teilen, Interoperabilität zu fördern und den Mehraufwand der Parallelentwicklung zu vermeiden.

Bibliographie

ADHO Special Interest Group. *Global Outlook::Digital Humanities(GO::DH).* http://www.globaloutlookdh.org (zugegriffen: 17.07.2023).

Arbeitsstelle kleine Fächer. 2020. *Dokumentation des Informations- und Vernetzungsworkshops „Digitalisierung in Lehre und Forschung kleiner Fächer”*.https://www.kleinefaecher.de/beitraege/blogbeitrag/dokumentation-des-informations-und-vernetzungsworkshops-digitalisierung-in-lehre-und-forschung-kle (zugegriffen: 17.07.2023).

Asef, Esther, and Cosima Wagner. 2018. “Workshop-Bericht ‘Nicht-lateinische Schriften in multilingualen Umgebungen: Forschungsdaten und Digital Humanities in den Regionalstudien’.” In *DHd Blog*. https://dhd-blog.org/?p=10669.

BMBF. “Kleine Fächer – Große Potenziale - BMBF.” Bundesministerium für Bildung und Forschung - BMBF. (zugegriffen: 17.07.2023). https://www.bmbf.de/bmbf/de/forschung/geistes-und-sozialwissenschaften/kleine-faecher/kleine-faecher_node.html.

Burr, Elisabeth. 2006. “Discussion Paper: Internationalization, Multi-lingual & Multi-cultural agenda”. In *Alliance of Digital Humanities Organizations (ADHO): Leadership*. https://adho.org/wp-content/uploads/2022/07/MLMC_DiscussionPaper_2006.pdf (zugegriffen: 17.07.2023).

Dombrowski, Quinn. 2020. *What’s a ‘Word’: Multilingual DH and the English Default.* https://quinndombrowski.com/blog/2020/10/15/whats-word-multilingual-dh-and-english-default (zugegriffen: 17.07.2023).

Fiormonte, Domenico und Jonathan Usher (Hrsg.). 2001. *New Media and the Humanities: Research and Applications*. Proceedings of the first seminar Computers, Literature and Philology, Edinburgh 7-9 September 1998. Oxford: Humanities Computing Unit, University of Oxford.

Fiormonte, Domenico (Hrsg.). 2003. *Informatica Umanistica*. Dalla ricerca all'insegnamento. Atti dei convegni Computers, Literature and Philology, Università di Roma "La Sapienza" (1999) e Università di Alicante, Spagna (2000). Roma: Bulzoni.

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.

Gil, Alex, und Élika Ortega. 2016. “Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing.” In *Doing Digital Humanities: Practice, Training, Research*, 22–34. Abingdon: Routledge.

Grallert, Till, Xenia Monika Kudela, Eliese-Sophia Lincke, Colinda Lindermann, Jana-Katharina Mende, Jonas Müller-Laackman, und Larissa Schmid. 2023. *Umgang mit Multilingualität im DACH und DHd Verband (v1.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.7957187.

Horváth, Alíz. 2023. “DH in Japanese Studies, Japanese Studies in DH: Recent Trends, Tools, and Concepts.” *International Journal of Digital Humanities* 4, no. 1: 213–23. https://doi.org/10.1007/s42803-022-00063-6.

Kim, Dorothy, und Jesse Stommel (Hrsg.). 2018. *Disrupting the Digital Humanities*. punctum books. https://doi.org/10.21983/P3.0230.1.00.

Nicholas, Gabriel, und Aliya Bathia. 2023. *Lost in Translation. Large Language Models in Non-English Content Analysis*. Center for Democracy & Technology. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (zugegriffen: 17.07.2023).

Nicolás, Carlota und Massimo Moneglia (Hrsg.). 2005. *La gestione unitaria dell'eredità culturale multilingue europea e la sua diffusione in rete*. Atti della conferenza internazionale CLiP 2003 Computers, Literature and Philology , Firenze 04.-06.12.2003. Firenze: Firenze University Press.

Spence, Paul. 2021. *Disrupting Digital Monolingualism: A Report on Multilingualism in Digital Theory and Practice*. London: Language Acts & Worldmaking project. https://doi.org/10.5281/zenodo.5743283.

Weingart, S.B., N. Eichmann-Kalwara und M. Lincoln, et al. 2020. *The Index of Digital Humanities Conferences*. Carnegie Mellon University. https://doi.org/10.34666/k1de-j489.

Vorträge

Agreement und Kookkurrenz bei unzuverlässigem Erzählen. Ziele, Herausforderungen und erste Ergebnisse aus dem Projekt CAUTION

Blessing, André
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID: 0000-0001-7573-578X

Jacke, Janina
janina.jacke@uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland
ORCID: 0000-0001-7217-3136

Kuhn, Jonas
jonas.kuhn@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

Ziel des Projekts CAUTION („Computer-aided Analysis of Unreliability and Truth in Fiction – Interconnecting and Operationalizing Narratology“) ist die computergestützte Auseinandersetzung mit dem erzähltheoretischen Konzept des unzuverlässigen Erzählens. Unzuverlässiges Erzählen liegt dann vor, wenn die fiktive Erzählinstanz eines fiktionalen Textes unwahre Äußerungen über die fiktive Welt des Textes tätigt (vgl. Shen 2011). Ziel des Projekts ist es, in einem experimentellen Mixed-Methods-Setting Erkenntnisziele auf unterschiedlichen Ebenen zu erreichen. Zum einen soll geprüft werden, ob eine computationelle Modellierung des Konzepts (oder zumindest eine literaturwissenschaftlich nützliche Annäherung, vgl. Jacke 2023) möglich ist. Zum anderen geht es um die Erforschung bestimmter theoretischer und methodologischer Fragen, die mit dem Konzept verbunden sind und deren Klärung eine potenzielle reflektierte computationelle Modellierung informieren kann: Wie interpretationsabhängig ist die Feststellung von unzuverlässigem Erzählen in literarischen Texten? Und welche Rolle spielen formale bzw. sprachliche Textmerkmale (textuelle Indikatoren) bei der Feststellung?

Nach einer genaueren Vorstellung der Erkenntnisziele (Abschnitt 2) und der sich daraus ergebenden Projektkonzeption (Abschnitt 3) sollen erste Ergebnisse der Auswertung im Projekt erstellter Annotationen – mit besonderem Fokus auf Inter-Annotator Agreement und Annotationskookkurrenz – präsentiert und diskutiert werden (Abschnitt 4). Abschließend wird ein vorläufiges Fazit zu den im Projekt zutage tretenden Herausforderungen bezüglich adäquater Standards und Methoden bei der Evaluation statistischer Ergebnisse in der computationellen Literaturwissenschaft gezogen (Abschnitt 5).

Erkenntnisinteressen und Hypothesen

Das Konzept des unzuverlässigen Erzählens stellt eine besondere Herausforderung für die computationelle Modellierung dar, denn es wird in der Literaturwissenschaft gemeinhin als stark interpretationsabhängiges Konzept verstanden (vgl. Yacobi 1981; Kindt 2008: 53-67). Gleichzeitig werden aber auch Listen von (teilweise sprachlichen) Indikatoren zusammengestellt, die auf unzuverlässiges Erzählen hindeuten können (vgl. Nünning 1998). Dabei lassen sich in der Regel aber keine genauere Angaben darüber finden, wie relevant oder verlässlich die einzelnen Indikatoren sind. Eine weitere Herausforderung besteht darin, dass das Konzept meist als Kategorie zur Einordnung von Texten oder Erzählfiguren genutzt wird, zugleich aber sowohl eine Verwendung als Analysekategorie zur Einordnung von Textsegmenten als auch eine graduelle Anwendung des Konzepts naheliegen (vgl. Jacke 2020: 173-183).

Vor diesem Hintergrund stehen im hier vorgestellten Projektteil von CAUTION folgende Fragen und Hypothesen im Zentrum:

1) (Dis-)Agreement bzw. Interpretationsabhängigkeit: Sollte unzuverlässiges Erzählen tatsächlich ein stark interpretationsabhängiges [1] Konzept sein, ist zu erwarten, dass die Einschätzung von Annotator:innen ungewöhnlich stark auseinander liegen, das Inter-Annotator Agreement also besonders niedrig ist. Dies soll in CAUTION überprüft werden. In diesem Zusammenhang wird auch untersucht, in welchem Verhältnis dazu das Agreement bei der Feststellung ausgewählter sprachnaher Phänomene steht, die in der literaturwissenschaftlichen Forschung als Indikatoren für Unzuverlässigkeit betrachtet werden. Eine Hypothese ist, dass sich zumindest für einige sprachliche Indikatoren ein vergleichsweise hohes Agreement erzielen lässt. Die Ergebnisse der Überprüfung dieser Hypothesen sind nicht nur literaturtheoretisch interessant, sondern erlauben auch vorläufige Schlüsse hinsichtlich der Frage, ob sich Unzuverlässigkeit oder die Indikatoren computationell operationalisieren lassen.

2) Relation zwischen Indikatoren und unzuverlässigem Erzählen: Sollte die Einordnung der in der Unzuverlässigkeitsforschung genannten Phänomene als Indikatoren tatsächlich zutreffen, müssten sich signifikante Relationen zwischen dem Vorkommen der Indikatoren und unzuverlässigem Erzählen andeuten. Falls sich eine ausrei-

chend hohe Übereinstimmung feststellen lässt, wäre das wiederum interpretationstheoretisch und -methodologisch interessant – und die computationelle Modellierung bestimmter Indikatoren könnte ggf. als literaturwissenschaftlich vertretbare Annäherung an eine Modellierung unzuverlässigen Erzählens selbst vorgenommen werden.

Projektkonzeption

Um im Hinblick auf die im vorangegangenen Abschnitt genannten Fragen zumindest erste Einblicke zu erzielen, wird eine Variante der Methode der kollaborativen Annotation (vgl. Gius und Jacke 2017) eingesetzt.[2] Das hat einen doppelten Vorteil: Zum einen können hiermit vorläufige Antworten auf die genannten Fragen hinsichtlich Agreement und Kookkurrenz statistisch errechnet und übersichtlich dargestellt werden. Zum anderen wird zugleich eine erste Datenbasis erstellt, die Experimente mit der computationellen Modellierung bzw. machine-learning-basierten Automatisierung geeigneter (Teil-)Phänomene erlaubt.

Das Kernkorpus für diesen Projektteil setzt sich aus neun [3] deutschsprachigen literarischen Erzählungen aus dem Zeitraum zwischen dem 19. Jahrhundert und der Gegenwart zusammen, von denen vier tendenziell als unzuverlässig gelesen werden, vier tendenziell zuverlässig erzählt sind, aber mit Unzuverlässigkeit verwandte Phänomene aufweisen (bspw. eine merkwürdige fiktive Welt oder eine stark personale Erzählinstanz), und eine keine derartige Besonderheit enthält (Abb 1).[4]

Text	Jahr	Tokens
Hoffmann: Der Sandmann (Teil 1)	1816	3597
Hoffmann: Der Sandmann (Teil 2)	1816	1268
Hoffmann: Der Sandmann (Teil 3)	1816	617
Hoffmann: Der Sandmann (Teil 4)	1816	9295
Eichendorff: Auch ich war in Arkadien	1834	7034
Ebner-Eschenbach: Die Spitzin	1901	4283
Schnitzler: Andreas Thameyers letzter Brief	1902	2932
Kafka: Ein Bericht für eine Akademie	1917	3776
Perutz: Nur ein Druck auf den Knopf	1930	4030
Fallada: Gute Krüseliner Wiese rechts	1991 (posthum)	3996
Kehlmann: Rosalie geht sterben	2009	7074
Bendixen: Meine falschen Eltern	2012	2217

Abb. 1: Korpusübersicht

Im Rahmen des annotationsbasierten Projektteils gibt es zwei Annotationsaufgaben: Die erste zielt auf die Feststellung ausgewählter Indikatoren, die andere auf die Identifikation von unzuverlässigem Erzählen.

A) Annotation von Indikatoren: Aus den Indikatorenlisten der Unzuverlässigkeitsforschung wurden sechs Phänomene ausgewählt. Bei allen handelt es sich um Eigenschaften der Erzählfigur, die sich (unserer Hypothese zufolge) einerseits weitgehend an sprachlichen Textmerkmalen festmachen lassen und die andererseits dafür sorgen, dass die Erzählfigur mit erhöhter Wahrscheinlichkeit unzuverlässig ist. Die ausgewählten Indikatoren bzw. Eigenschaften sind a) emotionale Erregung, b) Unsicherheit, c) die Absicht, besonders sicher zu wirken, d) Bewusstsein für eine kommunikative Situation, e) Abgelenktheit und f) die Absicht, abzulenken.[5] Nach einem automatischen *sentence splitting* sollen die Annotator:innen für jede der Eigenschaften und für jeden Satz der Erzählungen entscheiden, ob sie Anzeichen für das Vorliegen der Eigenschaft feststellen können. Dabei stehen die Werte „yes“, „no“ und „undecided“ zur Verfügung – mit letzterem Wert kann eine wahrgenommene Ambiguität eines Satzes im Hinblick auf die relevanten Eigenschaften markiert werden.

B) Annotation von Unzuverlässigkeit: In einem späteren, separaten Schritt werden die Annotator:innen gebeten, diejenigen Textstellen zu annotieren, an denen die Erzählfigur ihres Erachtens eine inkorrekte Äußerung über die fiktive Welt tätigt. Annotierte Einheiten sind dabei in der Regel Teilsätze, die eine (inkorrekte) Proposition ausdrücken. Als Werte können „yes“ und „undecided“ vergeben werden, nicht annotierte Stellen gelten als „no“. Der leicht geänderte Annotationsprozess ergibt sich daraus, dass in Bezug auf unzuverlässiges Erzählen eine satzbasierte Annotation unnötig ungenau wäre: Die Identifikation der inkorrekten Proposition spielt eine wichtige Rolle, und gerade in literarischen Texten werden pro Satz oft mehrere Propositionen ausgedrückt.

Jeder Korpustext wird von mindestens zwei Annotator:innen auf der Basis gemeinsamer Annotationsrichtlinien bearbeitet. Allgemeine Annotationsprobleme werden regelmäßig gemeinsam diskutiert und die Richtlinien werden entsprechend überarbeitet. Sobald jeder Text in einer ersten Runde bearbeitet worden ist, erfolgt eine Überarbeitungsrunde, die der finalen Version der Richtlinien folgt. Individuelle Annotationsentscheidungen werden nicht diskutiert oder von den Annotator:innen verglichen, aber es besteht die Möglichkeit, Gründe oder Unsicherheiten bei der Annotation als Kommentar im Text zu vermerken. [6]

Erste Auswertung vorläufiger Annotationsergebnisse

Zum Verfassungszeitpunkt dieses Beitrags ist die erste Annotationsaufgabe abgeschlossen. Für die zweite Aufgabe liegen Ergebnisse für einige der Korpustexte vor.[7] Im vorliegenden Abschnitt sollen erste Ergebnisse für die zwei oben genannten Fragekomplexe vorgestellt werden. Ein besonderer Fokus liegt dabei auf den Entscheidungsgründen für die genaue Konfiguration der Abfragen, auf den für die Deutung der Ergebnisse verwendeten Methoden sowie auf den Konsequenzen, die sich aus den Ergebnissen ziehen lassen.

1) (Dis-)Agreement bzw. Interpretationsabhängigkeit: Für die Berechnung des Inter-Annotator Agreements ist zunächst zu entscheiden, welches Maß gewählt werden soll. Für die Ergebnisse der ersten Annotationsaufgabe erscheint die Verwendung von *Cohen’s kappa* sinnvoll (vgl. Reiter und Konle 2022). [8] Bei der zweiten Annotationsaufgabe ist allerdings problematisch, dass für die nicht explizit ausgeführten „no“-Annotationen keine Segmentierung vorgenommen wurde und eine Berechnung des Agreements

mit *kappa* deswegen zu schwer interpretierbaren und mit der ersten Aufgabe nicht vergleichbaren Ergebnissen führt. Um das Agreement zwischen der ersten und der zweite Annotationsaufgabe vergleichen zu können, wurde eine alternative Agreement-Berechnung verwendet, die anzeigt, wie viel Prozent der positiven Annotationen bei paarweiser Auswertung übereinstimmend sind (Anzahl übereinstimmender positiver Annotationen) / (Anzahl übereinstimmender positiver Annotationen + Anzahl nicht übereinstimmender positiver Annotationen). [9] Das Inter-Annotator Agreement wurde dabei immer paarweise berechnet. Pro Werk und Kategorie ergeben sich dadurch drei bis sechs Agreement-Werte für jeweils drei bis vier Annotator:innen. Die Verteilungen dieser Werte werden in den folgenden Abbildungen in Boxplots dargestellt. Die "undecided"-Annotationen wurden in diesem ersten Zugang zur Datenanalyse als „yes" gewertet.[10]

Während sich das so errechnete Agreement auf diese Weise noch nicht uneingeschränkt mit üblicherweise in den Feldern der Computerlinguistik und der computationellen Literaturwissenschaft erzielten Werten vergleichen lässt, lassen sich zumindest projektintern interessante Beobachtungen treffen: So kann beispielsweise festgestellt werden, dass – entgegen unserer ursprünglichen Hypothese – unzuverlässiges Erzählen anscheinend mit einem höheren Agreement festgestellt wird als die sprachnahen Indikatoren (Abb 2).

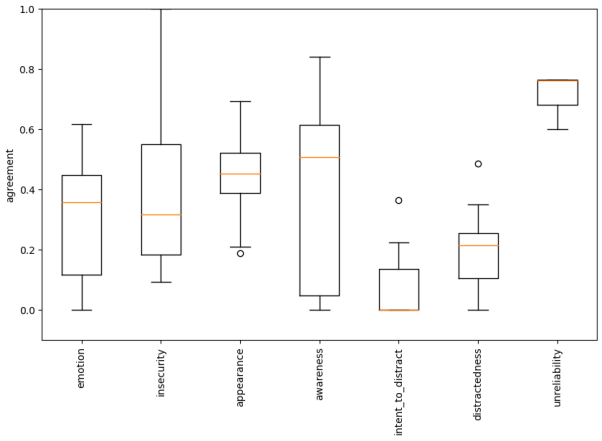


Abb. 2: Verhältnis Agreement Annotationsaufgaben A und B [11]

Mögliche Erklärungen könnten sein, dass Erzähler:inneneigeschaften wie Emotionalität sich schwieriger an genauen Textstellen festmachen lassen als eine inkorrekte Äußerung (also Unzuverlässigkeit). Außerdem ist anzunehmen, dass die Eigenschaften aus der ersten Annotationsaufgabe öfter in Abstufungen vorliegen und die Annotator:innen möglicherweise unterschiedliche intuitive Schwellenwerte haben, ab denen sie eine Annotation vornehmen.[12] Bei der zweiten Annotationsaufgabe ist die Entscheidung dagegen binär: Eine Äußerung ist entweder korrekt oder inkorrekt.

Auffällig ist weiterhin, dass – innerhalb der ersten Annotationsaufgabe (Abb. 3) – für die Eigenschaft *awareness of communicative situation* ein vergleichsweise hohes Agreement erzielt werden konnte, während das niedrigste Agreement für *intent to distract* verzeichnet wird.

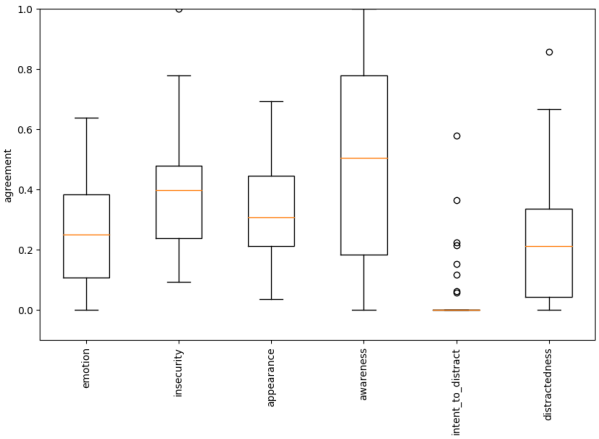


Abb. 3: Agreement Annotationsaufgabe A

Das hohe Agreement ließe sich dadurch erklären, dass das Bewusstsein einer Erzählinstanz, sich in einer kommunikativen Situation zu befinden, sich in den meisten Fällen tatsächlich an linguistischen Texteigenschaften festmachen lässt – beispielsweise an direkter Adressat:innenansprache unter Verwendung der zweiten Person bei Personalpronomen und Verbformen. Diese Hypothese ließe sich zukünftig durch Anwendung geeigneter Computermodelle auf das Korpus überprüfen, die auf die automatische Feststellung potenziell relevanter sprachlicher Eigenschaften zielen.[13] Insbesondere im Hinblick auf die Eigenschaft der Emotionalität liegen Modelle vor, die – aufgrund des mäßigen Agreements zwischen Annotator:innen – zumindest eindeutige Fälle der entsprechenden Phänomene möglicherweise hinreichend gut erkennen können (vgl. Klinger et al. 2020). Die niedrige Übereinstimmung hinsichtlich der Ablenkungsabsicht einer Erzählinstanz steht dagegen im Einklang mit der Analyse der Diskussionen im Annotator:innenteam, die persistente definitorische Unzulänglichkeiten des relevanten Analysekonzepts offenbaren.

2) Relation zwischen Indikatoren und Unzuverlässigkeit: Um die Frage zu untersuchen, ob die ausgewählten Indikatoren auffallend häufig mit der Feststellung von unzuverlässigem Erzählen korrelieren, wurden vorerst textstellenbasierte Kookkurenz betrachtet. Dabei wurde errechnet, wie oft die Indikatoren einerseits im Korpus (im Verhältnis zur Menge der Sätze) annotiert worden sind und wie oft sie andererseits an den Stellen festgestellt wurden, an denen auch Unzuverlässigkeit diagnostiziert wurde.[14] Um diese Frage frei von Agreement-Problemen zu halten, wurden die Kookkurrenzen jeweils nur annotator:innenintern errechnet und dann der Mittelwert gebildet. Ein Vergleich der relativen Häufigkeit der potenziellen Indikatorphänomene im Korpus (Abb. 4) mit ihrer Häufigkeit an Textstellen, an denen Unzuverlässigkeit diagnostiziert worden ist (Abb. 5), zeigt, dass das Vorkommen bei allen Kategorien (bis auf

insecurity) zumindest leicht erhöht ist. Deutlich erhöhtes Vorkommen lässt sich bei *emotion* sowie bei *appearance* (der Absicht, besonders sicher wirken zu wollen) feststellen.

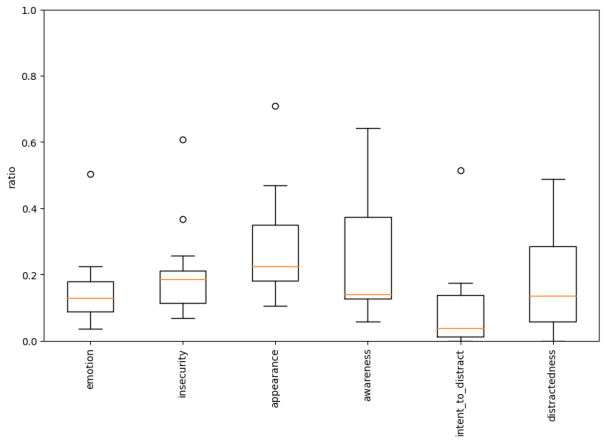


Abb. 4: Relative Häufigkeit der Indikatoren im Korpus (Aufgabe A)

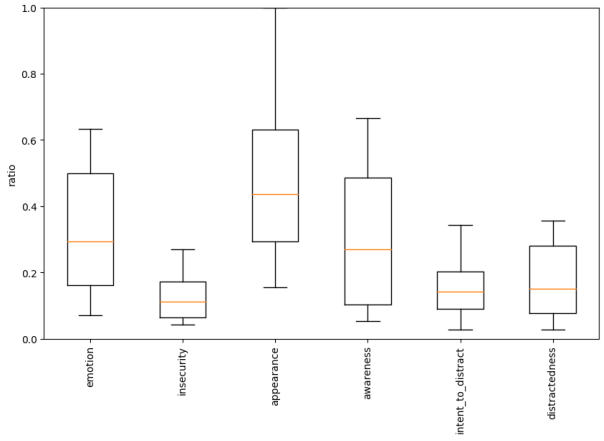


Abb. 5: Relative Häufigkeit der Indikatoren (Aufgabe A) bei Unzuverlässigkeit (Aufgabe B)

Um die Indikationskraft der Erzähler:inneneigenschaften weiter zu prüfen, könnte zukünftig genauer untersucht werden, ob eine bestimmte Kombination von Indikatorkategorien besonders häufig mit dem Auftreten von Unzuverlässigkeit korreliert. Zudem sollen neben der hier in einem ersten Zugang skizzierten textstellenbasierten Analyse der Kookkurrenzen auch die Gesamttexte untersucht werden, um festzustellen, ob die Indikatorphänomene gehäuft in Texten mit vielen Vorkommnissen von Unzuverlässigkeit auftreten – auch wenn die genauen Textstellen nicht notwendigerweise übereinstimmen. [15]

Fazit

Die hier präsentierten ersten Auswertungen der im Projekt CAUTION durchgeführten Annotationen zeigen, dass sich besondere Herausforderungen aus dem doppelten Erkenntnisinteresse ergeben: Es sollen sowohl computerlinguistische Erkenntnisse hinsichtlich der computationellen Modellierbarkeit von unzuverlässigem Erzählen als auch literaturtheoretische Einsichten in die Interpretationsabhängigkeit des Konzepts und seine Relation zu bestimmten sprachnahen Indikatoren gewonnen werden. Obwohl die Erkenntnisinteressen in weiten Teilen Schnittflächen aufweisen, ergeben sich bei der genaueren Konzeption der Annotationsaufgaben sowie bei der Auswertung tendenziell Diskrepanzen u.a. im Zusammenhang mit Korpusgröße, Annotationstools, annotierten (oder nicht-annotierten bzw. nicht-segmentierten) Einheiten, der Wahl eines Agreement-Maßes und der Beurteilung der Agreementwerte.

Dennoch zeichnen sich erste interessante Tendenzen ab. So ist aus literaturtheoretischer Perspektive bemerkenswert, dass bei der Feststellung des vermeintlich stark interpretationsabhängigen unzuverlässigen Erzählens tendenziell ein höheres Agreement erzielt werden konnte als bei der Identifikation vermeintlich sprachnaher Indikatoren. Darüber hinaus scheinen die Indikatoren insgesamt eine schwache positive Indikationskraft für Unzuverlässigkeit zu haben, wobei insbesondere Emotionalität und ein betont sicheres Auftreten der Erzählinstanz wichtige Rollen spielen könnten.

Nächste Schritte könnten sich der Analyse von textbasierten (im Gegensatz zu textstellenbasierten) Kookkurrenzen zuwenden. Außerdem ist die genauere Analyse von besonders auffälligen Einzeltexten und Textstellen interessant – beispielsweise dort, wo unzuverlässiges Erzählen ohne die untersuchten Indikatoren auftritt. Hierfür eignen sich besonders interaktive Visualisierungen, die Annotationen im Textverlauf abbilden und ein Close Reading von Text und Annotator:innenkommentaren zulassen (vgl. Münz-Manor und Marienberg-Milikowsky 2023). Weitere Auswertungen könnten sich der Frage zuwenden, ob die literaturwissenschaftlichen Einschätzungen zur Unzuverlässigkeit der Korpustexte mit der Häufigkeit von Unzuverlässigkeitsannotationen in den jeweiligen Texten im Einklang stehen und welche weiteren Faktoren zusätzlich zur Frequenz ggf. einzubeziehen wären.

Im Zusammenhang mit der computationellen Modellierbarkeit ist interessant, dass sich die Unzuverlässigkeitsannotationen durch ihr höheres Agreement tendenziell eher für ein Trainieren von Modellen eignen als – wie ursprünglich angenommen – die annotierten Indikatoren. Potenziell schwierig ist bei diesem Szenario aber die Tatsache, dass die Unzuverlässigkeitsannotationen auf dem Verstehen der ausgedrückten Proposition und einer Vorstellung von der erzählten Welt basieren. Vor diesem Hintergrund erscheint ein paralleles Experimentieren mit Modellen zur automatischen Feststellung relevanter Indikatoren sinnvoll.

Fußnoten

1. Interpretationsabhängigkeit liegt vor, wenn die Anwendung eines Konzepts in hohem Maße von (strittigen) extratextuellen Annahmen und/oder nicht-wahrheitserhaltenden Schlüssen wie Induktion oder Abduktion abhängig ist (vgl. Jacke, in Vorbereitung).
2. Wir verstehen das relevante Annotationssetting insofern als kollaborativ als hier mehrere Annotator:innen dieselben Texte unter derselben Aufgabenstellung bearbeiten, allgemeine Annotationsprobleme gemeinsam besprochen und darauf basierend die Annotationsguidelines angepasst werden. Dieses Vorgehen ist aber abzugrenzen von einer engeren Form der Kollaboration, bei der etwa konkrete Annotationsinstanzen gemeinsam erstellt werden oder sich die Annotator:innen über individuelle Annotationsentscheidungen systematisch austauschen und abstimmen.
3. Aus computationeller und statistischer Perspektive handelt es sich um ein extrem kleines Korpus, weshalb hierauf basierende Erkenntnisse und Modellierungen nur vorläufigen Charakter haben können. Da literaturwissenschaftliche Annotationen in der Regel sehr zeitintensiv sind, war aus pragmatischen Gründen die Bearbeitung eines größeren Korpus nicht möglich.
4. Dabei ist die Zuordnung der Texte zu den tendenziellen Eigenschaften wie folgt: Unzuverlässig erzählt sind die Texte von Hoffmann, Schnitzler, Perutz und Bendixen. Verwandte Phänomene enthalten die Texte von Eichendorff (Satire), Kafka (merkwürdige fiktive Welt), Kehlmann (Metalepse) und Fallada (stark personale Erzählinstanz). Der Text von Ebner-Eschenbach enthält keines dieser Phänomene.
5. Unsere Auswahl aus den zahlreichen in der Forschung zu unzuverlässigem Erzählen genannten Indikatoren ist zum einen bedingt durch den Versuch, damit unterschiedliche Formen unzuverlässigen Erzählens abzudecken (z.B. solche mit und solche ohne bewusste Täuschungsabsicht der Erzählfigur. Zum anderen wollten wir eine Skala zwischen möglichst formal-sprachlich operationalisierbaren Indikatoren und voraussetzungsreicheren Indikatoren abdecken.
6. Der hier vorgestellte wird durch drei weitere Projektteile ergänzt, von denen zwei stärker an computerlinguistisch-algorithmischen, der andere stärker an literaturwissenschaftlich-interpretationstheoretischen Erkenntnisinteressen ausgerichtet ist. Der erste computerlinguistische Projektteil untersucht existierende Modelle zur automatischen Feststellung von Textphänomenen, die den in der ersten Annotationsaufgabe annotierten Phänomenen nahekommen (z.B. *emotion* bzw. *sentiment analysis* oder *uncertainty detection*), um festzustellen, wie nah entsprechende automatische Annotationen den manuellen Annotationsergebnissen kommen. Der zweite computerlinguistische Projektteil prüft die automatische Erkennung unzuverlässiger Erzählungen mithilfe von *deep learning*-Verfahren, die auf kanonisch kategorisierte Textkorpora angewandt werden. Der literaturwissenschaftliche Projektteil beschäftigt sich mit der Frage, ob das Inter-Annotator Agreement bei der Feststellung unzuverlässigen Erzählens durch Analyse der zugrundeliegenden Argumentation und intensive Diskussion zwischen den Annotator:innen erhöht werden kann. Zum Verfassungszeitpunkt dieses Beitrags befinden sich diese drei Projektteile noch in Vorbereitung.
7. Das dieser Studie zugrunde liegende Datenset ist in seiner Konstitution komplex und kann hier nicht im Detail vorgestellt werden. Alle relevanten Zahlen und die Annotationsrichtlinien werden aber ab Februar 2024 unter zu finden sein. Um die Dimensionen grob einschätzen zu können, sei hier aber ein Beispiel genannt: Im ersten Teil des Textes *Der Sandmann* (3.597 Tokens, 192 Sätze), der von der literaturwissenschaftlichen Forschung tendenziell als unzuverlässig eingeschätzt wird, hat Annotator:in A im Rahmen der satzbasierten Annotation folgende Einschätzungen vorgenommen: emotionale Erregung 24-mal „yes“, 4-mal „undecided“; Unsicherheit 1-mal „yes“, 8-mal „undecided“; Absicht, sicher zu wirken 11-mal „yes“, 6-mal „undecided“; Bewusstsein einer kommunikativen Situation 21-mal „yes, 0-mal „undecided“; Absicht, abzulenken 0-mal „yes“, 0-mal „undecided“; Abgelenktheit 1-mal „yes“, 4-mal „undecided“. Annotator:in A hat im Rahmen der teilsatzbasierten Annotation für den ersten Teil des *Sandmanns* 17 Vorkommnisse eines *incorrect statement* (also unzuverlässiges Erzählen) festgestellt und 15-mal potenzielle Unzuverlässigkeit („undecided“). Was die relative Häufigkeit an Unzuverlässigkeitsannotationen angeht, liegt dieser Text gewissermaßen im Mittelfeld: Annotator:in A hat, um zwei Extreme zu nennen, in *Auch ich war in Arkadien* (7.034 Tokens, 316 Sätze) 228-mal unzuverlässiges Erzählen (plus 14-mal „undecided“) festgestellt und in *Die Spitzin* (4.382 Tokens, 230 Sätze) 6-mal unzuverlässiges Erzählen (plus 6-mal „undecided“). Uns ist bewusst, dass diese Studie aufgrund der kleinen Korpusgröße und Datenmenge lediglich erste Tendenzen aufweisen kann. Im Rahmen des Vortrags werden wir das Datenset etwas ausführlicher vorstellen.
8. Wir konnten für Annotationsaufgabe A in den meisten Fällen ein moderates Agreement erzielen mit *kappa*-Werten zwischen 0,4 und 0,6.
9. Es erfolgt also keine *chance correction*. Ein Vergleich der *kappa*-Werte mit dem hier verwendeten Agreementmaß ergibt aber, dass die Werte tendenziell vergleichbar sind.
10. Diese Entscheidung wurde hier in einem ersten Zugang getroffen, um die Stellen zu erfassen, an denen der Verdacht von Unzuverlässigkeit zumindest im Raum steht. Für Analysen, bei denen nunr tatsächlich übereinstimmende Kategorien („yes“ mit „yes“ und „undecided“ mit „undecided“) als Agreement gewertet wird, sind deutlich niedrigere Agreement-Werte zu erwarten.
11. Diese Grafik bezieht die Annotationen zu drei Texten ein, zu denen zum Verfassungszeitpunkt alle relevanten Annotationen vorliegen und auswertbar sind.

12. Beispielsweise könnte eine gewisse „Grundemotionalität“ (insbesondere bei personalen Erzählinstanzen) von einigen Annotator:innen als Default angenommen und entsprechend nicht als „yes“ annotiert werden, während andere Annotator:innen diese feststellbare leichte emotionale Erregung als ausreichend für eine „yes“-Annotation einordnen. Ein plausibler Standard für entsprechende Fälle lässt sich nicht oder nur schwer pauschal in Form von Annotationsrichtlinien festlegen. Bei der Frage nach der Inkorrektheit von Behauptungen dagegen erscheint zum einen das Denken in Graden grundsätzlich weniger naheliegend. Zum anderen könnte es den Annotator:innen leichter gefallen sein, sich auch in eventuell auftretenden Fällen einer nur geringen Abweichung der Erzähläußerungen von der fiktiven Wahrheit für die Annotation als *incorrect statement* zu entscheiden.
13. Konkret könnten regelbasierte Verfahren eingesetzt werden, um die genannten sprachlichen Merkmale zu annotieren (Personalpronomen und Verbformen in zweiter Person, zusätzlich z.B. Imperative, Appellative), so dass sich dann über Kookkurrenzanalysen feststellen ließe, ob eine hinreichende Kollokation mit den als "Bewusstsein einer kommunikativen Situation" annotierten Textstellen besteht.
14. Zusätzlich zur textstellenbasierten Analyse von Kookkurrenz ist eine textbasierte denkbar, in deren Rahmen die relative Häufigkeit des Auftretens von Indikatoren und Unzuverlässigkeit nicht spezifisch in Textpassagen, sondern in ganzen Texten untersucht wird.
15. In diesem Zusammenhang können auch die in diesem Beitrag noch nicht diskutierten textbasierten Annotationen herangezogen werden, bei denen pro annotierter Kategorie (d.h. bzgl. der sechs Indikatorkategorien und Unzuverlässigkeit) ein Prozentwert für den Gesamttext vergeben wurde, der das geschätzte Maß anzeigt, mit welchem die fragliche Eigenschaft in dem Text bzw. bei der Erzählinstanz vorliegt.

Bibliographie

Gius, Evelyn und Janina Jacke. 2017. „The Hermeneutic Profit of Annotation. On Preventing and Fostering Disagreement in Literary Analysis“. In *International Journal of Humanites and Arts Computing* 11 (2), 233-254.

Jacke, Janina. 2023. „Vom sprachlichen Indikator zum komplexen Phänomen? Operationalisierungsprobleme in der computationellen Literaturwissenschaft am Beispiel des unzuverlässigen Erzählens“. In Open Humanities, Open Culture. DHd 2023 Trier. Konferenzabstracts , 317-321. DOI: 10.5281/zenodo.7688632 .

Jacke, Janina. In Vorbereitung. „Operationalization and Interpretation Dependence in Computational Literary Studies“.

Kindt, Tom. 2008. *Unzuverlässiges Erzählen und literarische Moderne. Eine Untersuchung der Romane von Ernst Weiß*. Tübingen: Niemeyer.

Klinger, Roman, Evgeny Kim und Sebastian Padó. 2020. „Emotion Analysis for Literary Studies. Corpus Creation and Computational Modeling“. In Nils Reiter, Axel Pichler und Jonas Kuhn (Hg.): Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt . De Gruyter. 237-268. DOI: https://doi.org/10.1515/9783110693973-011 .

Münz-Manor, Ophir und Itay Marienberg-Milikowsky. 2023. „Visualization of Categorization: How to See the Wood and the Trees“. Digital Humanities Quarterly 17 (3). http://www.digitalhumanities.org/dhq/vol/17/3/000703/000703.html (zugegriffen: 24. November 2023).

Nünning, Ansgar. 1998. „'Unreliable Narration' zur Einführung: Grundzüge einer kognitiv-narratologischen Theorie und Analyse unglaubwürdigen Erzählens“. In Ansgar Nünning, Carola Surkamp und Bruno Zerweck (Hg.): *Unreliable Narration. Studien zur Theorie und Praxis unglaubwürdigen Erzählens in der englischsprachigen Erzählliteratur*. Trier: Wissenschaftlicher Verlag Trier, 3-39.

Reiter, Nils und Leonard Konle. 2022. „Messverfahren zum Inter-annotator-agreement (IAA). Eine Übersicht“. DARIAH-DE Working Papers 44. Göttingen. DOI: https://doi.org/10.47952/gro-publ-103 .

Shen, Dan. 2011. „Unreliability“. In Peter Hühn et al. (Hg.): the living handbook of narratology . Hamburg: Hamburg University. http://www.lhn.uni-hamburg.de/article/unreliability (zugegriffen: 18. Juli 2023).

Yacobi, Tamar. 1981. „Fictional Reliability as a Communicative Problem“. *Poetics Today* 2 (2), 113-126.

Applied Text as Graph (ATAG)

Kuczera, Andreas
andreas.kuczera@mni.thm.de
Technische Hochschule Mittelhessen, Deutschland
ORCID: 0000-0003-1020-507X

Stand der Forschung

Zahlreiche Konferenzen, Workshops, Artikel und Blogbeiträge in den letzten Jahren beschäftigten sich mit der Frage, was Text eigentlich ist. Dabei hat sich der Eindruck verfestigt, dass unser Verständnis von Text viele Interpretationen zulässt und daher ständig im Wandel ist. Dies spiegelt sich auch in den TEI-Richtlinien und den verschiedenen Wegen wider, wie sie angewendet werden. Dabei können Textedierende Module und Elemente sorgfältig anpassen, um zu einem Set von Kodierungsrichtlinien zu gelangen, das mit ihrer Interpretation und ihrem Forschungsinteresse am Quelltext übereinstimmt. Dennoch

unterscheidet sich oft die Art und Weise, wie die Textdaten auf einem Computer gespeichert werden, vom intellektuellen Verständnis des Textes des Editors. Das bedeutet, dass textliche Merkmale, die nicht natürlich in das Hierarchiemodell von XML passen, nur mit Hilfe von Umwegen oder zusätzlicher (vokabularspezifischer) Codierung adäquat digital dargestellt werden können. Je mehr zusätzliche Codierung erforderlich ist, desto komplizierter wird es, den Text zu kodieren, zu verarbeiten oder abzufragen. Auch in der Computerlinguistik und in den Computational Literary Studies wird das Problem der konkreten Verbindung von Theorie und Datenmodellierung diskutiert (Pichler/Reiter 2022; Bode 2023). Es ist davon auszugehen, dass die digitale Editionswissenschaft an der Schnittstelle von Wissenschaft und Technik mit demselben Problem en kämpft, weil Forschende immer nur lose und implizit wissen, von welchen Zeichen sie sprechen, wenn sie Interpretationen und Schlussfolgerungen äußern. Dies führt zu einer elementaren informationellen Ungenauigkeit, die sich auf alle Interpretationsebenen ausbreitet und die Arbeit hemmen. Nicht zuletzt deswegen werden in den digitalen Editionswissenschaften seit einiger Zeit die Vorteile von Standoff-Formaten und Text as Graph diskutiert. Zuletzt haben Bleeker et al. (2022) ihren Vorschlag zu einer Text as Graph Markup Language (TagML) vorgestellt und dem bisher führenden Standard TEI-XML gegenübergestellt. Standoff-Formate werden in der Computerlinguistik sehr häufig verwendet und sind ein sehr robustes Format. Allerdings kann bei ihnen in der Regel bereits a n notierte r Text später nicht mehr geändert werden. Bei TagML ist dies möglich, allerdings wird TagML bisher kaum produktiv eingesetzt. Das ursprünglich von Desmond Schmidt (Schmidt 2016) entwickelte und in (Neill/Kuczera 2019) vorgestellte Standoff-Property-System (SPO) vereint einen s tandoff-basierten Ansatz mit einer Labeled-Property-Graphdatenbank. Darauf aufbauend wird hier das Konzept Applied Text as Graph (ATAG) vorgestellt, das bereits im Rahmen des DFG-geförderten Projekts zum Liber Epistolarum der Hildegard von Bingen (https://liberepistolarum.mni.thm.de/home) eingesetzt wird (Kuczera 2020) . Die Software der Publikationsumgebung (vgl. Abb. 1) steht auf GitHub zur Nachnutzung zur Verfügung (https://github.com/digicademy/graph-dse).

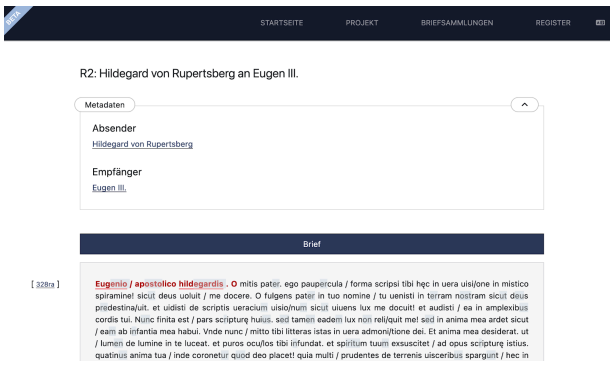


Abb 1: Der Anfang des Briefs R2 der Handschrift R des Liber Epistolarum der Hildegard von Bingen. URL: https://liberepistolarum.mni.thm.de/id/300dcfe1-9f1a-4e21-914d-4730fd85f1d2 (abgerufen am 6.7.2023).

Auf Grundlage dieser Publikationsumgebung wurde im folgenden die Webseite des DFG-Projekts zu den Sozinianischen Briefwechseln (https://gepris.dfg.de/gepris/projekt/3245185) von einer Typo3-basierten Webseite auf eine graphbasierte Publikationsumgebung umgestellt. Dabei wurden die in TEI-XML vorliegenden Briefe mit Hilfe eines TEI2json-Konverters (https://gitlab.rlp.net/adwmainz/digicademy/ sbw/tei2json) in das SPO-JSON-Format konvertiert und konnten so sehr einfach in die graphbasierte Publikationsumgebung eingespielt werden.

Auch wenn auf der Publikationsseite alles reibungslos funktioniert, haben sich im Hildegard-Projekt im Verlauf des Projekts doch einige Herausforderungen ergeben. So werden im Editionssystem Codex (vgl. Abb. 2) Texte, die bearbeitet werden, komplett in den Browser geladen. Bei Briefen mit einer Länge bis zu zehn DINA4-Seiten (nur als Größenvergleich) ist das noch realisierbar. Bei langen Kapiteln mit vielleicht 40 Seiten stößt das System an seine Grenzen, da die Bearbeitung bei größeren Textlängen sehr träge und langsam wird.

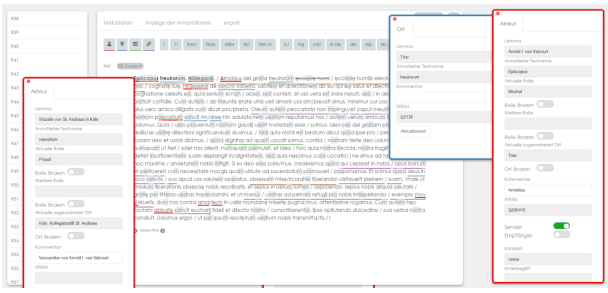


Abb. 2: Das graphbasierte Editionssystem Codex (Neill/Kuczera 2019).

Applied Text as Graph (ATAG)

Mit *Applied Text as Graph* (ATAG) schlagen wir ein neues Konzept von Text as Graph vor, das nicht nur denselben Grad notwendiger Flexibilität mitbringt, wie man ihn

bei TagML findet, sondern durch seine Anlehnung an SPO auch eine performantere technische Umsetzung in digitalen Editionsprojekten ermöglicht.

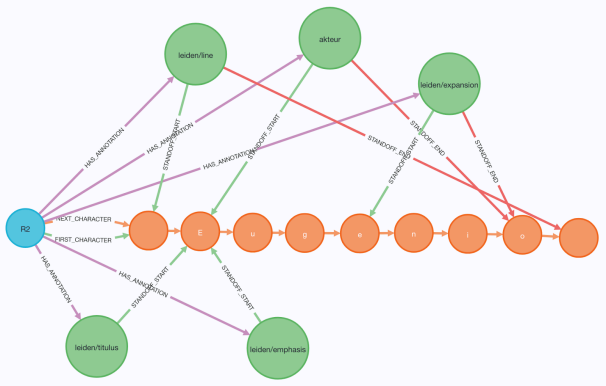


Abb 3.: Das Graphmodell des ersten Wortes des Briefes R2 aus der Handschrift R des Liber Epistolarum der Hildegard von Bingen. Der Grundtext ist normalisiert, die Annotation leiden/expansion zeigt an, dass im Rahmen der Transkription die Zeichen ‘enio’ ergänzt wurden (Quelle: Autor).

Die Grundlagen sind:

- Ein linearer Text (Textstream) bestehend aus Zeichen und Leerzeichen. Dieser lineare Text wird als Textstück bezeichnet.
- Die Zeichen (und Leerzeichen) eines Textstückes werden in Zeichenknoten (Orange Knoten in Abb. 3) abgebildet, die untereinander mit NEXT-Kanten verbunden sind.
- Jedes Textstück beginnt mit einem Textknoten (Blauer Knoten in Abb. 3). Von diesem Textknoten geht eine FIRST_CHARACTER-Kante und eine NEXT-Kante zum ersten Zeichen der Zeichenkette und eine LAST_CHARACTER-Kante zum letzten Zeichen der Zeichenkette.
- Die kleinste Granularität ist die Zeichenebene, deren Reihenfolge mit NEXT-Kanten festgehalten wird.
- Jeder Zeichenknoten ist über eine UUID eindeutig identifizierbar und über eine persistent stabile und auflösende URI über das Internet adressierbar.
- Die Ketten von Zeichenknoten können mit Annotationsknoten annotiert werden (Grüne Knoten in Abb. 3).
- Die Annotationsknoten sind über FIRST_CHARACTER- und LAST_CHARACTER-Kanten mit den Zeichenknoten verbunden und machen die Reichweite einer Annotation explizit.
- Eine Annotation kann mit einem weiteren Textstück verbunden werden, in dem z.B. eine alternative Lesart oder ein Kommentar enthalten ist. Damit ergibt sich ein Netzwerk von Texten und Annotationen mit Texten.
- Der Grundtext (auf den Begriff Basistext wird bewusst verzichtet, da keinerlei Hierarchie hergestellt werden soll) und die Annotationen bilden die Grundlage dieses Textmodellierungssystems.
- Das Ende einer Zeichenknotenkette kann mit einer NEXT-Kante mit dem ersten Zeichenknoten einer weiteren Zeichenknotenkette verbunden werden und gibt damit eine mögliche Leserichtung wieder.
- Das System macht keine Vorgaben, wie z.B. die Metadaten von Texten festgehalten werden. Diese könnten beispielsweise gemäß der gut dokumentierten TEI-Richtlinien in einem Metadatenknoten festgehalten werden, der mit dem Textknoten verbunden ist. Oder die Metadaten werden als Properties direkt im Textknoten abgespeichert.
- Das System macht keine Vorgaben, was ein „Zeichen“ definiert. Diese Definition findet im Projekt im Anwendungszusammenhang statt.
- Festzuhalten bleibt also, dass es lineare Textstücke gibt, die Annotationen haben können und diese Annotationen können wiederum einen Text haben, der Annotationen hat.
- Es ergibt sich ein Netzwerk von Texten.

Das Modell geht davon aus, dass es keine z.B. Diskontinuitäten, keine Nichtlinearitäten oder Löschungen im Text gibt, sondern betrachtet sie als konstitutive Schritte auf dem Weg zu einem lesbaren Text. Die Textphänomene wie z.B. Diskontinuitäten, Löschungen oder Nichtlinearitäten werden in den Annotationen verzeichnet. Im Unterschied zu TEI-XML, wo bei der Verwendung des <choice>-Elements keine Vorgabe für das Lesen des Textes vorgegeben wird, ist bei ATAG die Entscheidung für einen Basistext konstitutiv. Allerdings kann der Nutzer selbst entscheiden, was in den Basistext und was in die Annotationen kommt. Darüber hinaus ermöglicht die Verwendung eines Labeled-Property-Graphen (LPG) die Verwendung von weiteren Properties in den Zeichen- und Annotationsknoten. Damit kann sehr flexibel modelliert werden.

Abbildung 3 zeigt das ATAG-Graphmodell des ersten Wortes des Briefes R2 aus der Handschrift R des Liber Epistolarum der Hildegard von Bingen. Abb. 1 zeigt den gleichen Brief auf der Webseite. Der Grundtext ist normalisiert, die Annotation leiden/expansion zeigt an, dass im Original die Zeichen ‘enio’ fehlen. In Abb. 3 sind auch die weiteren Annotationen zu sehen, die diesen Textbereich betreffen.

Als zweites Beispiel wird eine Textstelle aus einem Brief des Sozinianerprojekts vorgestellt.

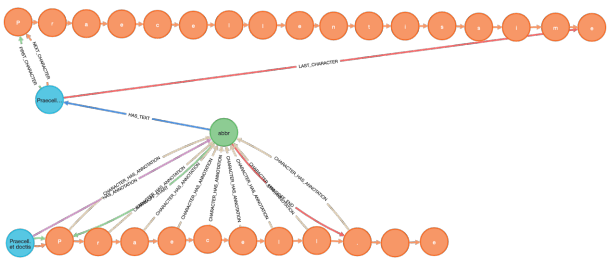


Abb. 4.: Das erste Wort des Briefes von Johannes Müller (Hamburg) an Stanisław Lubieniecki (Hamburg) vom 5. Januar 1665 (https://sozinianer.m-ni.thm.de/view/ed_m3l_ldy _ vkb) abgerufen am 15.7.2023.

Links unten in Abb. 4 befindet sich der blaue Textknoten des Briefes, von dem die Zeichenkette “Praecelli.” startet. Die Zeichen von “P” bis “.” sind von der Annotation *abbr* (grüner Knoten in Abb. 4) umfasst. Vom Annotationsknoten geht eine Kante zum oberen blauen Textknoten, von dem die Zeichenknotenkette mit der normalisierten Fassung des Wortes „Praecellentissime“ startet. Hier wird gut sichtbar, wie sich das Muster zu größeren Strukturen zusammensetzt. Festzuhalten bleibt auch, dass im Sozinianerprojekt im Grundtext die Originalschreibweise festgehalten wird, während die Normalisierung in der Annotation liegt. Im Hildegardprojekt ist hingegen der Grundtext normalisiert und die Ergänzungen annotiert. Diesbezüglich ist das Textmodell agnostisch. Ein Projekt muss allerdings zu Beginn die Entscheidung treffen, wie der Text modelliert werden soll, und sich anschließend daran halten. Prinzipiell sind die verschiedenen Varianten über Graphalgorithmen inneinander überführbar. Daher wird hier auch keine Position in der langen Diskussion eingenommen, was genau ein „Text“ ist, den diese Definition liegt bei den Nutzenden.

Webbasierter Editor für ATAG

Wie bereits oben erwähnt, sind standoff-basierte Textformate schon lange bekannt und werden umfangreich genutzt. Allerdings besteht bei vielen Ansätzen das Problem, dass der Basistext später nicht geändert werden kann, da sich die Indizes sonst verschieben. Mit ATAG wird das Ändern von bereits annotiertem Text möglich.

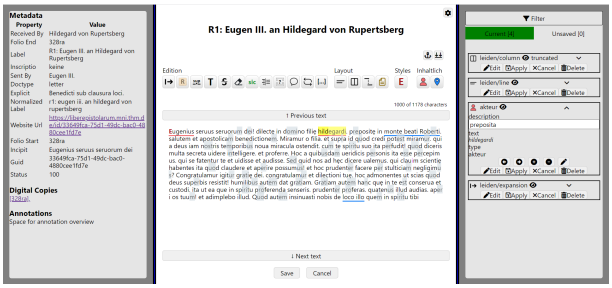


Abb. 5.: Screenshot des ATAG-Editors, der gerade im Rahmen einer Masterarbeit entwickelt wird.

Momentan entsteht im Rahmen einer Masterarbeit ein webbasierter Editor für ATAG. Die Herausforderungen liegen hier vor allem in der Visualisierung der zahlreichen Annotationsebenen und -möglichkeiten. Es müssen sinnvolle Nutzungsszenarien entwickelt werden, die auf die jeweiligen Bedürfnisse der Nutzer zugeschnitten sind. Das Textmodell von ATAG ist sehr eng verwandt mit SPO, die beiden sind direkt ineinander überführbar. SPO wird in den Publikationsumgebungen des Sozinianer-Projekts und des Hildegard-Projekts schon produktiv eingesetzt. Mit einem Algorithmus kann aus SPO-Texten ATAG erstellt werden. ATAG bildet dann die Grundlage für das Edieren oder das Auswerten der annotierten Texte. Sind die Änderungen im Graph gespeichert, können die SPO-Angaben in den Text- und Annotationsknoten neu berechnet werden. Damit können die Änderungen auch direkt in die Publikationsumgebung übernommen werden. Festzuhalten ist, dass die Wahl von Zeichenknoten als kleinste Einheit dem leichteren Management im webbasierten Editor geschuldet ist. Abbildung 4 zu „Praecell.“ zeigt die direkt Verbindung zwischen Zeichenknoten und Annotationsknoten über HAS_ANNOTATION-Kanten. Damit ist für jedes Zeichen unmittelbar feststellbar, mit welchen Annotationen es verbunden ist. Diese Information erleichtert die Darstellung im webbasierten Editor. Es lassen sich ausgehend von den Zeichenknoten aber jederzeit Annotationen erstellen, die dann die Tokenebene abbilden und weitere Zusatzinformationen, wie z.B. das Lemma des Tokens, enthalten können. Grundlage ist aber immer der Grundtext mit den Zeichenknoten.

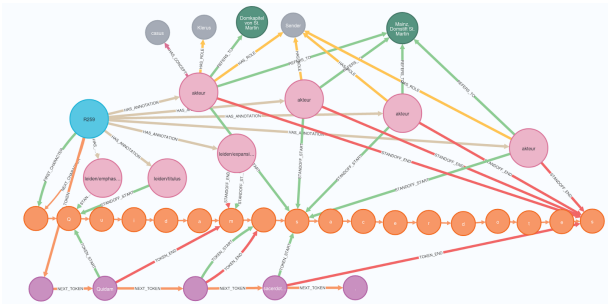


Abb. 6.: Ein Textstück aus dem Brief 259 des Manuskripts R des Liber Epistolarum der Hildegard von Bingen mit Zeichen- (orange) und Tokenebene (fliederfarben).

Abschließend wird in Abbildung 5 der Brief R259 mit einer sehr stark annotierten Stelle, gezeigt bei dem der Zeichenkette “sacerdotes” neben den verschiedenen Layoutannotationen noch vier Akteursannotationen zugeordnet werden, ohne dass das Modell an Klarheit verliert. Die fliederfarbenen Knoten am unteren Rand des Bildes zeigen die Tokenknoten, die jeweils mit der Zeichenebene verbunden sind. Hier wird deutlich, wie das Modell unerschiedliche Granularitätsstufen gemeinsam modellieren kann.

Für einen Test zur Performance wurden die über 200.000 Volltextregesten der Regesta Imperii Online in Neo4j importiert und in ATAG umgewandelt. Der Prozess dauerte auf einer Standardinstallation von Neo4j Desktop ca. 100 Minuten.

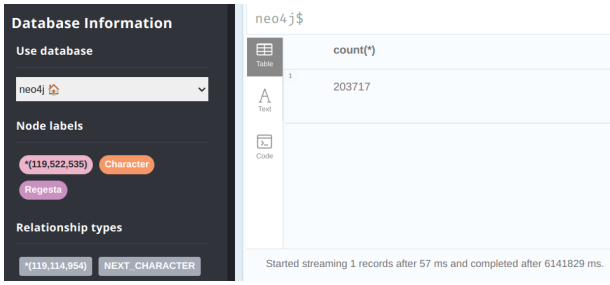


Abb. 7.: Ergebnis des Imports der 203717 Regesten der Regesta Imperii in ATAG.

Nach dem Importprozess hatte die Datenbank knapp 120 Millionen Knoten und knapp 120 Millionen Kanten. Auf die Geschwindigkeit der Queries hatte dies jedoch keine Auswirkungen.

Zusammenfassung

Mit ATAG, oder Applied Text as Graph können Text und Annotationen strukturiert modelliert werden. Im Kern basiert ATAG auf drei Hauptelementen: einem Textknoten, der den Beginn eines Textblocks kennzeichnet, einer Kette von individuellen Zeichenknoten, die den Text ausmachen, und Annotationen, die mit einem anderen Textknoten verknüpft werden können. Eine der flexiblen Eigenschaften von ATAG besteht darin, dass es zwar eine grundlegende Zeichenkette erfordert, jedoch keine Beschränkungen dafür auferlegt, was diese Kette enthalten sollte. Bemerkenswert ist, dass die Basiskette von Zeichen in ATAG dynamisch ist. Sie kann auch nach der Annotation bearbeitet oder modifiziert werden. Dies ist besonders nützlich in kollaborativen Umgebungen oder wenn Informationen aktualisiert werden. Darüber hinaus ist ATAG darauf ausgelegt, eine nahezu unbegrenzte Anzahl paralleler und sich überlappender Annotationen zu verarbeiten, ähnlich den bestehenden Standoff-Systemen, die in der Textanalyse verwendet werden. In Bezug auf zukünftige Arbeiten in ATAG gibt es mehrere Richtungen. Das Framework erforscht die Integration der Standards der Text Encoding Initiative (TEI), um seine Fähigkeiten weiter zu bereichern. Darüber hinaus wird an der Entwicklung eines webbasierten Editors gearbeitet, der die Interaktion mit ATAG benutzerfreundlicher macht. Zusammen mit dem generischen Publikationssystem stehen dann alle für eine graphbasierte digitale Edition notwendigen Komponenten zur Verfügung.

Bibliographie

Bleeker, Elli, Ronald Haentjens Dekker, Bram Buitendijk. 2023. “Texts as Hypergraphs: An Intuitive Representation of Interpretations of Text”, Journal of the Text Encoding Initiative, Issue 14 | April 2021-March 2023, Online since 08 June 2022, connection on 13 March 2023. URL: http://journals.openedition.org/jtei/3919 ; DOI: https://doi.org/10.4000/jtei.3919

Bode, Katherin. 2022. “Doing (Computational) Literary Studies”, New Literary History 53.4-54.1 (2022-23): 531-558.

Kuczera, Andreas. 2022. “TEI Beyond XML – Digital Scholarly Editions as Provenance Knowledge Graphs.” In Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): Graph Technologies in the Humanities - Proceedings 2020, published at http://ceur-ws.org/Vol-3110 , 2022. http://ceur-ws.org/Vol-3110/paper6.pdf .

Neill, Iian, Andreas Kuczera. 2019. “The Codex – an Atlas of Relations.” In Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) text/html Format. DOI: 10.17175/sb004_008.

Pichler, Axel, und Nils Reiter. 2022. “From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities.” Journal of Cultural Analytics, vol. 7, no. 4, Dec. 2022, doi:10.22148/001c.57195.

Schmidt, Desmond. 2016. “Using standoff properties for marking-up historical documents in the humanities.“ In it – Information Technology 58 (2016), H. 2, S. 63–69. DOI: 10.1515/itit-2015-0030

Automatisierung und CI/CD in digitalen Editionen am Beispiel des Conciliator

Schrader, Oliver
oliver.schrader@uni-koeln.de
Cologne Center for eHumanities (CCeH), Universität zu Köln, Deutschland
ORCID: 0000-0003-0481-395X

Gassmann, Sebastian
gassmann@uni-bonn.de
Institute for Medical Humanities, Rheinische Friedrich-Wilhelms-Universität Bonn, Deutschland
ORCID: 0009-0001-9396-8824

Gengnagel, Tessa
tessa.gengnagel@uni-koeln.de
Cologne Center for eHumanities (CCeH), Universität zu Köln, Deutschland
ORCID: 0000-0001-8820-5112

Einordnung

Gemäß dem Manifest für digitale Editionen, das im Anschluss an die letztjährige DHd-Tagung entstanden ist, gehen digitale Editionen „weit über die gewohnten gedruckten Ausgaben hinaus“ (IDE, 2022). Dies wird nicht zuletzt damit begründet, dass „digitale Editionen nicht mit der Publikation beendet sind, sondern als fortlaufender Prozess offener, weiter und umfassender gedacht und geplant werden müssen“ (IDE, 2022).

Unser Beitrag greift diesen Grundgedanken auf. Wir verstehen eine digitale Edition nicht nur als digitale Speicherung und Präsentation der Daten (zur ‚Datenhaftigkeit' digitaler Editionen, s. auch Stäcker, 2020), sondern wesentlich als computergestützte und zunehmend automatisierte Verarbeitung, Aufbereitung und Erzeugung von Daten.

Das Projekt, in dessen Rahmen wir den im Folgenden vorgestellten Editionsprozess erproben und weiterentwickeln, ist die DFG-geförderte Edition des *Conciliator* von Petrus von Abano.[1] Der *Conciliator differentiarum philosophorum et praecipue medicorum* ist das Hauptwerk des Arztes, Astronomen und Naturphilosophen Petrus von Abano, in zweiter Redaktion fertiggestellt im Jahr 1310 (Kaiser und Schenkel, 2019). Das Werk legt in mehrfacher Hinsicht nahe, es digital zu erschließen. Petrus' Umgang mit zahlreichen Zitaten insbesondere aus antiken griechischen und aus arabischen Quellen wird besser ersichtlich, wenn Quellen ausführlich wiedergegeben und, wo es solche gibt, mit deren digitalen Editionen verknüpft werden können (s. zum Paradigma der ‚Rekontextualisierung' Meier und Viehauser, 2020). Auch die Beigabe eines Kommentars, einer Übersetzung und – unter Berücksichtigung der „Zuwendung zum Material" (Mertens, 2021, Abs. 100) im Digitalen – die Verlinkung auf im IIIF-Standard verfügbare Digitalisate der Handschriften und Drucke sind geplant. Zudem umfasst das Projekt vorerst nur zehn der insgesamt 210 Kapitel des Werks, sodass die Voraussetzungen für eine spätere nahtlose Ergänzung bedacht werden sollten. Wesentlich über diese Aspekte hinaus geht jedoch der Umstand, dass die Überlieferung in 13 vollständigen, dazu mehreren unvollständigen Handschriften sowie 13 Druckausgaben von 1472 bis 1565 hohen Kollationsaufwand verursacht, der durch die Verwendung von Software stark reduziert werden kann. Dazu wurde im Projekt ein eigenes Setup von automatisierten Verarbeitungsschritten implementiert, das sich am Ansatz der Continuous Integration/Continuous Delivery orientiert.

Grundgedanken: CI/CD

Die Verknüpfung der genannten Komponenten – Edition, Variantenapparat, Quellenapparat, Übersetzung, Stellenkommentar – zu einer homogenen digitalen Edition erzeugt eine hohe Komplexität: allem voran die Erstellung und Verwaltung der IDs für Verweise der Elemente untereinander ist aufwendig und fehleranfällig und erschwert nachträgliche Änderungen, weil sämtliche Abhängigkeiten korrekt aktualisiert werden müssen. Wenn andererseits die Verknüpfung nur einmalig nach Erledigung der Vorarbeiten erfolgen soll, können Zwischenergebnisse jeweils nur isoliert betrachtet werden; damit wiederum nimmt man sich die Möglichkeit, frühzeitig das entstehende Gesamtbild zu prüfen und für die restliche Arbeit zu nutzen. Ideal wäre also ein Setup, in welchem die unterschiedlichen Bestandteile der Edition unabhängig voneinander entstehen können und zugleich Zwischenergebnisse jederzeit so zusammengefügt werden können, dass ein funktionsfähiges Zwischenprodukt entsteht.

In die digitale Edition des Conciliator haben wir deshalb einige Konzepte aus der modernen Softwareentwicklung übernommen, die unter den Begriff „Continuous Integration, Continuous Delivery" (CI/CD) fallen.[2] Im Detail sind damit viele verschiedene (auch regelmäßig neue) Technologien verbunden, die in unserem Zusammenhang keine Rolle spielen – der Grundgedanke lässt sich aber gewinnbringend auf das Feld der Editionsphilologie übertragen.[3]

- „Continuous Integration" (CI) bezeichnet das Prinzip, die Ergebnisse einzelner Arbeitsschritte laufend in das wachsende Produkt so einzufügen, dass funktionale (nutzbare) Zwischenstände entstehen. Diese Integration soll so weit automatisiert werden, dass dafür kaum bzw. zuletzt gar kein manuelles Zutun mehr nötig ist.
- „Continuous Delivery" (CD) schließt daran an mit dem Gedanken, dass die regelmäßige Veröffentlichung bzw. bei Software das Einspielen und Starten (Deployment) der jeweiligen Zwischenstände ebenso automatisiert wird wie zuvor die Integration. Oft ist es damit nur noch ein Befehl – „ein Klick" – von der veränderten Datengrundlage zum aktualisierten Produkt.

Im Fall der Conciliator-Edition bedeutet das beispielsweise: Auch ohne Quellenapparat und Übersetzung soll aus einem vorläufig konstituierten Text, basierend auf wenigen transkribierten Zeugen, automatisch eine nutzbare Edition erzeugt werden können, die zumindest den Text darstellt und bislang resultierende Varianten im Apparat verzeichnet. Wenn ein Zitat identifiziert wird, wenn eine Teilübersetzung erstellt oder eine Sachanmerkung hinzugefügt wird, soll jedes Mal automatisch eine aktualisierte Edition mit den neuen Inhalten erzeugt werden können, seien sie auch noch so unvollständig. Doch nicht nur bei Erweiterungen, auch bei Änderungen: Wenn Transkriptionen korrigiert oder Änderungen am textus constitutus vorgenommen werden, soll die Edition mit einem angepassten Apparat neu erstellt werden – wiederum automatisch, was insbesondere eine vollständig softwaregestützte Kollation und Apparaterzeugung voraussetzt. Damit sind einige Möglichkeiten von Continuous Integration beschrieben, um deren Verwirklichung es im Folgenden gehen soll.

Continuous Delivery hieße darüber hinaus im Wesentlichen, die neueste Version der Edition jeweils unmittelbar online zugänglich zu machen; das ist jedoch im momentanen Stadium unseres Projekts noch nicht relevant und wird deswegen hier nicht vertieft. Ein einfaches – insbesondere einfach zu nutzendes – Konzept hierfür bietet z.B. GitHub Pages unter dem Motto: „Just edit, push, and your changes are live."[4]

Technische Details

Um die beschriebenen Vorteile ausnutzen zu können, müssen wir einige Voraussetzungen erfüllen:

Grundvoraussetzung ist eine vollständig digitale Datenhaltung. Wir bevorzugen dabei einfache, in der Programmierung etablierte Dateiformate wie XML, JSON und YAML. XML kommt tendenziell dort zum Einsatz, wo Daten wenig verarbeitet werden müssen (z.B. Beschreibungen von Textzeugen im TEI/XML-Standard) oder bereits auf die spätere Darstellung hin ausgezeichnet werden (z.B. Editionstext und Stellenkommentar mit den in HTML verallgemeinerten typographischen Möglichkeiten); JSON und YAML sind die Dateiformate für Transkriptionen einzelner Textzeugen, für Konfigurationsdateien und für tendenziell typographisch anspruchslose bzw. noch stark weiterverarbeitete Informationen.

Zur Versionskontrolle verwenden wir Git (https://git-scm.com/) mit einer vom Cologne Center for eHumanities (CCeH) verwalteten GitLab-Instanz als zentralem Repository. Über die Git History ist implizit bereits eine Möglichkeit zur stabilen Versionierung der Edition gegeben, sollte später keine andere Lösung vorteilhaft sein.

Eine weitere Voraussetzung folgt aus der Idee von CI und bildet den Kern unserer Neuerung: Die gesamte Edition – die Repräsentation – muss aus den jeweils gegebenen Daten vollständig abgeleitet werden können. Dabei sollen die nötigen Verweise sowie alle prinzipiell ableitbaren Informationen automatisch ergänzt werden, denn gerade deren dauernde manuelle Pflege würde einen erheblichen Mehraufwand bedeuten.

Der Build-Prozess für die Edition besteht aus wesentlich zwei Schritten:

1. Die in verschiedenen Formaten abgelegten Daten werden von einem Programm[5] eingelesen und in mehreren Teilschritten verarbeitet: Zunächst wird jedem Wort aus dem Editionstext eine ID zugeordnet, um eine punktgenaue Referenzierung zu ermöglichen; daraufhin werden die Textzeugen gemäß einem Satz von vorgegebenen Regeln normalisiert, dann aus Editionstext und normalisierten Textzeugen der Variantenapparat berechnet und das Ergebnis TEI-konform als Dokument mit Standoff-Markup (d.h. mit „variantEncoding“ nach der Double-end-point-Methode, external) strukturiert (zum Thema Standoff-Markup in digitalen Editionen s. Spadini und Turska, 2019; Klug, 2021). Angereichert um Standoff für Quellen, Übersetzungen, Anmerkungen nach demselben Prinzip, wird alles zusammen in einer einzigen Datei abgelegt, die damit die gesamte Edition i.S.d. Information enthält. Das Datei-Format ist momentan ein TEI-ähnliches XML (d.h. semantisch an TEI orientiert, aber strukturell unter Aufbrechung der Hierarchie-Anforderungen von XML als reines Standoff gestaltet), kann sich aber z.B. noch zu GraphML (oder einem Äquivalent) ändern, um auch im Zwischenprodukt einem Standard zu folgen, der eine einfache Verarbeitung in anderen technischen Umgebungen erlaubt.
2. Im zweiten Schritt erfolgt eine XSL-Transformation, die die Editionsdaten aus dem ersten Schritt in eine Standalone-Website im HTML-Format überführt und damit den CI-Prozess abschließt. Der gesamte Prozess dauert nur wenige Sekunden bis Minuten (abhängig vom Textumfang und der Zahl der Zeugen) und kann beliebig häufig angestoßen werden (z.B. automatisiert bei jedem Commit/Merge in GitLab). Aus Gründen der angestrebten Einfachheit und Langzeitverfügbarkeit arbeiten wir momentan darauf hin, selbstgenügsame HTML-Dateien zu erzeugen, wie sie sich seit den ersten Tagen des Internet als überlebensfähig erwiesen haben. Die Integration von JavaScript-Code schafft Raum für die Implementierung auch komplexer Funktionen für die Suche, Konfiguration (des angezeigten Inhalts, der Darstellungform), punktgenaue Zitierung, etc. – eine *static website* muss keineswegs „statisch“ erscheinen, nur verzichten wir auf eine dynamische Generierung der Website im Backend zum Zeitpunkt eines Abrufs. Das Ausgabeformat ist aber generell sekundär und es wäre möglich, weitere Formate anzubieten, über die unsere Edition z.B. als Applied Text as Graph (ATAG: https://git.thm.de/aksz15/atag) genutzt werden kann.

Eine Prämisse, die aus der automatischen Verarbeitung folgt, ist der vollständige Verzicht auf nachträgliche manuelle Eingriffe in Berechnungsergebnisse. Alle Informationen und editorischen Entscheidungen müssen explizit in den Ausgangsdaten enthalten sein oder daraus berechnet werden – bei gleichzeitiger größtmöglicher Unabhängigkeit der einzelnen Komponenten voneinander (z.B. folgt die Transkription so genau wie möglich der originalen Schreibweise, während die Kollation und der Apparat basierend auf normalisierten Schreibweisen erzeugt werden). Derartig voneinander getrennt bearbeiten wir den Editionstext, die Textzeugen (jeweils in einzelnen Dateien transkribiert), Quellennachweise, Übersetzung und Anmerkungen.

Dabei ist es (noch) nicht möglich, diese Bestandteile gänzlich unabhängig, d.h. ohne jegliche manuelle Verknüpfung, zu behandeln, denn jeder Paratext, der auf eine bestimmte Stelle des Textes bezogen ist, setzt eine Referenzierung voraus. An den folgenden Beispielen lassen sich unterschiedliche Strategien beschreiben, die das bisher erreichte Ausmaß an Automatisierung im Build-Prozess illustrieren:

- **Kollation** und **kritischer Apparat** können mittlerweile dank einer selbstentwickelten Software vollständig automatisiert erstellt werden. Es existiert bereits Software, die wie ChrysoCollate (https://cental.uclouvain.be/chrysocollate/) oder Collation Editor (https://github.com/itsee-birmingham/collation_editor_core) die manuelle Kollation erheblich erleichtert oder wie CollateX (https://collatex.net/) sogar großteils automatisiert, auch wenn sie bisher in Editionsprojekten nicht flächendeckend eingesetzt wird – zumal einige Ansätze wie Juxta mittlerweile obsolet sind (für einen Einsatz von CollateX s. Hulle, 2019; für eine allgemeinere Problematisierung Nury, 2019). Keines der bestehenden Tools bietet jedoch bislang den Funktionsumfang und

das Maß an Automatisierung, um sich in unseren CI/CD-Ansatz gut integrieren zu lassen.

- **XML-IDs**, die der internen Verknüpfung dienen, können deterministisch durch einen Algorithmus – z.B. wesentlich durch Durchzählen von Elementen – automatisch erzeugt werden. Anfangs manuell vergebene Präfixe für größere Abschnitte und den jeweiligen Zeugen schaffen eine gewisse Ordnung, wären aber entbehrlich.
- **Erweiterungen** um neue Zeugen oder neue Textabschnitte erfordern nur die einmalige Aktualisierung einer Konfigurationsdatei um die entsprechenden Dateipfade, die der Software vorgeben, welche Dateien zu verarbeiten sind.
- **Normalisierung** und Ausscheiden unwichtiger Varianten beruhen auf einem Satz von Ersetzungsregeln, der laufend ergänzt wird. Jede Normalisierungsregel (z.B. basal: „Lies jedes *v* als *u*.“; komplexer: „Lies *ncium* am Wortende als *ntium*, jedoch nicht bei *lancium* und *teruncium*.“) wird einmalig definiert und danach – wie eine *regular expression* – automatisch auf alle passenden Kontexte angewandt (unbenommen der Möglichkeit der Definition von Ausnahmen in den Regeln oder auch punktuellen „Überschreibens“ in den Ausgangsdaten). Nachvollziehbarkeit der erfolgten Ersetzungen bietet eine Log-Datei; zusätzliche Sicherheit wird durch eine im Projektverlauf wachsende „Allow-List“ geschaffen, die nur bestätigte Ersetzungen zulässt und weitere sich aus den Regeln ergebende zur Prüfung durch die Herausgeber vorlegt (und dann ggf. in die Liste aufnimmt). Hier ist also ebenfalls eine fast vollständige Automatisierung erreicht.
- **Standoff-Markup** (Quellen, Übersetzungen und Anmerkungen) weisen bislang den geringsten Grad an Automatisierung auf, weil die Referenzierung auf den Editionstext manuell eingerichtet werden muss. Da die XML-IDs mit jedem Build-Prozess neu erzeugt werden und sich bei geändertem Editionstext ändern können (obwohl sie in gewissen Grenzen stabil sind und die Links deshalb meist nicht durch Änderungen an entfernten Stellen im Editionstext beeinflusst werden), werden sie im Build-Prozess validiert; im Fehlerfall erscheint eine Warnung und die Referenzierung muss manuell aktualisiert werden.

Wichtig zu bemerken ist der Umstand, dass hierbei niemals eine bestehende XML- bzw. HTML-Datei verändert, sondern jedes Mal alles neu berechnet und erzeugt wird. Deswegen verzichten wir auf spätere manuelle Anpassungen – sie würden einfach überschrieben werden. Stattdessen müssen spätere Entscheidungen, wie beschrieben, vorweggenommen werden. Gerade das erlaubt, je allgemeiner Regeln formuliert werden können, eine umso stärkere Reduzierung der expliziten Abhängigkeiten zu Gunsten eines einfacheren und flexibleren Arbeitsprozesses. Konkrete Verknüpfungen werden so weit wie möglich durch logische Abhängigkeiten ersetzt.

Der beschriebene Build-Prozess kann von allen Projektmitgliedern auf dem eigenen Computer ausgeführt werden und erweist sich in der täglichen Arbeit als äußerst hilfreich. Zudem ist er in eine GitLab-Pipeline integriert und erzeugt nach jedem Push die jeweiligen Artefakte (oder schlägt fehl – und dient damit als Qualitätskontrolle). Die größten Vorteile einer so weitreichenden Automatisierung sind:

1. **Übersicht**: Die Edition ist selbst die beste Darstellung der Zwischenergebnisse. Je früher Teile der Edition auf die vorgesehene Art repräsentiert werden können, desto besser – beispielsweise für die Arbeit an Übersetzung und Kommentierung, aber auch für eine stetige Verbesserung der Repräsentation selbst.
2. **Flexibilität**: Abhängigkeiten zwischen einzelnen Arbeitsschritten werden reduziert. Eigentlich ‚frühere‘ Arbeitsschritte wie Textkonstitution und Festlegung der Normalisierungsregeln können mit eigentlich ‚späteren‘ wie Kollation und Erstellung des Apparats abwechseln, eine Revision ist jederzeit möglich.
3. **Zeitersparnis**: Am leichtesten sind Arbeitsschritte automatisierbar, die von konkreten Regeln abhängen. Derartige Routineaufgaben, wie sie etwa bei der Normalisierung divergierender lateinischer Schreibweisen entstehen, fallen oft in sehr großer Zahl an und sind entsprechend zeitaufwendig. Computer erledigen das in Sekunden(bruchteilen).
4. **Qualität**: Die beschriebene Flexibilität gilt auch für Fehlerkorrekturen, bei denen die Notwendigkeit entfällt, die teils weitreichenden Auswirkungen von Fehlern manuell zu korrigieren: Ist der Fehler selbst verbessert, erledigt sich alles weitere von selbst im nächsten Build-Prozess. Korrekte Algorithmen vorausgesetzt, entstehen auch erst keine Fehler bei automatisierten Teilschritten.

Schlussbemerkungen

Der *Conciliator* des Petrus von Abano ist aufgrund seiner scholastischen, einheitlichen Strukturierung besonders gut für den vorgestellten Workflow geeignet (die *differentiae* und ihre Gliederungspunkte bilden natürliche kleine *increments* und die Menge der Textzeugen behindert nicht einen frühen Beginn der Textkonstitution und Kollation). Dasselbe gilt für alles, was kleine, abstrahierbare Einheiten enthält (wie Briefcorpora, die Gegenstand vieler digitaler Editionsprojekte sind, s. Dumont, 2023) – aber letztlich ist die Textgrundlage nachrangig, denn man kann auch willkürlich Grenzen für Zwischenergebnisse setzen.

Unser Prozess richtet sich vielmehr aus an grundlegenden Einsichten:

- **Zwischenergebnisse** sind wertvoll für die weitere Arbeit und sollten deswegen möglichst früh gut darstellbar sein.
- Die **Komplexität** digitaler Editionen kann viel Aufwand für eigentlich rein formales Beiwerk bedeuten

und sollte deswegen weitgehend reduziert werden, um die inhaltliche Arbeit nicht zu belasten.
- Viele Einzelschritte sind formalisierbar, eine passende Materialgrundlage vorausgesetzt, und eignen sich damit für weitgehende **Automatisierung**.
- Fehler passieren: Es ist sinnvoll, typische **Fehlerquellen** mit Hilfe von Software auszuschalten, und es ist wichtig, Fehler unkompliziert korrigieren und dokumentieren zu können.

Unser Beitrag wird diesen Arbeitsprozess und die bereits vorliegenden Ergebnisse vorstellen sowie weitere Erfahrungen aus der laufenden Arbeit berücksichtigen. Andere prozessorientierte Ansätze wie bei Cugliana und Barabucci 2021 haben in jüngerer Vergangenheit ebenfalls auf das Potential automatisierter Prozesse in digitalen Editionsvorhaben hingewiesen und können als Vergleichsfolie dienen. Trotzdem ist insgesamt festzuhalten, dass das Potential einer Anwendung von Prinzipien aus dem Software Development auf digitale Editionen bei weitem noch nicht ausgeschöpft ist und sich insbesondere angesichts des Konferenzthemas die Frage stellt, ob und inwiefern es sich hierbei um ein bestimmendes Paradigma des kommenden Jahrzehnts handeln wird. Aufgrund unserer guten Erfahrungen möchten wir die Position vertreten, dass die Umsetzbarkeit solcher Unterfangen sich vor allen Dingen aus einem Verständnis des zu edierenden Materials sowie einer Kenntnis außerakademischer Arbeitsprozesse speist und bei sorgfältiger Anwendung die beschriebenen Vorteile mit sich bringen kann.

Fußnoten

1. DFG-Projekt „Die philosophischen Grundlagen der Medizin – Digitale kritische Edition des Conciliator (Differentiae 1 bis 10) des Petrus von Abano“, Laufzeit 02/2022–01/2025, Projektleitung Dr. Christian Kaiser (Institute for Medical Humanities, Rheinische Friedrich-Wilhelms-Universität Bonn), (abgerufen am 06.12.2023).
2. Für ausführlichere Informationen s. z.B. (abgerufen am 06.12.2023).
3. Der von uns vorgestellte Ansatz beansprucht keine prinzipielle Neuheit für den Transferansatz als solchen, der z.B. schon von Safaryan, Andrews und Atayan 2019 formuliert worden ist und im Projekt „The Chronicle of Matthew of Edessa“ (, abgerufen am 06.12.2023) verfolgt wird. Anders als dort, wo mehrere automatische Verarbeitungsschritte in einer Pipeline ausgeführt werden, anschließend jedoch auch wieder manuelle Eingriffe erfolgen, verfolgen wir das Ziel einer vollständigen Automatisierung der Ableitung der Edition (nicht nur der Darstellung, sondern auch der Varianten unterschiedlicher Zeugen, d.h. die Edition konstituierender Daten) aus den Anfangsdaten. Darin liegt überdies eine Abgrenzung zu virtuellen Forschungs- und Editionsumgebungen wie TextGridLab, die die technische Umsetzung prinzipiell ihrer Anlage nach unterstützen, aber in sich keine editionskonstituierenden (‘inhaltlichen’) Aufgaben übernehmen.
4. (abgerufen am 06.12.2023).
5. Im Rahmen des Projekts entsteht eine in der Programmiersprache Haskell geschriebene Software mit dem Arbeitstitel „Hkoll“, die außer der ursprünglich namensgebenden Aufgabe der automatischen Kollation inzwischen weitere Teilaufgaben ausführt und integriert.

Bibliographie

Cugliana, Elisa und Gioele Barabucci. 2021. „Signs of the Times: Medieval Punctuation, Diplomatic Encoding and Rendition.“ *Journal of the Text Encoding Initiative* 14. https://doi.org/10.4000/jtei.3715.

Dumont, Stefan. 2023. „Briefeditionen vernetzen.“ In *Digitale Literaturwissenschaft: Germanistische Symposien*, hg. von Fotis Jannidis, 729–750. Stuttgart: Metzler.

IDE. 2022. „Manifest für digitale Editionen.“ https://dhd-blog.org/?p=17563 (abgerufen am 06.12.2023).

Kaiser, Christian und Peter Schenkel. 2019. „Pietro d’Abano über die Bedeutung der theoretischen Wissenschaften für den Arzt (mit einer kritischen Edition und Übersetzung der Differentia prima des Conciliator)“ In *Die nackte Wahrheit und ihre Schleier. Weisheit und Philosophie in Mittelalter und Früher Neuzeit - Studien zum Gedenken an Thomas Ricklin* (Dokimion, Bd. 42), hg. von Christian Kaiser, Leo Maier und Oliver Maximilian Schrader, 191–247. Münster: Aschendorff.

Klug, Helmut W. 2021. „Stand-off-Markup.“ In *KONDE Weißbuch*, hg. von Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt ‚Kompetenznetzwerk Digitale Edition‘. Handle: https://hdl.handle.net/11471/562.50.171. PID: o:konde.171 (abgerufen am 06.12.2023).

Meier, Simon und Gabriel Viehauser. 2020. „Rekontextualisierung als Forschungsparadigma des Digitalen?“ In *Rekontextualisierung als Forschungsparadigma des Digitalen* (Schriften des Instituts für Dokumentologie und Editorik, Bd. 14), hg. von Simon Meier, Gabriel Viehhauser und Patrick Sahle, 1–20. Norderstedt: Books on Demand.

Mertens, Ina. 2021. „Zwei Seiten einer Medaille – IIIF und die Arbeit mit digitalen Bildbeständen.“ *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2021_002.

Nury, Elisa. 2019. „Towards a Model of (Variant) Readings.“ In *Versioning Cultural Objects: Digital Approaches* (Schriften des Instituts für Dokumentologie und Editorik, Bd. 12), hg. von Roman Bleier und Sean Winslow, 3–23. Norderstedt: Books on Demand.

Safaryan, Anahit, Andrews, Tara L. und Tatevik Atayan. 2019. „Continuous Integration Systems for Critical Edition: The Chronicle of Matthew of Edessa.“ Paper, *Digital Humanities 2019*, Utrecht, Niederlande. https://doi.org/10.5281/zenodo.5498352.

Spadini, Elena und Magdalena Turska. 2019. „XML-TEI Stand-off Markup: One Step Beyond.“ *Digital Philology: A Journal of Medieval Cultures* 8/2: 225–239. http://doi.org/10.1353/dph.2019.0025.

Stäcker, Thomas. 2020. „›A Digital Edition is not Visible‹ – Some Thoughts on the Nature and Persistence of Digital Editions.“ *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2020_005.

Van Hulle, Dirk. 2019. „Digitizing Beckett.“ In *The New Samuel Beckett Studies*, hg. von Jean-Michel Rabaté, 19–35. Cambridge: Cambridge University Press.

Über die Ordnung von materiellen und digitalen Dingen: Zur multi-klassifikatorischen Visualisierung der Bibliotheca Eugeniana

Windhager, Florian
florian.windhager@donau-uni.ac.at
Universität für Weiterbildung Krems, Österreich
ORCID: 0000-0002-5170-2243

Tartler, Annerose
annerose.tartler@onb.ac.at
Österreichische Nationalbibliothek, Österreich

Mayer, Simon
simon.mayer@onb.ac.at
Österreichische Nationalbibliothek, Österreich

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

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

Hintergrund & Motivation

Prinz Eugen von Savoyen (1663–1736) ist als militärischer Stratege und Feldherr des Habsburgerreiches vielen Zeitgenoss:innen bis heute ein Begriff. Neben seiner militärischen Tätigkeit trat er auch als Sammler und Mäzen der Künste in Erscheinung, der eine der bedeutendsten barocken Büchersammlungen aufbaute, die “Bibliotheca Eugeniana”. Der Wert dieser Sammlung mit ihren rund 15.000 Büchern wurde nach dem Tode des Prinzen höher als dessen Wiener Schloss Belvedere veranschlagt, und von dessen Erbin für eine Leibrente von 10.000 Gulden an die österreichische Hofbibliothek verkauft. Bis heute ist sie als zentraler Bestand im Prunksaal der österreichischen Nationalbibliothek (ÖNB) zu finden und wurde als bedeutendes Zeugnis des Wissensstandes ihrer Zeit 2014 zum UNESCO Weltkulturerbe erklärt.

Trotz der historischen Bedeutung dieser Sammlung steht die bibliotheks- und kulturwissenschaftliche Analyse und Konsolidierung ihrer Bestände noch aus: Wie strukturierten die ursprünglichen Kataloge und Klassifikationen diesen Bestand — und welche Entsprechungen finden diese im modernen Katalog der Nationalbibliothek? Wie groß ist mit Blick auf die mitunter abenteuerliche Erhaltungsgeschichte der Sammlung ihr aktueller Umfang? Welche Bücher zählen zur genuin-historischen Sammlung, welche Exemplare wurden im Laufe der Zeit anderen Sammlungen und Aufstellungen zugeordnet, gingen verloren oder wurden fälschlich in den Bestand der Eugeniana integriert? Können die Farben und Wappen der historischen Ledereinbände die präzise Zuordnung unterstützen und lassen sie einen Rückschluss auf die fachliche Einordnung der Bücher zu (vgl. Mauthe, 2010, S. 194)? Neben ihrer Bedeutung für Wissenschaft und Forschung, ist darüber hinaus die öffentliche Bekanntheit dieses historischen Weltkulturerbebestandes in Österreich sehr gering.

Das Projekt Bibliotheca Eugeniana Digital (https://labs.onb.ac.at/bed/) hat neben der Beantwortung dieser wissenschaftlichen Fragen die digitale Rekonstruktion und Visualisierung von Prinz Eugen von Savoyens Büchersammlung zum Ziel. Kombinierte Methoden aus den digitalen Geisteswissenschaften und Data Sciences werden dazu genutzt, die Zusammensetzung und Geschichte der barocken Sammlung zu rekonstruieren, diese für die Wissenschaft zugänglich zu machen und sie in geeigneter Form der Öffentlichkeit zu kommunizieren. Die Materialgrundlage hierfür liefern die digitalisierten historischen Buchbestände der Österreichischen Nationalbibliothek und deren Erfassung in einem handschriftlichen Katalog sowie im modernen Bibliothekskatalog der ÖNB.

Datenaufbereitung und Rekonstruktion von multiplen historischen Wissensordnungen

Eine besondere Herausforderung für die Visualisierung der barocken Büchersammlung ergibt sich aus der parallelen Präsenz von multiplen Wissensordnungen: Diese sind 1) durch einen handschriftlichen historischen Katalog gegeben, 2) durch eine materielle Klassifikation über farb-

lich differenzierte Supralibros-Einbände, sowie 3) durch die (partielle) Re-Klassifikation der Bibliotheca Eugeniana-Bücher im modernen Katalog der österreichischen Nationalbibliothek.

Analyse des historischen Katalogs

Mittels automatisierter Handschriftenerkennung wird im Projekt der thematisch geordnete Teil des handschriftlichen Sammlungskatalogs der Bibliotheca Eugeniana in eine digitale Edition transformiert. Zu diesem Zweck wird über die HTR-Plattform Transkribus ein eigenes Modell auf die im Katalog vorkommende Handschrift trainiert und das so erzeugte Modell auf den gesamten Katalog angewendet. Um die Fehlerhäufigkeit insbesondere für vereinzelt vorkommende abweichende Schreiberhände zu minimieren, werden die Bände in Folge manuell nachkorrigiert.

Im nächsten Schritt werden mittels verschiedener Suchheuristiken die einzelnen Bucheinträge des historischen Katalogs im modernen Katalog der ÖNB gesucht. Als Ergebnis dieser Vergleiche lässt sich ein Überblick über sich noch im Bestand der ÖNB befindliche oder ausgeschiedene bzw. verkaufte Bücher aus der Sammlung Prinz Eugens gewinnen. Zusätzlich können in der digitalen Edition Querverweise zum modernen Bibliothekskatalog der ÖNB und den digitalisierten Büchern hergestellt werden.

Analyse der historischen Bucheinbände

Ein großer Teil der Bände in der Bibliotheca Eugeniana trägt drei verschiedene Varianten des Wappens von Prinz Eugen auf dem Ledereinband. Mit Hilfe von Machine Learning werden diese Supralibros im gesamten digitalisierten historischen Buchbestand der ÖNB gesucht und datengetrieben der Sammlung zugeordnet (basierend auf Krickl et al., 2022). Dies soll ermöglichen, die verschiedenen Arten von Wappen sowie verschiedene Einbandfarben den (vermuteten) Wissensklassen zuzuordnen. Einschränkend ist anzumerken, dass sich auch Einbände ohne das typische Wappen des Prinzen in der Sammlung befinden und besonders wertvolle Handschriften nicht digitalisiert wurden, die über diesen Ansatz daher nicht zu erfassen sind.

Analyse des modernen Katalogs

Zur Kennzeichnung der Bücher aus der Bibliotheca Eugeniana im Gesamtbestand der ÖNB wurden Signaturen, die mit “BE” beginnen, eingesetzt und diese entsprechend dieser Signatur im Prunksaal aufgestellt. Die Korrekturbedürftigkeit dieser Zuordnung von Signaturen im modernen Katalog erschließt sich durch den Blick auf eine erste Visualisierung aller Bücher mit BE-Signatur (Abb. 1). Die Daten zeigen, dass BE-Signaturen nicht exklusiv für die Bibliotheca Eugeniana genutzt wurden, sondern im Gegenteil mehr als 9000 Bücher mit einem Erscheinungsdatum nach dem Tod von Prinz Eugen eine BE-Signatur tragen, was auf eine Quote von mindestens 45% an unkonventionellen Anwendungen der Signatur hindeutet. Im Gegenzug ordnet der moderne Katalog vielen historischen Objekten Schlagworte der modernen Klassifikation zu, die für die Entwicklung von Distant Reading-Perspektiven als hoch relevant erscheinen.

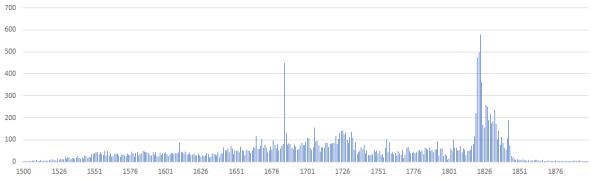


Abbildung 1: Anzahl der Einträge mit BE-Signatur im modernen Bibliothekskatalog der ÖNB.

Visualisierung von materiellen und digitalen Wissensordnungen

Im Sinne eines multimodalen Forschungsansatzes werden mit Blick auf die Bestands- und Klassifikationsdaten multiple Visualisierungen entwickelt, um die Zusammensetzung der Bibliotheca Eugeniana zu analysieren und die Erkenntnisse des Projektes Bibliotheca Eugeniana Digital auch an eine breitere Öffentlichkeit zu kommunizieren. Einerseits sollen damit Forschungsfragen zu Umfang und Aufbau der Sammlung erkundet werden (vgl. Windhager et al., 2018 für einen Überblick zu Techniken der Sammlungsvisualisierung). Andererseits soll mit Techniken des visualisierungsgestützten Storytellings (vgl. Kusnick et al., 2021) ein erhöhtes Bewusstsein für die Existenz und Relevanz der Sammlung in der breiten Öffentlichkeit geschaffen werden.

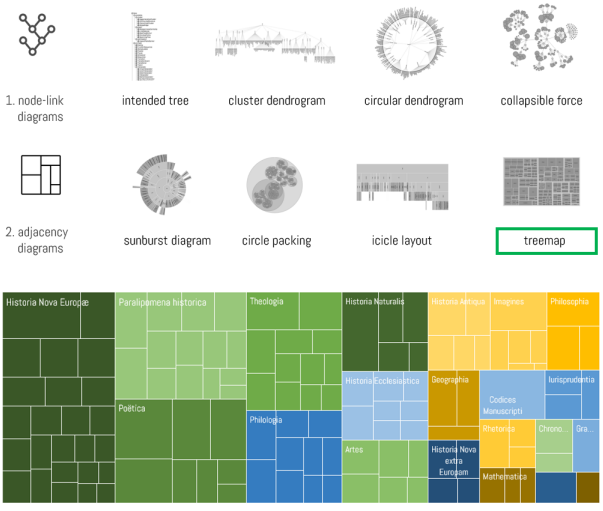


Abbildung 2: Optionen der Visualisierung hierarchischer Wissensklassifikationen (oben) und Treemap-Visualisierung der historischen Klassifikation der Bibliotheca Eugeniana basierend auf dem handschriftlichen Katalog.

Zur Visualisierung von Taxonomien stehen verschiedene Methoden mit divergierenden Profilen zur Verfügung (Schulz, 2011). Abbildung 2 (oben) listet eine Auswahl dieser Optionen, die aktuell für die Visualisierung der Wissensordnungen der Bibliotheca Eugeniana evaluiert werden. Der untere Teil der Abbildung 2 zeigt eine mögliche Repräsentation der zwei Ebenen der historischen Taxonomie aus dem handschriftlichen Katalog mit einer Treemap. Eine Herausforderung mit Implikationen für viele weitere DH-Datensätze von historischer Komplexität stellt in der Folge die visuelle Repräsentation multipler Klassifikationen und von klassifikatorischer Unsicherheit dar, die zu den Zielen des Projekts zählen. Während bereits die Gestaltung von produktiven Sammlungsvisualisierungen komplexe Prozesse des Co-Designs von Expert:innen erfordert (Dörk et al., 2020), so multiplizieren klassifikatorische Pluralität (Bornhofen & Düring, 2020) und Datenunsicherheit (Windhager et al., 2019) die Komplexität von erforderlichen Designentscheidungen. In diesem Kontext wird eine multi-perspektivische Designstrategie sowohl die Erzeugung von multiplen Ansichten auf den Korpus der Bibliothek informieren (Multiplikation von analytischen Perspektiven und Einstiegspunkten für Forscher:innen, sowie Auskopplung einer narrativen Visualisierung zu Vermittlungszwecken), wie auch die Visualisierung von multiplen Wissensordnungen und von typologischer Unsicherheit.

Ein wesentlicher Zugang zur Bibliotheca Eugeniana für die Öffentlichkeit ist aktuell der Prunksaal, in dessen Mitteloval sich bis heute ein Großteil der Sammlung befindet. Die materielle Anordnung der Bücher soll daher auch ein wesentliches Gestaltungselement für einen explorativen, visuellen Zugang zur digitalen Bibliotheca Eugeniana sein. Abbildung 3 präsentiert eine konzeptuelle Skizze des geplanten Interfaces, welches über koordinierte Ansichten neben der materiellen Aufstellung der Bibliothek im Prunksaal (links) auch chronologische, geographische und multiple klassifikatorische Zusammenhänge erkundbar machen wird. Ausgehend von der Zuordnung der Bücher zu den Regalen, können so verschiedene Charakteristika der Sammlung (etwa die historische oder moderne Klassifikation, die Farbe der Bucheinbände, oder die Arten der Supralibros) exploriert werden. Zum Zeitpunkt der Präsentation wird ein ausführlicher Designentwurf dieses Interfaces bereits eine Diskussion der gewählten Designstrategien erlauben. Dieses Interface soll sowohl für eine Erkundung der Sammlung im digitalen Raum zur Verfügung stehen als auch den Besuch des Prunksaals durch vertiefende Einsichten in diese bedeutende Büchersammlung in den digitalen Raum erweitern.

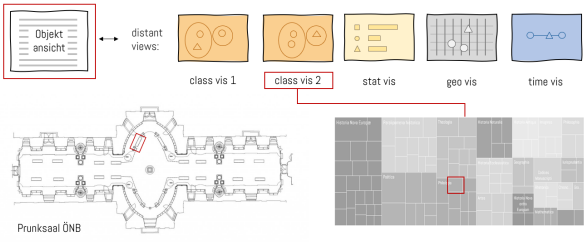


Abbildung 3: Konzept für ein multi-perspektivisches Sammlungs-Interface, welches einen Blick auf die materielle Anordnung der Bibliotheca Eugeniana im Prunksaal der ÖNB mit diagrammatischen Perspektiven auf die Sammlung über koordinierte Ansichten verbindet.

Diskussion

Als historische Sammlung mit einer komplexen Zusammensetzung und Geschichte vereint die Bibliotheca Eugeniana viele Faktoren und Herausforderungen, die auch den Einsatz von komplexe(re)n Designprozessen bei ihrer digitalen Rekonstruktion und Visualisierung nahelegen: Die Präsenz von multiplen Klassifikationen durch den historischen Katalog, die Supralibros-Farbklassifikation und den modernen Katalog (GND) sollte nicht nur auf einzelne Objekte angewandt werden, sondern auch in ihrem Wechselspiel im Sinne des Distant Readings gemeinsam visualisiert werden. Zudem werden diese mit dem prominenten materiellen Aufstellungskontext im Prunksaal der ÖNB in Zusammenhang gesetzt, um Zusammenhänge von taxonomischen, diagrammatischen und implizit-räumlichen Anordnungen zu erhellen. Weitere Herausforderungen ergeben sich aus multiplen Unsicherheiten (z.B. Differenzen und nicht empirisch geprüfte Hypothesen in der historischen Sekundärliteratur über die Sammlung, aber auch statistische Unsicherheiten der Machine Learning Klassifikation), und die sehr heterogenen Zielgruppen - von Forscher:innen bis zur interessierten Öffentlichkeit. Als solches manifestiert die Bibliotheca Eugeniana einen komplexen Forschungsgegenstand und Datenbestand, wie er in zahlreichen DH-Projekten vorliegt. Das Projekt Bibliotheca Eugeniana Digital will in diesem Kontext einen wesentlichen Beitrag leisten für die Diskussion von zukünftigen Forschungs- und Visualisierungsstrategien, die über eine reine Datenbereitstellung für die Forschung oder eine Reduktion der Informationen für die interessierte Öffentlichkeit hinausgeht. Der im Projekt entwickelte vielseitige, explorative Zugang eröffnet neue Möglichkeiten, um sich den “postdigitalen” und multi-perspektivischen Herausforderungen im Umgang mit historisch-kultureller Datenkomplexität zu stellen.

Danksagung

Das Projekt Bibliotheca Eugeniana Digital wird durch die österreichische Akademie der Wissenschaften im Rahmen des Calls godigital!3.0 gefördert. Die Autor:innen möchten sich herzlich beim gesamten Projektteam bedanken, insbesondere bei Martin Krickl für seinen Beitrag zum Zustandekommen dieses Projektes.

Bibliographie

Bornhofen, S., Düring, M. Exploring dynamic multilayer graphs for digital humanities. Appl Netw Sci 5, 54 (2020). https://doi.org/10.1007/s41109-020-00295-x

Dörk, M., Müller, B., Stange, J. E., Herseni, J., & Dittrich, K. (2020). Co-Designing Visualizations for Information Seeking and Knowledge Management. Open Information Science, 4(1), 217-235.

Krickl, M., Mayer, S., & Zangger, E. (2022). Mit Machine Learning auf der Suche nach Provenienzen – ein Use Case der Bildklassifikation an der Österreichischen Nationalbibliothek. Bibliothek Forschung und Praxis, 46(1), 227-238. https://doi.org/10.1515/bfp-2021-0090

Mauthe, Gabriele. Die Bibliotheca Eugeniana im europäischen Zeitvergleich. In: Agnes Husslein-Arco und Marie-Louise von Plessen (Hrsg.): Prinz Eugen. Feldherr, Philosoph und Kunstfreund. Katalogbuch zur Ausstellung im Belvedere Wien, 2010. HIRMER VERLAG, pp. 190-220.

Kusnick, J., S. Jaenicke, C. Doppler, K. Seirafi, J. Liem, F. Windhager, & E. Mayr (2021). Report on narrative visualization techniques for OPDB data. Technical report, InTaVia project, 2021. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e47d9524&appId=PPGMS

Schulz, H. J. (2011). Treevis. net: A tree visualization reference. IEEE Computer Graphics and Applications, 31(6), 11-15.

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

Windhager, F., Salisu, S., & Mayr, E. (2019). Exhibiting uncertainty: Visualizing data quality indicators for cultural collections. In Informatics (Vol. 6, No. 3, p. 29). MDPI.

Closing the Gap in Non-Latin-Script Data: Pragmatic Approaches for Increasing Awareness

Beers, Theodore
theo.beers@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0002-5129-5748

Kudela, Xenia Monika
m.kudela@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0002-4582-0542

Müller-Laackman, Jonas
jonas.mueller-laackman@sub.uni-hamburg.de
Staats- und Universitätsbibliothek Hamburg, Deutschland
ORCID: 0000-0003-2279-6751

Multilingual and multi-script research in the Digital Humanities (DH) is still located at the margins of the field. Most projects are either focused on Latin-script materials or need to rely on Latin-script-based solutions. As a result, it is often necessary for DH researchers to build individual workarounds to mitigate the lack of support for non-Latin scripts (NLS). There is, furthermore, a severe lack of visibility and topic-related networks, and often no proper support group, for DH practitioners working with NLS materials.[1] Thus, Multilingual Digital Humanities (MLDH) topics remain marginalized, to the point that even projects active *in* the field have no comprehensive perspective *on* the field.

The purpose of the project “Closing the Gap in Non-Latin Script Data” (https://m-l-d-h.github.io/Closing-The-Gap-In-Non-Latin-Script-Data/) is to intervene here to improve the visibility of research relating to non-Latin scripts in the Digital Humanities, particularly in Germany, and to enhance collaboration among projects dealing with Arabic and similar languages. This serves three goals. First, we want to provide an overview of the state of the field in Germany and thereby draw further attention to projects on the margins of German DH. Second, we seek to promote communication and exchange among researchers, which will facilitate the necessary advancements in the field. Third, on a political level, we aim to challenge the hegemony of English (and other European languages) in DH. The dominance of the mono-cultural view and the need to address it critically has been articulated by many researchers (see, e.g., Fiormonte 2012); however, lack of visibility, difficulty in accessing sources, and inadequacy of research tools are still the daily reality in research on non-European languages—particularly those written in non-Latin scripts.

The central goal of our effort is thus to raise awareness for the field and to give an overview of projects, initiatives, infrastructures, methods, and workflows that are being implemented in Multilingual DH in Germany. We also have the opportunity to leverage the data and expertise that we acquire during the course of this project to establish a set of best practices and guidelines for researchers engaged in the field. Finally, we hope to serve as an exemplar of the practices that we advocate, with a special emphasis on the principles of open and FAIR science. As the core team of “Closing the Gap” is based at the Seminar for Semitic and Arabic Studies at the Freie Universität Berlin (https://www.geschkult.fu-berlin.de/e/semiarab/arabistik/), we started with a focus mainly on projects working with Arabic, but

we have expanded the data model to include many more languages.

As of November 2023, our database covers 159 projects and initiatives around the world, with a geographic focus on Germany. These projects, in turn, work with a total of 112 different languages—most of them written in non-Latin scripts. “Closing the Gap” is integrated in a network of institutions and initiatives, such as the Multilingual DH Lab of the Ada Lovelace Center for Digital Humanities at the Freie Universität Berlin (https://www.ada.fu-berlin.de/), the Department of Digital Scholarship Services at the State and University Library of Hamburg (https://www.sub.uni-hamburg.de/service/digitale-forschungsdienste.html), and the Multilingual DH working group (AG) within the DHd (https://m-l-d-h.github.io/DHd-AG/). Our project therefore serves as a “micro-hub” for NLS-related Multilingual DH.

Many of the projects that we have catalogued are also embedded in the disciplinary background of the so-called “Kleine Fächer” (https://www.kleinefaecher.de/), which often involve NLS languages and suffer from a lack of recognition in the larger academic landscape. “Closing the Gap” aims to address this issue by providing visibility through our database, as well as by showing the interconnectedness of projects in Germany and across Europe.

When we started developing workflows for our project, we aimed for pragmatic solutions that allowed us to commit to OpenScience and FAIR principles from the very beginning, without needing to rely on an institutional infrastructure that was still unable to meet our needs. It was also important for us to act efficiently, given a limited period of funding. While one can debate the risks of conducting academic work on corporate-owned platforms, we chose to build our database in a public GitHub repository (https://github.com/M-L-D-H/Closing-The-Gap-In-Non-Latin-Script-Data), and to use GitHub Pages to host a web frontend that gives users an easy way of exploring the data.

Our main work consists of three pillars:

1. A **data model** (https://github.com/M-L-D-H/Closing-The-Gap-In-Non-Latin-Script-Data/tree/master/SCHEMATA) that allows us to encode most of the information about a project that one would need in order to describe it concisely, while including enough detail on locations, people, topics, and tech stack (among other things). Data is written in the JSON format, which is well-suited to version control and the Pull Request workflow on GitHub. We use an open-source library called Zod (https://zod.dev/) to ensure that all JSON files follow our schema. Whenever a new Pull Request is opened in the repository, this validation process is run automatically. That is, we have put in place a continuous integration (CI) system to ensure that our data remains clean and consistent. In the schema itself, we try to follow established standards, to the extent practical, such as the TaDiRAH taxonomy for keywords (https://de.dariah.eu/en/tadirah). Entities are linked to authority files—e.g., through GND and VIAF—as well as to services like ORCID and ROR. Having a well-documented and reliable data model was one of the main tasks for the first phase of “Closing the Gap.”
2. A **web frontend** (https://m-l-d-h.github.io/Closing-The-Gap-In-Non-Latin-Script-Data/) that shows our data in an approachable manner. The site was first developed in Vue (https://vuejs.org/) and was recently reimplemented with SvelteKit (https://kit.svelte.dev/). Each time that the frontend is loaded, it reads the latest version of our database directly from the GitHub repository, so that everything stays up to date. Since we do not rely on a traditional web server infrastructure, but simply on a directory of JSON files addressed by UUID, one can clone our frontend and database at any time to save a copy of the entire project. One goal for the frontend was that it should be user-friendly, with flat hierarchies and simple search functions by full text, keyword, and language. We have also begun to offer data visualizations, such as a map and a timeline of projects. (This will be developed further in the next funding phase.) The website provides information on how to collaborate with us, and it includes a form that can be used by researchers who want to contribute new data but are not accustomed to working with JSON directly.
3. Raising awareness and building a **network**. Our goals are not limited to gathering information about projects and creating a “network” of NLS-related initiatives in the form of a database. Rather, we hope to foster such networks also in the real world! Given that “Closing the Gap” is completely open and interested in collaboration, we would like to address the DH and DHd communities and invite them to join us in raising awareness of issues pertaining to NLS languages and Multilingual DH. This could mean providing information for our project, or hinting toward initiatives that we have not yet covered. We recently launched a blog aimed at sharing our insights on best practices and facilitating discussion of crucial topics within the field (https://ctg.hypotheses.org/). In the upcoming project phase, our team will organize a series of events, inviting prominent NLS DH researchers to participate in experiential knowledge exchange, thereby fostering enhanced cooperation opportunities. We see collaboration among researchers as the first step toward decolonizing the Multilingual Digital Humanities and “disrupting digital monolingualism.”[2]

Since we strive to keep our work as transparent as possible, we decided to discuss all major project-related matters openly via the Issues function on GitHub. Anyone from the outside can therefore follow and comprehend our decisions. Better still, interested users have the ability to participate in these discussions, to suggest new features or ask questions to be answered by the core team or other collaborators. This way, we also engage in new ways of dealing with pro-

blems and failures, especially regarding the applicability of our stack to non-Latin-script textual data. We hope to raise awareness that a research process is not only a group effort, but that it is often non-linear as well.

In accordance with the conference theme of "DH Quo Vadis," we want to present our ideas and workflows at DHd 2024 and to assess preliminary results of the project at the end of its first funding period. Furthermore, with the second phase of funding recently confirmed, this is an opportune time for us to share our vision of the future and the issues that we hope to address in the next two years, such as better data visualization, how to expand the multilingual use of TaDiRAH, or the potential implementation of a knowledge graph based on our data.

Promoting awareness of the issues facing Multilingual DH will be an essential part of securing the future of the Digital Humanities as a broad and interdisciplinary field *par excellence*. We intend to discuss the advantages and downsides of our approach—both on the technical side, with regard to our stack and the use of GitHub as a database host, and in terms of cultivating a real-life network of researchers who can assist one another.

Fußnoten

1. At the DHd conference in 2022, a working group for Multilingual DH in Germany was founded, aiming to provide a space for networking and to act as such a support group.
2. King's College 2020; see: https://www.k-cl.ac.uk/events/disrupting-digital-monolingualism (accessed July 17, 2023).

Bibliographie

Asef, Esther, and Cosima Wagner. 2018. "Workshop-Bericht: 'Nicht-lateinische Schriften in multilingualen Umgebungen: Forschungsdaten und Digital Humanities in den Regionalstudien.'" *DHd Blog*. https://dhd-blog.org/?p=10669 (zugegriffen: 17. Juli 2023).

BMBF. n.d. *Kleine Fächer – Große Potenziale – BMBF*. Bundesministerium für Bildung und Forschung – BMBF. https://www.bmbf.de/bmbf/de/forschung/geistes-und-sozialwissenschaften/kleine-faecher/kleine-faecher_node.html (zugegriffen: 17. Juli 2023).

Fiormonte, Domenico. 2017. "Digital Humanities and the Geopolitics of Knowledge." In *Digital Studies / Le Champ Numérique* 7 (1). https://doi.org/10.16995/dscn.274.

———. 2021. "Taxation Against Overrepresentation? The Consequences of Monolingualism for Digital Humanities." In *Alternative Historiographies of the Digital Humanities*, 333–76. Earth: punctum books. https://doi.org/10.53288/0274.1.00.

Fiormonte, Domenico, Sukanta Chaudhuri, and Paola Ricaurte, eds. 2022. "Introduction." In *Global Debates in the Digital Humanities*, ix–xxxiii. Minneapolis: University of Minnesota Press.

Ghorbaninejad, Masoud, Nathan P. Gibson, and David Joseph Wrisley. 2023. "Right-to-Left (RTL) Text: Digital Humanists Plus Half a Billion Users." In *Debates in the Digital Humanities 2023*. Minnesota: University of Minnesota Press.

Gil, Alex, and Élika Ortega. 2016. "Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing." In *Doing Digital Humanities*. London: Routledge.

Grallert, Till, Xenia Monika Kudela, Eliese-Sophia Lincke, Colinda Lindermann, Jana-Katharina Mende, Jonas Müller-Laackman, and Larissa Schmid. 2023. *Umgang mit Multilingualität im DACH und DHd Verband (v1.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.7957187.

Ortega, Élika. 2014. "Multilingualism in DH." *Disrupting the Digital Humanities*. https://web.archive.org/web/20210424073656/https://www.disruptingdh.com/multilingualism-in-dh/ (zugegriffen: 17. Juli 2023).

Spence, Paul. 2021. *Disrupting Digital Monolingualism: A report on multilingualism in digital theory and practice*. London: Language Acts and Worldmaking. https://doi.org/10.5281/zenodo.5743283.

Co-Kreativität digital erschließen: Über die Annotation komplexer ästhetischer Phänomene

Bauer, Matthias
m.bauer@uni-tuebingen.de
Universität Tübingen, Deutschland
ORCID: 0000-0003-0395-0629

Göggelmann, Michael
michael.goeggelmann@smail.uni-koeln.de
Universität Tübingen; Universität zu Köln, Deutschland

Rogalski, Sara
sara.rogalski@uni-tuebingen.de
Universität Tübingen, Deutschland

Wetzel, Sandra
sandra-madeleine.wetzel@uni-tuebingen.de
Universität Tübingen, Deutschland

Zirker, Angelika
angelika.zirker@uni-tuebingen.de
Universität Tübingen, Deutschland
ORCID: 0000-0001-6819-3871

Forschungsfrage

Obwohl gemeinschaftliche Autorschaft in der frühen Neuzeit häufig der Normalfall war, geht die Forschungsliteratur weiterhin und gerade dann von einem Konzept der Einzelautorschaft aus, wenn sie sich lediglich auf die Identifizierung der Anteile individueller Autoren (insbesondere Shakespeares) an Gemeinschaftswerken fokussiert (z.B. Vickers, 2002). Das Aufkommen und die Verbesserung digitaler Methoden hat leider nur zu einer Bagatellisierung des Konzepts gemeinschaftlicher Autorschaft geführt; stilometrische Untersuchungen von dramatischen Texten befördern die Vorstellung von gemeinschaftlicher Autorschaft als Summe von Einzelautorschaften und reduzieren die Autor- und Urheberschaft auf den Stil. Gemeinschaftliche Autorschaft ist aber mehr als die Summe ihrer Teile; um sie zu erforschen, braucht man eine Idee davon, was an ihr anders ist. Um dies herauszufinden, sollte man wissen, wie in der Frühen Neuzeit selbst darüber gedacht wurde. Im Projekt zur Ästhetik gemeinschaftlicher Autorschaft[1] haben wir uns zum Ziel gesetzt, das bislang wenig beachtete, inhärente Konzept der Co-Kreativität, das der Praxis gemeinschaftlicher Autorschaft zugrunde liegt, aus englischen literarischen Texten der Frühen Neuzeit zu erschließen. Wir untersuchen daher implizite und explizite Reflexionen co-kreativer Prozesse, die wir z.B. in Metaphern des Gebens und Nehmens finden (etwa George Herberts Widmung seines Gedichtbandes an Gott „The Dedication," Z. 1-2: „from [whom] they came / and must return" (Herbert, 2008, 44; s. dazu Bauer et al., 2023a). Um über einzelne, qualitative Studien hinaus auch einen Blick für wiederkehrende Reflexionsfiguren und weitgreifende Konzepte zu gewinnen, nutzen wir das Annotationsprogramm *CorefAnnotator*[2]; damit können wir in größeren Korpora Reflexionen über Co-Kreativität digital erfassen und systematisieren. Eine solche Systematisierung ästhetischer Reflexion ist allerdings auch eine große Herausforderung für digitale Methoden (siehe Heiniger et al., 2022).

Material

Wir entschieden uns zu Beginn der Arbeit für kurze, aber reflexionsdichte Texte, d.h. Gedichte, um möglichst schnell einen Einblick in verschiedene Autoren und Werke zu erlangen. Bei der Aufbereitung der unserer Arbeit zugrunde liegenden Gedichtkorpora griffen wir überwiegend auf editierte, in digitaler Form vorliegende Werke bekannter frühneuzeitlicher Autoren zurück und bereiteten diese auf Basis intern entwickelter Richtlinien für die Weiterarbeit im *CorefAnnotator* vor.[3] Zunächst beschränkten wir unsere Auswahl auf zuvor von der Annotationsgruppe ausgewählte Fallbeispiele, d.h. wir annotierten zunächst fünf im Hinblick auf die Reflexionsdichte exemplarische Gedichte einiger Dichter. Diese Beispielgedichte dienten zur ersten Entwicklung der Annotationsrichtlinien. Im nächsten Schritt arbeiteten wir ausschließlich mit zufällig gewählten, aber in Proportion zur Korpusgröße des Autors stehenden Gedichten fünf bekannter frühneuzeitlicher Autoren: Edmund Spenser, George Herbert, Henry Vaughan, John Donne und William Shakespeare. Diese 100 Gedichte, aufgeteilt in zwei Korpora mit je 50 Gedichten, dienen nun als Grundkorpus für die Entwicklung, Erprobung und Überarbeitung unserer Annotationsrichtlinien; das annotierte erste Korpus bildet den Datensatz für die ersten Auswertungen.

Methode

Aufgrund der Komplexität des Phänomens, das wir untersuchen, nutzten wir den *CorefAnnotator* von Anfang an als heuristisches Instrument, das uns in erster Linie zur Erfassung des Konzepts der Co-Kreativität diente. Es benötigte zunächst einen Fehlversuch, um eine Methode zu finden, die uns ermöglichte, Reflexionen über Co-Kreativität in Texten nicht nur zu lokalisieren und konkret in *items* zu verankern, sondern auch zu beschreiben und in Zusammenhang zu bringen.

Der erste Versuch, co-kreative Reflexionen in Gedichten zu annotieren, basierte auf dem Expertenvorwissen aller Annotator:innen. Wir suchten zunächst nach Bausteinen, die kombiniert, so lautete unsere Hypothese, ein bestimmtes Konzept der Co-Kreativität bildeten. Wenn also z.B. Herbert von einem Austausch der Gedichte zwischen Gott und ihm spricht, könnte man dieses Konzept als aus den Bausteinen des Gebens und Nehmens gebildet analysieren. Die Herausforderung dieser Herangehensweise lag darin, die Bausteine ohne Kenntnis des im Gedicht anzutreffenden Konzepts zu identifizieren. Ein erster Versuch bestand darin, Synonyme bereits bekannter Bausteine in unseren Fallbeispielkorpora ausfindig zu machen und manuell 50 *tokens* vor und nach dem Beleg zu untersuchen, um herauszufinden, ob gewisse Bausteine häufig zusammen mit bestimmten anderen Ausdrücken eine Reflexion über Co-Kreativität bilden. L eider konnte jedoch auf diese Weise kein für die Reflexion von Co-Kreativität charakteristisches Vokabular ermittelt werden. Folglich richteten wir unsere Untersuchung neu aus und entwickelten das Komponenten-Modell.

Dieses Modell geht davon aus, dass die Reflexion über co-kreative Prozesse einen Sonderfall der Reflexion über Produktionsvorgänge bildet, an denen mehrere Akteure beteiligt sind. Dementsprechend erfolgt die Annotation co-kreativer Reflexionen in mehreren Schritten: zunächst muss der im Text erwähnte Akt der Produktion oder das produzierte Artefakt (A) in den Blick genommen werden, dann die Akteure, auf denen die Co-Kreativität (CO) beruht, so-

wie im letzten Schritt die damit verbundene Prädikation (P), d.h. das, was über die Produktion gesagt wird. Das Zusammenspiel aller drei Komponenten bilden eine co-kreative Konstellation. Verdeutlicht werden kann dies am Beispiel von George Herberts Gedicht „A True Hymne“: Hier heißt es in Z.17-18: „Although the verse be somewhat scant, / God doth supplie the want“ (Herbert, 2008, 574). „Verse“ ist das geschaffene Artefakt (A), das auf dem (hier implizierten) Akt des Schreibens basiert; die Co-Kreativität (CO) besteht in der Zusammenarbeit des Sprechers mit Gott, und die Prädikation (P) ist „supplie the want,“ d.h. die Beschreibung der Leistung des Co-Autors. Co-Kreativität wird hier demnach als Aktivität beschrieben, in der ein Beteiligte:r die Mängel des oder der anderen ausgleicht. Die Annotation sieht wie folgt aus: „Although the [verse] $_{A}$ be somewhat scant, / [God] $_{CO}$ doth [supplie the want] $_{P}$.“

Ein erstes zentrales Ergebnis des Entwicklungsprozesses der Annotationsrichtlinien für die erste Komponente, die Akte und Artefakte (A), war die Feststellung, dass wir einen ganz offenen und neuen Minimalkonsens dafür schaffen mussten, worin ‚Gemachtheit‘ besteht. Als Grundlage für die Annotationsrichtlinien gilt daher:

Act: the process of making an object

Annotate all activities which imply that something is (being) made.

Artefact: a made object

Annotate all items mentioned in the text which have been made, i.e. the product as well as the object that is being changed (e.g. in “[she] […] humbled harts brings captiues vnto thee,” there are two artefacts: “humbled harts” and “captives”). What is made can be abstract (e.g., beauty in Herbert’s “Death” (“and all thy bones with beautie shall be clad,” 20) is used concretely in this making process) and need not be a (physical) object. The requirements are a) that madeness is expressly mentioned in the context and/or b) that the madeness of the product self-evident (the object has been produced, e.g. “composition” or “book”). One of the two requirements must unequivocally apply.

Die Annotation der Akte und Artefakte erforderte zudem ausführliche Angaben zu ihrer Verankerung im Text, wie z.B. die Regelung, dass wir maximal annotieren, d.h. dass wir alle syntaktischen Elemente, die einen Akt oder ein Artefakt spezifizieren, mitannotieren. Darüber hinaus führen die Annotationsrichtlinien eine Reihe von Spezialfällen auf, wie z.B. die Annotation von hypothetischen oder destruktiven Akten und Artefakten.

Wenn alle Akte und Artefakte in einem Korpus erfasst und nummeriert wurden, erfolgt im zweiten Schritt die Annotation der Komponente „CO“. Die Annotationsrichtlinien sehen hier vor:

CO: one of at least two co-creators

Annotate all agents that (hypothetically) participate(d) in an act or in the creation of an artefact if, and only if, the process involves at least two agents.

Da in einem Gedicht mehrere Akte und Artefakte mit unterschiedlichen COs auftreten können, gilt es an dieser Stelle im Annotationsprozess, die einzelnen Komponenten und ihre Zugehörigkeit zu einer Konstellation über verbindende Marker kenntlich zu machen.

Die dritte Komponente, die Annotation der Prädikation (P), befindet sich momentan noch in der Erprobung. Grundsätzlich soll P eine Verhältnisbestimmung der einzelnen Komponenten beitragen. Über die Erfragung des Verhältnisses zwischen „A“ und „CO“ kommt man zur Prädikation. In Donnes „A Valediction of Weeping“ (2008, 112) stehen der Sprecher und die Geliebte z.B. im wechselseitigen Abhängigkeitsverhältnis („Since thou and I sigh one anothers breath,“ Z. 26). Diese wechselseitige Abhängigkeit zwischen den Beteiligten bildet also eine Aussage über einen co-kreativen Prozess, der nach erfolgter Annotation mit anderen Aussagen über ähnliche Prozesse bzw. mit ähnlichen Aussagen über andere co-kreative Prozesse verglichen werden kann. Das Ziel der P-Annotationen ist es, möglichst viele Aussagen über die Verhältnisse der annotierten *items* treffen zu können, um später in der Auswertung des Datensatzes konkret nach bestimmten Konstellationen zu suchen.

Bislang wurden in unserem ersten Korpus mit fünfzig Gedichten alle Akte und Artefakte sowie aller CO-Akteure annotiert. Während die Erprobung der P-Annotationsrichtlinien läuft und das zweite 50er Korpus auf A und CO hin annotiert wird, können uns erste Auswertungen der vorhandenen A- und CO-Annotationen bereits Erkenntnisse liefern, die ohne diesen digitalen Zugang nicht ersichtlich wären und zudem das Potential dieser Annotationsmethode und des *CorefAnnotator* deutlich zeigen. Während wir natürlich über Möglichkeiten der Automatisierung nachdenken, bereitet neben der geringen Größe des Korpus auch die Komplexität der Phänomene aktuell noch Herausforderungen, sodass dieser Aspekt der Annotationsarbeit momentan außerhalb unseres Fokus liegt.

Ergebnisse

Die im *CorefAnnotator* angefertigten Annotationen wurden exportiert und anschließend so aufbereitet, dass alle co-kreativen Konstellationen über die Verankerung der Marker offengelegt und analysiert werden konnten. Im ausgewerteten Korpus sind 311 (von insgesamt 1011) Akte und Artefakte an co-kreativen Konstellationen beteiligt, was bedeutet, dass 30,76% aller erwähnten Akte und Artefakte co-kreativ entstanden sind. Abbildung 1 zeigt zudem deutlich, dass (bis auf acht Ausnahmen) in allen Fällen der co-kreativen Aktivität zwischen zwei Akteuren mindestens zwei Akte/Artefakte geschaffen werden.

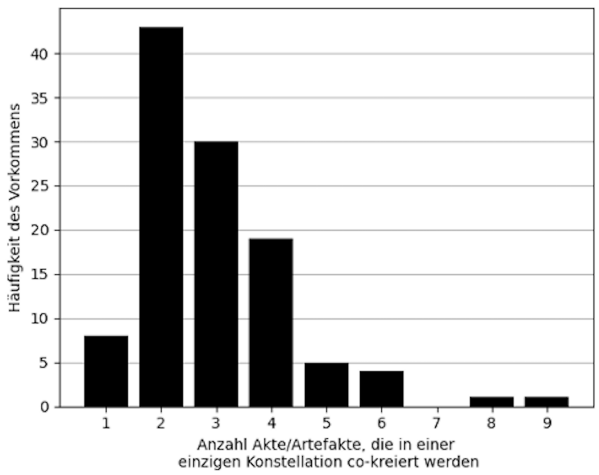


Abbildung 1: Die Anzahl der Akte/Artefakte, die in einer Konstellation co-kreiert werden

Die Anzahl an Akten und Artefakten, die im Rahmen einer einzigen co-kreativen Aktivität erschaffen werden, bezeugt die Komplexität des untersuchten Phänomens. In manchen Fällen werden sogar bis zu neun verschiedene Akte und Artefakte in einer co-kreativen Konstellation kreiert. Die Vielschichtigkeit der Reflexionen, die wir zu erfassen suchen, wird auch in der folgenden Abbildung deutlich, die unterschiedliche CO-Konstellationen gemäß der Häufigkeit ihres Auftretens sortiert:

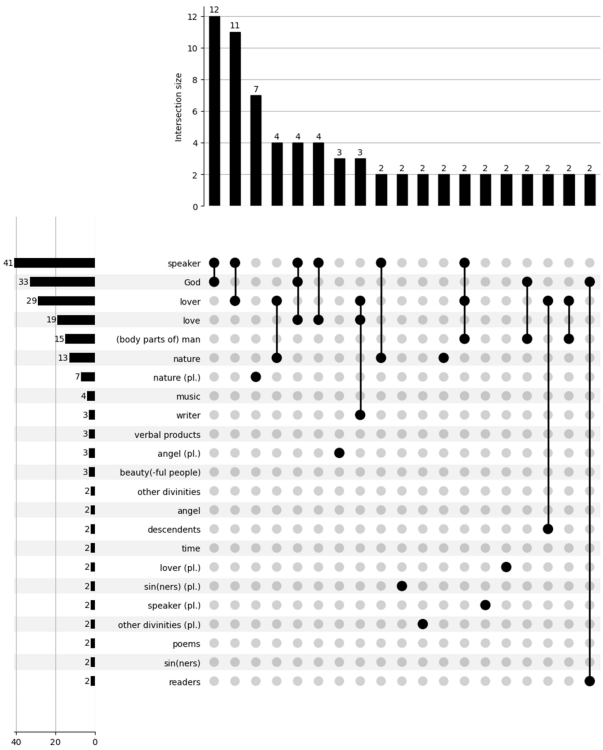


Abbildung 2: CO-Konstellationen nach Häufigkeit des Auftretens

Die Abbildung zeigt auf, welche CO-Akteure besonders häufig miteinander produktiv tätig sind: der Sprecher des Gedichts etwa tritt 41x als CO im Korpus auf, Gott 33x; gemeinsam sind sie in dieser Konstellation 12x co-kreativ tätig. Das Diagramm verschafft also einen ersten Eindruck in die unterschiedlichen CO-Konstellationen zwischen zwei oder mehr Akteuren, die wir in Reflexionen über Co-Kreativität antreffen.

Die Datenlage zeigt insgesamt, dass diese Reflexionen in doppelter Hinsicht hochkomplex sind: sie involvieren mehrere COs in unterschiedlichen Konstellationen, und es werden meist die Schöpfungsprozesse von mehr als zwei Akten und Artefakten reflektiert.

Unsere Annotationsmethode ermöglicht uns aber nicht nur die Erfassung der beteiligten COs und geschaffenen As, obgleich allein diese Daten bereits erkenntnisreich sind, wenn beispielsweise Untersuchungen zu den am häufigsten auftretenden Kollaborationspartnern Gottes erwünscht sind. Unser komplexes Annotationssystem bietet darüber hinaus die Möglichkeit, über die Annotation zusätzlicher Eigenschaften den Konstellationen bestimmte Kategorisierungen zuzuteilen. In diesem Korpus führten wir z.B. den sog. Marker „enabling“ ein, da der heuristische Prozess des Annotierens und Revidierens der Annotationsrichtlinien bereits zur Aufstellung einer Hypothese führte: es schien, als würde Co-Kreativität zwischen Gott und Mensch auf einer Abhängigkeit des Menschen von Gott beruhen. Diese Abhängigkeit konnten wir weiter als eine Befähigung des Menschen zum Kreieren durch Gott spezifizieren. Das ganze Korpus wurde dahingehend untersucht und alle co-kreativen Reflexionen in denen ein CO das andere CO zum schöpferischen Prozess befähigt, erhielten den Marker „enabling.“ Da es sich nicht um eine spezifische Aussage über den Produktionsprozess handelt, wurde diese Abhängigkeit zwischen den Personen nicht als P annotiert.

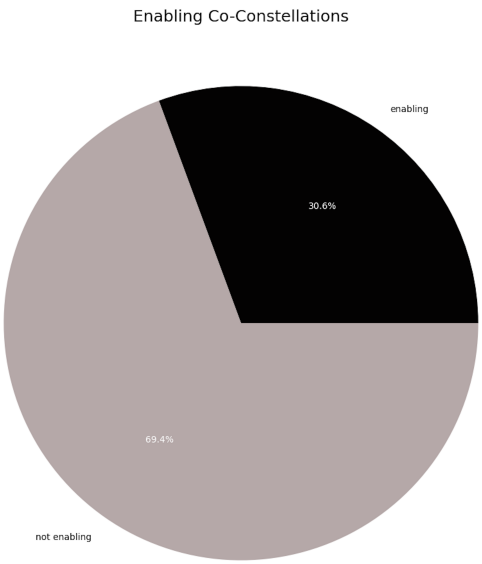


Abbildung 3: Gegenüberstellung aller "enabling"- und "non-enabling"-Konstellationen

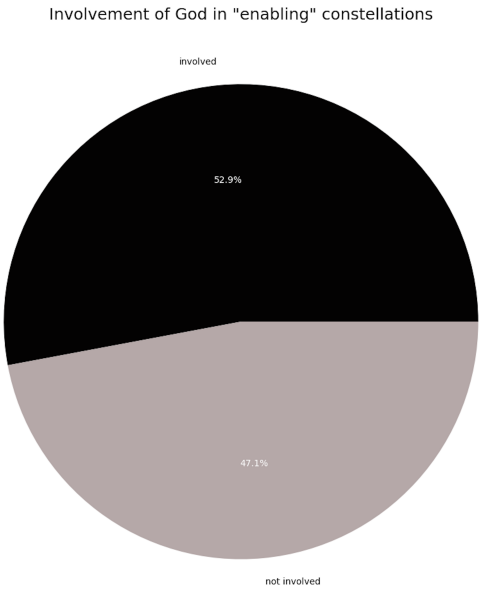


Abbildung 4: Gegenüberstellung aller "enabling"- und "non-enabling"-Konstellationen an denen Gott beteiligt ist

Die Kreisdiagramme zeigen, dass Gott in über 52% aller „enabling"-Konstellationen, die fast ein Drittel aller co-kreativen Konstellationen ausmachen, partizipiert. Damit können wir eine gängige Annahme, dass Gott und Mensch in der frühen Neuzeit nicht als kreative Partner gedacht wurden, auf Grundlage unserer Annotationsmethode und der Auswertung widerlegen. Erkenntnisse dieser Art legen den Baustein für ein breiteres Verständnis eines vormodernen Konzepts der Co-Kreativität. Die Grundlage unserer digitalen Arbeit bilden die Konstellationen aus As und COs, die über Marker um Eigenschaften ergänzt werden, bspw. durch den „enabling"-Marker oder Marker, die Selbstreferenzialität oder Metaphorik kennzeichnen. Die Einführung unserer dritten Komponente, der Prädikation P, wird zusätzliche Eigenschaften der co-kreativen Konstellationen aufzeigen. Der größte Gewinn dieser Annotationsarbeit ist also der große Datensatz an bereits erfassten co-kreativen Reflexionen, die mit Eigenschaftsmarkern versehen und deren Verhältnisse zueinander bestimmt wurden. Diese aufbereiteten Daten lassen uns erkennen, wie über co-kreative Prozesse gesprochen und gedacht wurde; das so entstandene Bild wird helfen, auch die Praxis der kreativen Zusammenarbeit neu zu betrachen. Die vorgestellte komplexe Annotationsmethode erlaubt es nicht nur, komplizierte Reflexionen in Einzelkomponenten herunterzubrechen und zu systematisieren. Sie bietet darüber hinaus methodische Ansätze für die Annotation weiterer komplexer literarischer Phänomene.

Fußnoten

1. Das Projekt C05 Die Ästhetik gemeinschaftlicher Autorschaft in der englischen Literatur der Frühen Neuzeit ist im SFB 1391 Andere Ästhetik angesiedelt; https://uni-tuebingen.de/forschung/forschungsschwerpunkte/sonderforschungsbereiche/sfb-andere-aesthetik/.
2. S. Reiter, 2018. Das Tool *CorefAnnotator* und seine Versionen sind online nachverfolgbar und stehen zum Download bereit: https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/corefannotator/
3. Die bisherigen Annotationsrunden wurden mitsamt den zugehörigen Richtlinien in Zenodo archiviert und öffentlich zugänglich gemacht. Siehe hierzu Zirker et al., 2023.

Bibliographie

Bauer, Matthias, Sarah Briest, Sara Rogalski, und Angelika Zirker. 2023. "Geben und Nehmen. Eine Reflexionsfigur gemeinschaftlicher Autorschaft in der englischen Literatur der Frühen Neuzeit." In *Plurale Autorschaft: Ästhetik der Co-Kreativität in der Vormoderne* , hg. von Stefanie Gropper, Anna Pawlak, Anja Wolkenhauer und Angelika Zirker, 31-52. Berlin, Boston: De Gruyter. DOI: 10.1515/9783110755763-003 .

Heiniger, Anna Katharina, Nils Reiter, Nathalie Wiedmer, Stefanie Gropper und Angelika Zirker. 2022. "Kann man Ästhetik zählen? Systematische Annotation und quantitative Analyse von Erzählerbemerkungen in den *Isländersagas* . " In *Andere Ästhetik: Grundlagen – Fragen – Perspektiven* , hg. von Annette Gerok-Reiter, Jörg Robert, Matthias Bauer und Anna Pawlak, 283-308. Berlin, Boston: De Gruyter.

Donne, John. 2008. "A Valediction of Weeping." In *The Major Works including* Songs and Sonnets *and sermons* , hg. von John Carey, 112. Oxford, New York: Oxford UP.

Herbert, George. 2007. “A True Hymn.” In *The English Poems of George Herbert* , hg. von Helen Wilcox, 574. Cambridge: Cambridge UP.

Herbert, George. 2007. “The Dedication.” In *The English Poems of George Herbert* , hg. von Helen Wilcox, 44. Cambridge: Cambridge UP.

Jowett, John. 2013. “Shakespeare as Collaborator.” In *Shakespeare Beyond Doubt: Evidence, Argument, Controversy* , hg. von Paul Edmondson und Stanley Wells, 88-99. Cambridge: Cambridge UP.

Reiter, Nils. 2018. “CorefAnnotator - A New Annotation Tool for Entity References.” In *Abstracts of EADH: Data in the Digital Humanities* . DOI: 10.18419/opus-10144.

Van es, Bart. 2013. *Shakespeare in Company* . Oxford: Oxford UP, 2013.

Vickers, Brian. 2002. *Shakespeare, Co-Author. A Historical Study of Five Collaborative Plays* . Oxford, Oxford UP.

Zirker, Angelika, Matthias Bauer, Sara Rogalski, Sandra-Madeleine Wetzel und Alexa König. 2023. SFB1391 C05 Annotationen [Data set]. Zenodo. DOI: 10.5281/zenodo.7701515

Communities, Harvesting, and CGIF: Building the Research Data Graph at NFDI4Culture

Steller, Jonatan Jalle
jonatan.steller@adwmainz.de
Academy of Sciences and Literature Mainz, Germany
ORCID: 0000-0002-5101-5275

Söhn, Linnaea Charlotte
linnaea.soehn@adwmainz.de
Academy of Sciences and Literature Mainz, Germany
ORCID: 0000-0001-8341-1187

Tolksdorf, Julia
julia.tolksdorf@adwmainz.de
Academy of Sciences and Literature Mainz, Germany
ORCID: 0000-0002-0495-5897

Bruns, Oleksandra
oleksandra.bruns@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany; Karlsruhe Institute of Technology (AIFB), Germany
ORCID: 0000-0002-8501-6700

Tietz, Tabea
tabea.tietz@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany; Karlsruhe Institute of Technology (AIFB), Germany
ORCID: 0000-0002-1648-1684

Posthumus, Etienne
etienne.posthumus@partners.fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
ORCID: 0000-0002-0006-7542

Fliegl, Heike
heike.fliegl@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
ORCID: 0000-0002-7541-115X

Pittroff, Sarah
sarah.pittroff@adwmainz.de
Academy of Sciences and Literature Mainz, Germany
ORCID: 0000-0001-5134-1081

Sack, Harald
harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany; Karlsruhe Institute of Technology (AIFB), Germany
ORCID: 0000-0001-7069-9804

Schrade, Torsten
torsten.schrade@adwmainz.de
Academy of Sciences and Literature Mainz, Germany
ORCID: 0000-0002-0953-2818

Problem: truly linked research data

As a consortium of the *Nationale Forschungsdateninfrastruktur* (NFDI), NFDI4Culture is tasked with developing solutions to systematically make accessible and interconnect the rich decentralised research data available from various providers across its domains. These include architectural studies, art history, musicology, performing arts, and media studies. The overarching goal is to make such data usable for further research in the long term.

Research data in the NFDI4Culture domains largely exists in silos. Even though a large number of data providers subscribe to the use of authority files and controlled vocabularies like the GND, VIAF, Wikidata, or Iconclass to structure their research data, the resources they publish are not automatically ‘linked’ to the full extent of 5-star Linked

Open Data (LOD) (cf. Berners-Lee 2009). While many data providers support their users in getting from an individual resource to authority data, the reverse research path across individual repositories is largely obscured.

NFDI4Culture is building an information system that should enable users to find highly specific resources like, for example, images, objects, and 3D models depicting a specific motif based on authority files and controlled vocabularies. As such, it should also allow participating projects to retrieve related data from other participants in order to connect data based on information such as time, location, resource type, or motif and make them accessible for further research – even beyond the boundaries of individual research domains. The information system is further required to produce FAIR research data, i.e. data that is findable, accessible, interoperable, and reusable (cf. Wilkinson et al. 2016).

The paper is structured in the following manner. First, it reviews existing solutions for interconnecting research data (section 2). Then, an outline of the approach we chose to satisfy the above requirements is given (3). The next section discusses the implementation of the 'Research Data Graph' introduced in this paper (4). The final section outlines ongoing work to enhance and promote the presented solution across and beyond NFDI (5).

Review: centralised and federated infrastructures

Multiple approaches are possible to interconnect research data. Centralised infrastructures, for example, contain large amounts of data in a single location, and participating projects need to compile and contribute their data regularly for users to be able to find up-to-date content. Federated infrastructures, on the other hand, may have overarching interfaces but directly pass on requests to the participating data providers and need to collect and output their responses to queries.

Classical examples of centralised information systems in the culture domain are the German Digital Library (DDB) or Europeana(cf. Deutsche Digitale Bibliothek, n.d.; Europeana, n.d.). They ingest large amounts of data about physical objects via a network of aggregators such as museums and other libraries. In the case of Europeana, they ingest community standards such as LIDO and transform the data into the Europeana data model(cf. Isaac 2013, 4–6). While the object-focused data model is too restrictive for all the domains NFDI4Culture covers, Europeana's centralised approach enables them to semantically enhance data by applying a number of vocabularies to each record(cf. Isaac et al. 2015, 2).

Compared to centralised systems, federated infrastructures emphasise a shared API over shared formats. This requires more effort from individual data providers to implement a reliable endpoint, but has the benefit of providing information that is as up-to-date as a data provider is willing and able to deliver. In addition, fully federated systems can be less strict about the licence that data is made available under. The CLARIN Federated Content Search (FCS), for example, requires participants to implement an endpoint for the Search/Retrieve via URL (SRU) protocol and the Contextual Query Language (CQL) with responses serialised as XML(cf. CLARIN, n.d.), but does not require providers to specify a licence that governs how their data may be reused.[1] The technology is being reused by the NFDI consortium Text+ to interconnect linguistic data(cf. Körner et al. 2023), but does not naturally lend itself to NFDI4Culture due to CQL's limitation to text corpora. More recent approaches on this side of the spectrum require REST APIs, as in the case of the FCS developed in the ELEXIS lexicography project (cf. ELEXIS 2022), or SPARQL endpoints, which have federation built into the standard (cf. Prud'hommeaux and Buil-Aranda 2013).

Two existing projects stick out due to their hybrid approaches, which served as inspiration for NFDI4Culture. Firstly, Wikidata combines its centralised storage with participatory data management and the option to query its data via SPARQL (cf. Vrandečić and Krötzsch 2014).[2] Secondly, correspSearch allows participants to hand in correspondence metadata in a limited TEI XML format called CMIF in an effort to allow scholars to find correspondence data across corpora (cf. Dumont 2022).

Solution: the Research Data Graph

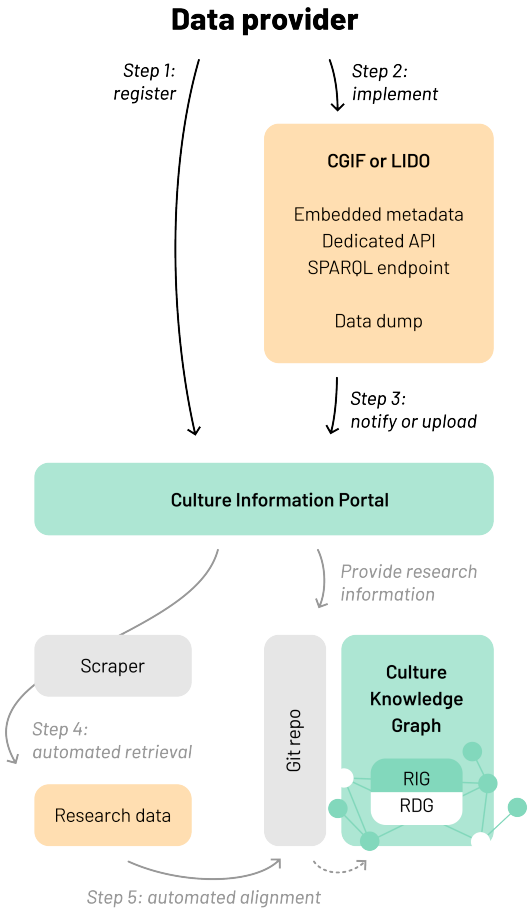
The solution implemented by NFDI4Culture aims to combine the authority and extensibility of a centralised repository with the diversity of federated APIs. The so-called Research Data Graph (RDG) organises a limited set of metadata on research data from participating providers into a knowledge graph. The goal is to provide data that is as granular as possible, but without demanding a specific level of detail: while the Corpus Vitrearum Germany, for example, provides metadata on individual images of stained-glass windows, a repository service like RADAR4Culture only has metadata on entire data sets which they store. Both of these data types are clearly marked as such and thus live next to each other in the RDG. The metadata from various contributors is connected to institutional data already available in the Research Information Graph (RIG), which is collated based on the data stored in NFDI4Culture's Culture Information Portal (cf. Tietz, Bruns, Söhn et al. 2023; Tietz, Bruns, Fliegl et al. 2023).[3] The RDG and the RIG together form the Culture Knowledge Graph.

To get metadata into the RDG, providers may implement the lightweight, RDF-based Culture Graph Interchange Format (CGIF) (cf. Bruns, Posthumus, Sack et al. 2023). We designed CGIF by reusing a narrow set of schema.org classes and properties. Resources can be classified as any resource class schema.org provides.[4] In addition to an identifier of the data provider and the data set, it mainly consists of a feed of individual resources with URIs enhanced by date ranges and keywords to express, for example,

time, place, and motif. The keywords are IDs from authority files and controlled vocabularies such as VIAF, GND, Wikidata, Getty AAT, Iconclass, and GeoNames (cf. BARTOC, n.d.), which are used in the graph to connect resources across data providers.

As a hybrid of centralised and federated approaches, CGIF may be provided either as embedded metadata, a dedicated API, or a SPARQL endpoint/query that can be harvested periodically, or as a data dump in any RDF serialisation. The goal behind this decision is to make data contributions as easy as possible: regardless of whether a project is fully engaged in LOD and able to SPARQL, uses a content management system with limited access to its inner workings, or uses a workflow based on transforming data from XML, CSV, JSON, or other sources into various formats. As an alternative route, a conversion of LIDO to CGIF has been implemented to utilise existing, fine-grained object metadata available across projects and organisations participating in NFDI4Culture.[5]

The combined triples of the Culture Knowledge Graph (RDG and RIG taken together) are made available via the Culture Information Portal. It hosts a triple store with a SPARQL endpoint, which is also available through a search interface and may be used to query data based on, for example, one of the keywords, a specific data type, a time period, or institutions. The endpoint may also be queried by other websites to retrieve information such as related entries based on a keyword. The Culture Information Portal does not host images or other files harvested via one of the options listed above, but includes, for example, URLs of preview images and IIIF manifests, if available.



The current process to add research data to the Culture Knowledge Graph.

Implementation: technological choices

To allow for a broad range of resource types, schema.org was chosen over other data interchange options like LIDO(cf. Coburn et al. 2021), which is restricted to information about physical objects, or CIDOC CRM (cf. Bekiari et al. 2022), which requires much more elaborate data than many projects under the umbrella of NFDI4Culture are able to provide. The CGIF is designed to be an intermediary that allows speedy information retrieval, and thus as an abstract addition to more specific formats across various data domains. Using schema.org for the high-level purpose of interconnecting diverse sets of research data has precedent (i.e. Verma et al. 2022, 1065, 1071). Compared to solutions like CLARIN's Component Metadata (cf. Windhouwer and Goosen 2022), schema.org is already widely used by private-sector search providers and goes beyond linguistic data. Projects which implement it as embedded metadata also make their content more machine-readable outside the realm of academic data repositories.

As fig. 1 illustrates, the process to add data to the RDG currently begins with a data provider (or someone acting on their behalf) registering their data set (and the institutions involved, if not available yet) in the Culture Information Portal (1). They implement CGIF or LIDO (2) and notify the portal when they are ready or upload their transformed research data as a data dump (3). The portal starts a custom scraper to generate triples from embedded metadata, a dedicated API, RDF behind a SPARQL endpoint, or a file dump (4).[6] In a last step, the data is filtered according to the CGIF specification, aligned with the existing RIG data and the NFDICORE ontology (cf. Bruns, Sack, Posthumus et al. 2023), and saved in a Git repository, which is then ingested into the combined knowledge graph (5). The Git repository is used to version-control triples that are being ingested and to allow for reproducing the entire graph.

While the CGIF was designed as part of a low-threshold ingest pipeline, community members are invited to use, reuse, or enhance the open-source components of the Culture Knowledge Graph. The custom scraping mechanism we currently use, for example, may also be used outside the portal to harvest paginated RDF data or to test and troubleshoot CGIF implementations: the Hydra Scraper (cf. Steller

2023) was originally developed as part of the Corpus Vitrearum, one of NFDI4Culture's participants. The scraper is based on the Python library RDFLib (cf. RDFLib team 2023), due to its compatibility with various RDF dialects. Pyoxigraph (cf. Oxigraph contributors 2023) is used as a triple store due to its speed.

The Culture Knowledge Graph builds on Semantic Web technologies. The alignment routine is necessary to allow for a lightweight interchange format that is compatible with search engine optimisation *and* a graph that is easy and uniform to query via the portal's SPARQL endpoint. As part of the alignment, irrelevant triples are filtered and some of the schema.org literals are converted to allow for reasoning via SPARQL. The schema.org property “temporalCoverage,” for example, is used in the CGIF to mark a resource's temporal origin. For time-based reasoning, however, the property needs to be transformed from a string into at least two standard XML Schema “dateTime” values and may even be automatically mapped to a vocabulary in the future. Additional automated clean-ups and filters may become necessary as we proceed with integrating metadata from further sources.

Road ahead: accessibility and network effect

As both the graph itself and the harvesting pipeline are operational, we are now focusing on two areas to iterate upon and improve the Research Data Graph. On the one hand, we are looking to enable scholars to more easily retrieve the data they require by improving the search frontend available in the Culture Information Portal. Since SPARQL is a powerful but also challenging interface for those who are unfamiliar with RDF, we are experimenting with visual interfaces to build the highly specific queries scholars require to retrieve the right information.

On the other hand, our efforts now focus on working with individual projects, communities, and other NFDI consortia to help them contribute data, to come up with sample data transformations, and to make full use of the portal's SPARQL capabilities in web applications. To help connect the vast amount of LIDO data available across NFDI4Culture, for example, we are trialling automated transformations of the relevant metadata into CGIF via a plain ElementTree retrieval in Python, but may yet decide to make this transformation more reusable by reimplementing it in RML (cf. Dimou and Vander Sande 2022) or the web service XTriples (cf. Schrade 2019). In the same vein we are trialling automated transformations for further community standards. Since a number of participants in NFDI4Culture use Wikibase, we are also working towards a best practice for integrating CGIF classes and properties with data managed in Wikibase instances. Last but not least, we are discussing the Research Data Graph with other NFDI consortia and the international Semantic Web community aiming at further adoption, participation, and contribution.

Fußnoten

1. Europeana and Wikidata, on the other hand, only allow CC0-licensed content.
2. Some recent infrastructure initiatives, like the DraCor project as part of CLS INFRA, rely on Wikidata as a community data repository with a SPARQL endpoint (cf. Fischer et al. 2019, 4).
3. In the following, ‘research information’ refers to metadata on organisations, funding, publications, and sometimes whole data sets. The Research Information Graph aims for compatibility with services like the OpenAIRE Graph (cf. Manghi et al. 2019) by implementing the CERIF data model. ‘Research data,’ on the other hand, here refers to more granular items in larger data sets. The distinction between the two can be blurry, however, when it comes to metadata ingested from long-term storage repositories like RADAR4Culture.
4. The schema.org classes and properties have already become a de-facto standard for providing machine-readable data in websites and may, for example, provide structured data about persons, creative works, and intangible entities. They were originally produced by large corporations such as Google, Yahoo!, and Microsoft, but are now shaped and extended by a lively community.
5. See section 5 for efforts to engage with providers who use further community standards such as Wikibase.
6. If an endpoint is used to harvest the data, a modification date in the CGIF implementation is periodically checked to see if it needs to reindex a feed and update the graph.

Bibliographie

BARTOC. n.d. “Vocabularies”. Basic Register of Thesauri, Ontologies & Classifications. Accessed 3 July 2023. https://bartoc.org/vocabularies.

Bekiari, Chryssoula, George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead and Athanasios Velios. 2022. *Definition of the CIDOC Conceptual Reference Model.* V. 7.1.2, June. Accessed 11 July 2023. https://www.cidoc-crm.org/sites/default/files/cidoc_crm_v7.1.2.pdf.

Berners-Lee, Tim. 2009. “Linked Data”. Design Issues, 18 June 2009. Accessed 30 June 2023. https://www.w3.org/DesignIssues/LinkedData.html.

Bruns, Oleksandra, Etienne Posthumus, Harald Sack, Linnaea Söhn, Torsten Schrade, Jonatan Jalle Steller, Tabea Tietz and Julia Tolksdorf. 2023. *Culture Graph Interchange Format Specification.* V. 1.1.0. Mainz, 22 September 2023. Accessed 2 October 2023. https://doi.org/10.5281/zenodo.8369661.

Bruns, Oleksandra, Harald Sack, Etienne Posthumus, Tabea Tietz, and Jörg Waitelonis. 2023. *NFDICORE Ontology.* V. 1.1.0. Karlsruhe, 27 February 2023. Accessed

1 December 2023. https://github.com/ISE-FIZKarlsruhe/nfdicore.

CLARIN. n.d. “Federated Content Search (CLARIN-FCS): Technical Details”. CLARIN ERIC. Accessed 1 July 2023. https://www.clarin.eu/content/federated-content-search-clarin-fcs-technical-details.

Coburn, Erin, Richard Light, Jutta Lindenthal, Gordon McKenna, Regine Stein, Axel Vitzthum and Michelle Weidling. 2021. *LIDO Schema.* V. 1.1, 30 December 2021. Accessed 11 July 2023. https://lido-schema.org/schema/v1.1/lido-v1.1.html.

Deutsche Digitale Bibliothek. n.d. “Kultur und Wissen online”. Deutsche Digitale Bibliothek. Accessed 3 July 2023. https://www.deutsche-digitale-bibliothek.de.

Dimou, Anastasia, and Miel Vander Sande. 2022. *RDF Mapping Language (RML).* In collaboration with Ben De Meester, Pieter Heyvaert and Thomas Delva. V. 1.1.1, 16 November 2022. Accessed 3 July 2023. https://rml.io/specs/rml.

Dumont, Stefan. 2022. “Correspondence Metadata Interchange Format”. CorrespSearch, 6 March 2022. Accessed 1 July 2023. https://correspsearch.net/en/documentation.html.

ELEXIS. 2022. “ELEXIS Protocol for Accessing Dictionaries”. GitHub, 20 April 2022. Accessed 1 July 2023. https://elexis-eu.github.io/elexis-rest.

Europeana. n.d. “Discover Europe’s Digital Cultural Heritage”. Europeana. Accessed 3 July 2023. https://www.europeana.eu.

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*, 1–6. Utrecht: Zenodo. Accessed 2 October 2023. https://doi.org/10.5281/zenodo.4284002.

Isaac, Antoine, ed. 2013. *Europeana Data Model Primer*. The Hague: Europeana, 14 July 2013. Accessed 29 June 2023. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf.

Isaac, Antoine, Hugo Manguinhas, Valentine Charles and Juliane Stiller, eds. 2015. *Selecting Target Datasets for Semantic Enrichment.* The Hague: Europeana, 29 October 2015. Accessed 29 June 2023. https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/EvaluationEnrichment_SelectingDatasets_102015.pdf.

Körner, Erik, Thomas Eckart, Axel Herold, Frank Wiegand, Frank Michaelis, Matthias Bremm, Louis Cotgrove, Thorsten Trippel and Felix Rau. 2023. *Federated Content Search for Lexical Resources (LexFCS): Specification.* Genève: Zenodo, 9 May 2023. Accessed 1 July 2023. https://doi.org/10.5281/zenodo.7986303.

Manghi, Paolo, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, and Pedro Principe. 2019. “The OpenAIRE Research Graph Data Model”. Zenodo. https://doi.org/10.5281/zenodo.2643199.

Oxigraph contributors. 2023. *Pyoxigraph.* V. 0.3.18, 13 June 2023. Accessed 11 July 2023. https://github.com/oxigraph/oxigraph/tree/main/python.

Prud’hommeaux, Eric, and Carlos Buil-Aranda. 2013. *SPARQL 1.1 Federated Query*. In collaboration with Andy Seaborne, Axel Polleres, Lee Feigenbaum and Gregory Todd Williams. V. 1.1, 21 March 2013. Accessed 1 July 2023. http://www.w3.org/TR/2013/REC-sparql11-federated-query-20130321.

RDFLib team. 2023. *RDFLib.* V. 6.3.2, 26 March 2023. Accessed 11 July 2023. https://github.com/RDFLib/rdflib.

Schrade, Torsten. 2019. *XTriples.* V. 1.4.0, 25 March 2019. Accessed 3 July 2023. https://doi.org/10.5281/zenodo.2604986.

Steller, Jonatan Jalle. 2023. *Hydra Scraper*. V. 0.8.4. Mainz, 22 October 2023. Accessed 1 December 2023. https://github.com/digicademy/hydra-scraper.

Tietz, Tabea, Oleksandra Bruns, Heike Fliegl, Etienne Posthumus, Torsten Schrade and Harald Sack. 2023. “Knowledge Graph-basierte Forschungsdatenintegration in NFDI4Culture”. In *DHd2023: Open Humanities, Open Culture: Konferenzabstracts*, 181–185. Trier: Zenodo, 10 March 2023. Accessed 2 July 2023. https://doi.org/10.5281/zenodo.7715509.

Tietz, Tabea, Oleksandra Bruns, Linnaea Söhn, Julia Tolksdorf, Etienne Posthumus, Jonatan Jalle Steller, Heike Fliegl et al. 2023. “From Floppy Disks to 5-Star LOD: FAIR Research Infrastructure for NFDI4Culture”. In *DaMaLOS 2023: 3rd Workshop on Metadata and Research (objects) Management for Linked Open Science*, 1–12. Hersonissos: Publisso, 16 June 2023. Accessed 2 July 2023. https://doi.org/10.4126/FRL01-006444986.

Verma, Shilpa, Rajesh Bhatia, Sandeep Harit and Sanjay Batish. 2022. “Scholarly Knowledge Graphs through Structuring Scholarly Communication: A Review”. *Complex & Intelligent Systems* 9, no. 1 (9 August 2022): 1059–1095. ISSN: 2198-6053, accessed 2 July 2023. https://doi.org/10.1007/s40747-022-00806-6.

Vrandečić, Denny, and Markus Krötzsch. 2014. “Wikidata: A Free Collaborative Knowledgebase”. *Communications of the ACM* 57, no. 10 (23 September 2014): 78–85. ISSN: 0001-0782, accessed 1 July 2023. https://doi.org/10.1145/2629489.

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. 160018 (15 March 2016): 1–9. ISSN: 2052-4463, accessed 30 June 2023. https://doi.org/10.1038/sdata.2016.18.

Windhouwer, Menzo, and Twan Goosen. 2022. “Component Metadata Infrastructure”. In *CLARIN: The*

Infrastructure for Language Resources, 191–222. Berlin: De Gruyter. https://doi.org/10.1515/9783110767377-008.

Computational Game Studies? Drei Annäherungsperspektiven

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities, Universität Leipzig, Deutschland
ORCID: 0000-0003-1354-9089

Piontkowitz, Vera
vera.piontkowitz@uni-leipzig.de
Computational Humanities, Universität Leipzig, Deutschland
ORCID: 0000-0003-3605-3609

Das Big Tent Digital Humanities

Die starke Heterogenität der Digital Humanities (DH) ist mittlerweile zu einem eigenen Topos geworden. Beim Versuch einer Definition wird gerne auf die Website *whatisdigitalhumanities.com*, mit ihren 817 unterschiedlichen Definitionsansätzen, verwiesen, um die Undefinierbarkeit des Fachs zu illustrieren. Ein weiterer Versuch, die inhaltliche Breite der DH abzubilden, zeigt sich in der oft gebrauchten *big tent*-Metapher, welche auf dem Motto der ADHO DH-Konferenz 2013 in Nebraska basiert. Gleichzeitig gibt es kritische Stimmen, die einen übermäßig inklusiven Ansatz problematisch sehen (vgl. Terras, 2011), weil die DH dadurch Gefahr laufen, zu einer leeren Worthülse zu verkommen. Um dem entgegenzuwirken, gibt es zwischenzeitlich Versuche, dieses *big tent* grundlegend zu strukturieren. So schlägt etwa Burghardt (2020) – basierend auf Roths (2019) Dreiteilung – die folgenden vier Teilbereiche der DH vor: (1) digitized humanities, (2) humanities of the digital, (3) public humanities und (4) computational humanities (vgl. Abb. 1).

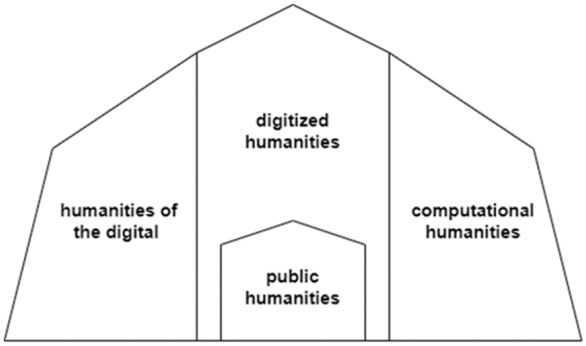


Abbildung 1: Das Big Tent Digital Humanities mit seinen vier Teilbereichen.

Game Studies und Digital Humanities

Als genuin digitales Kulturphänomen haben Computerspiele das Potenzial sich unmittelbar in das Big Tent der Digital Humanities einzufügen. Tatsächlich aber besteht das Feld der Game Studies schon mindestens zwei Dekaden, (vgl. etwa Unterhuber, 2021; Inderst, 2022), und das weitestgehend parallel zu den DH. Ganz ähnlich wie die DH, sind auch die Game Studies von einem hohen Maß an Interdisziplinarität geprägt. Die Methodik in den Game Studies ist jedoch zum allergrößten Teil nicht-digital, sondern überwiegend qualitativ-hermeneutisch, individuell geprägt durch die jeweiligen fachlichen Standards der unterschiedlichen beteiligten Disziplinen, etwa die Medien-, Kultur-, Literatur- und Sprachwissenschaft, Soziologie, Psychologie etc. Das genaue Verhältnis von Game Studies und Digital Humanities wurde bislang allenfalls episodisch diskutiert, etwa in einem Vortrag von Astrid Ensslin mit dem Titel „Video Games in/as Digital Humanities? Corpora, Code and Critical Co-Design”[1]. Nachfolgend wollen wir deshalb ein explizites Mapping von bestehenden Game Studies-Beiträgen auf einzelne Teilbereiche des Big Tent DH versuchen. Dabei fällt auf, dass sich drei der vier Teilbereiche recht unmittelbar auf die bestehenden Game Studies abbilden lassen. Man könnte also – die Game Studies mögen das kritisch sehen – durchaus argumentieren, dass nach dem Big Tent-Modell ein Großteil der bestehenden Game Studies-Forschung problemlos dem breiten Feld der DH zugeordnet werden kann. Nachfolgend finden sich Beispiele für bestehende Game Studies-Publikationen, jeweils abgebildet auf die Big Tent-Bereiche „Humanities of the Digital“, „Digitized Humanities“ und „Public Humanities“:

(1) Humanities of the Digital: Diese DH-Kategorie beschreibt die geistes- und kulturwissenschaftliche Auseinandersetzung mit digitalen Kulturartefakten und Kommunikation im digitalen Raum, so auch mit Videospielen. Beispielhaft seien hier eine Untersuchung von Thach (2021) zur Repräsentation von trans-Charakteren in Videospielen oder eine Abhandlung zu kolonialer Ideologie im

Spiel “The Legend of Zelda” (Hutchinson, 2021) genannt. Auch literaturwissenschaftliche Untersuchungen von Spielen, wie etwa die Analyse von Rhetorik und Narrativität in “Portal” (Wendler, 2014), können hier beispielhaft aufgeführt werden. Diese Spielart der DH scheint aktuell auf den Großteil bestehender Game Studies-Forschung zuzutreffen.

(2) Digitized Humanities: Hiermit sind die Entwicklung und Anwendung von Methoden der Digitalisierung, Sammlung und Modellierung geistes-, kultur- und sozialwissenschaftlicher Daten sowie deren Analyse und Visualisierung gemeint. Innerhalb der Game Studies sind an dieser Stelle beispielsweise die Einbindung materieller und immaterieller Kulturartefakte in Videospiele (vgl. Vosinakis, Avradinis, & Koutsabasis, 2018) und deren Analyse (Balela & Mundy, 2015) zu nennen. Darüber hinaus stellen Videospiele selbst kulturelles Erbe im Sinne von born digital-Medien dar, die ihrerseits komplexer Archivierungslösungen bedürfen. Der Erhalt des digitalen Kulturerbes wurde 2014 von der Gesellschaft für Informatik als eine von fünf *Grand Challenges*[2] der Informatik formuliert und Ansätze der Archivierung von Videospielen werden ständig weiterentwickelt (vgl. etwa Harkai, 2022). Ein weiteres Beispiel für diese Kategorie stellen textkritische (digitale) Editionen digitaler Fachmagazine (sogenannter DiskMags) dar (Roeder & Rettinghaus, 2020).

(3) *Public Humanities:* Diese Kategorie umfasst all jene Projekte, die sich der Wissensvermittlung und Wissenschaftskommunikation widmen. Als Beispiel aus den Game Studies sei hier etwa die Arbeit von Aitkin (2004) genannt, der schon früh das Potenzial von Videospielen für die Wissenschaftskommunikation thematisiert. Ein weiteres Beispiel stellt das Spiel “Walden, a game” dar, welches einen Roman des Schriftstellers Henry David Thoreau als Videospiel verarbeitet und Forschung über dessen Leben, Schriften, Beziehungen und Glauben einbezieht (Fullerton, 2020).

Herausforderungen und Potenziale von Computational Game Studies

Für die obigen drei Bereiche finden sich vielfältige weitere Beispiele in einschlägigen Game Studies-Publikationsorganen, wie etwa dem Online Journal *Game Studies*[3] oder den regelmäßig publizierten Proceedings der DiGRA-Konferenz (*Digital Games Research Association*)[4]. Auffällig ist, dass es für den spezifischen Big Tent-Bereich der Computational Humanities (CH) kaum Entsprechungen in den Game Studies gibt. Dies ist zunächst auch nicht weiter verwunderlich, denn die CH verstehen sich primär als angewandte Informatik oder Datenwissenschaft, die Methoden aus dem Bereich der Statistik und des maschinellen Lernens auf große Korpora sozio-kultureller Artefakte wie etwa Bücher, Bilder, Lieder oder Filme anwendet. Computerspiele lassen sich jedoch nicht ohne Weiteres mit diesen Methoden analysieren, da sie sich durch ihre prozedurale und interaktive Natur (vgl. Bogost, 2007) sowie der Diskrepanz zwischen Spieldesign und dem tatsächlichen Spielen eines Spiels (vgl. Hutchinson, 2021) wesentlich von anderen Medientypen unterscheiden. Zentral für die im Folgenden vorgestellten Annäherungsperspektiven an einen Computational Game Studies-Ansatz ist die Komplexitätsreduktion des multimodalen Phänomens Computerspiel, für das Hawreliak (2018) insgesamt fünf relevante semiotische Modi benennt: Text, Bild, Audio, Haptik und Prozeduralität. Während in der Multimodalitätsforschung (Bateman, Wildfeuer & Hiippala, 2017) die ganzheitliche Betrachtung und Interaktion zwischen einzelnen Modi eine zentrale Forderung sind, so greifen wir für eine erste Annäherung zunächst drei Modi heraus und betrachten diese jeweils als isoliertes Phänomen. Konkret diskutieren wir dabei die Perspektive der (i) Sprache von/in Computerspielen, (ii) Musik in Spielen und (iii) Computerspiele und ihre Remediatisierung als Videos.

Ludophilologie und Ludolinguistik: Die Sprache von/in Computerspielen

Innerhalb der Game Studies wird bei der Betrachtung natürlicher Sprache zwischen ‘ *Orthogame-Sprache’* – *im* Spiel verwendete Sprache und Text – und ‘ *Paragame-Sprache’* – Sprache *über* ein Spiel – unterschieden (Ensslin, 2012). Beispiele für die Anwendung computergestützter Methoden zur Analyse von Computerspielsprache finden sich in beiden Kategorien, beginnend mit Orthogame-Sprache. So stellt etwa Aycock (2017) ein System vor, welches sowohl die statische als auch dynamische Analyse von Text-Adventure-Spielen mithilfe computergestützter Methoden erlaubt. Um die Repräsentation von Gender in Videospielen in der Breite zu analysieren, wendet Heritage (2020; 2022) korpuslinguistische Methoden auf Korpora von in Spielen verwendeter Sprache an. Zahlreiche Beispiele finden sich auch für die computergestützte Analyse von Paragame-Sprache. Ensslin (2012) etwa reflektiert in ihrem Buch die Erstellung und Nutzung eines Korpus, bestehend aus Artikeln aus Videospielmagazinen, Spieleforen und -chats sowie Live-Konversationen aus dem Gameplay. Das *Game Walkthrough Corpus* (Burghardt & Tiepmar, 2021) erlaubt es Forscher:innen, Walkthrough-Texte und Metadaten von über 6.000 Spielen computergestützt als textuelle Abstraktion und Zusammenfassung von Computerspielen zu untersuchen. Weitere Untersuchungsgegenstände der Ludophilologie sind beispielsweise Tweets über Spiele (Wallner et al., 2019), Wikipedia-Artikel zu Spielen (Ryan et al., 2015), Spielerezensionen (Wang und Goh, 2020) oder Publikationen im Bereich der Game Studies (Coavoux, Boutet & Zabban 2017).

Ludomusikologie: Musikanalyse bei Computerspielen

Während sich für die computergestützte Analyse der Sprache von Computerspielen vielfältige Forschungsbeispiele in der Literatur finden, gilt es für den Bereich der Musik zunächst einen Methodentransfer aus der "Computational Musicology" (vgl. Meredith, 2016; Müller, 2021) zu bewerkstelligen, um so bestehende qualitative Ansätze im Bereich der Ludomusikologie (vgl. etwa Summers & Hannigan, 2016; Kamp et al., 2016) um quantitative Verfahren zu erweitern. Zum Standardrepertoire der Audiosignalverarbeitung gehören hier etwa die Spektrogrammanalyse, mit deren Hilfe man grobe Strukturen und Muster in Musikstücken erkennen kann, sowie auch die Chroma Feature-Analyse, die etwa zur Vorhersage von Akkordfolgen benutzt werden kann. Abbildung 2 zeigt eine beispielhafte Analyse von unterschiedlichen Musikstücken aus dem Super Mario-Franchise.

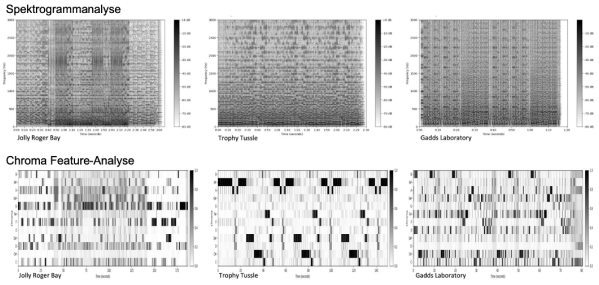


Abbildung 2: Spektrogrammanalyse und Chroma Feature-Analyse von drei Stücken aus dem Super Mario-Franchise (Super Mario 64, Super Mario Smash Bros. und Luigi's Mansion), angefertigt von Jonas Wernicke im Rahmen des Seminars "Computational Game Studies" im Sommersemester 2023, Seminarleitung Manuel Burghardt.

Während mit den genannten Methoden Audio-Features aus beliebigen Musikstücken extrahiert werden können, so kann die Erstellung eines größeren Song-Korpus mühsam sein, insbesondere, wenn die Stücke nicht unmittelbar auf Online-Plattformen verfügbar sind. Eine Alternative bietet hier die Arbeit mit bereits bestehenden Metadaten und Audio-Features, wie sie bspw. über die Spotify Web-API[5] verfügbar sind (vgl. Abb. 3). So können Songs aus unterschiedlichen Spielen etwa anhand ihres Tempos, ihrer Tanzbarkeit (*danceability*) und ihres Sentiment-Werts (*valence*) verglichen werden.

	Tempo	**Danceability** (1=High, 0=Low)	**Valence (Mood)** (1=Positive, 0=Negative)
Super Mario Bros.	100.014 bpm	0.768	0.975
Main Theme of Final Fantasy	84.015 bpm	0.743	0.445
Snake Eater (Metal Gear Solid 3)	179.562 bpm	0.267	0.272

Abbildung 3: Drei beispielhafte Audio-Features für drei Computerspiele, die über die Spotify-API extrahiert wurden. Das Beispiel ist Teil einer Studie von Nezar Rajeh im Rahmen des Seminars "Computational Game Studies" im Sommersemester 2023, Seminarleitung Manuel Burghardt.

Alle bislang gezeigten Features könnten künftig in musik-stilometrische (vgl. Backer & Van Kranenburg, 2005) Analysen von Game Music einfließen. Ein weiterer Zugang zur musikwissenschaftlichen Analyse von Spielemusik eröffnet sich für bestimmte Teilbereiche zudem direkt über die Code-Ebene der häufig digital geborenen Stücke. Besonders reizvoll sind hier sogenannte Chiptunes, die ihren Ursprung in der Retro-Computing-Szene haben. Für den in den 1980ern sehr populären Heimcomputer Commodore C64 und seinen ikonischen Soundchip SID (Sound Interface Device) (vgl. McAlpine, 2018; Rettinghaus, 2018) etwa findet sich ein umfangreiches Online-Archiv namens High Voltage SID Collection (HVSC)[6], mit 57.576 einzelnen Stücken (Stand 01.12.2023, Version HVSC #79). Diese Stücke können im nativen SID-Format heruntergeladen und dann in emulierten SID-Playern abgespielt werden. Aktuell versuchen wir das SID-Format in ein Standard-Format zu transformieren, um innerhalb des obigen Korpus Muster melodischer Ähnlichkeit (vgl. Velardo, Vallati und Jan, 2016) zu detektieren und so bspw. den Einfluss und Stil einzelner bekannter Komponisten der Chiptune-Szene, wie etwa Rob Hubbard, Chris Hülsbeck oder David Whittaker, zu untersuchen.

Cineludologie und 'Video'-Spiele: Distant Viewing von Games

Als letzte Analyseperspektive im Sinne von Computational Game Studies fokussieren wir in diesem Abschnitt schließlich auf die visuell-ästhetische Ebene von Spielen (vgl. Grotkopp et al., 2019). Um hier die Komplexität von Interaktivität und individuellem Spielerlebnis zu reduzieren, schlagen wir eine Remediatisierung von Computerspielen als Videos vor. Solche Game Videos sind bspw. als Let's Plays (Ackermann, 2017) und als andere Formaten auf Plattformen wie YouTube und Twitch.TV massenhaft verfügbar. Durch die Betrachtung von Computerspielen als 'Video'-Spiele, eröffnet sich eine filmwissenschaftliche Perspektive auf Games, die Raucher (2021) als "cineludisch" bezeichnet, und die gleichzeitig die Anwendung bestehender Tools und Methoden aus dem Bereich des *distant viewing* (Arnold & Tilton, 2019) erlaubt. Im Bereich der Spieleanalyse gibt es hier bislang keine quantitativen Ansätze, wir schlagen aber in Erweiterung der Methodik der visuellen Stilometrie – wie sie bspw. für die Analyse von impressionistischen Gemälden (Hanchao & Hughes, 2011) und Graphic Novels herangezogen wurde (Dunst & Hartel, 2018) – vor, auch visuelle Features aus Spielen zu extrahieren und diese damit stilistisch vergleichbar zu machen. Denkbar ist hier etwa der Einsatz von Deep Learning-Methoden der *object detection* (vgl. Abb. 4) oder des *automatic captioning* (vgl. Abb. 5). Derart visuelle Parameter können dann wiederum herangezogen werden, um stilistisch-ästhetische Gemeinsamkeiten und Unterschiede in Spielen zu untersuchen. Dies kann etwa in Form einer interaktiven Netzwerkanalyse (vgl. Abb. 6) geschehen, welche in einem ersten Schritt visuell ähnliche Frames einzelner

Spiele clustert (Analogie zu filmischen Einstellungen/Szenen) und dann einen Graph modelliert, bei dem die Knoten jene Frame-Cluster sind und die Kanten visuelle Ähnlichkeit abbilden, bspw. über erkannte Objekte und *captions*, oder aber auch andere Parameter wie Farb- und Kontrastwerte.

Abbildung 4: *Object detection* im Nintendo Switch-Spiel “The Legend of Zelda: Breath of the Wild” (2017) mit dem OWL-ViT-Modell (Vision Transformer for Open-World Localization) (Minderer et al., 2021).

```
frame 151: [{'generated_text': 'a screenshot of a man standing in front of a machine'}]
```

Abbildung 5: *Captioning* eines Frames aus dem C64-Spiel “Zak McKracken and the Alien Mindbenders” (1988) mit dem CLIP-Modell (Contrastive Language Image Pre-training) (Radford et al., 2021). Zunächst wird hier ein Mann, der vor einer Art von Maschine steht erkannt. Bemerkenswert ist weiterhin, dass in der automatisch erstellten Caption angenommen wird, dass hier ein Screenshot vorliegt, was vermutlich durch die niedrig aufgelöste Pixelgrafik zu erklären ist, die für Retro Games, wie dem hier vorliegenden, charakteristisch ist.

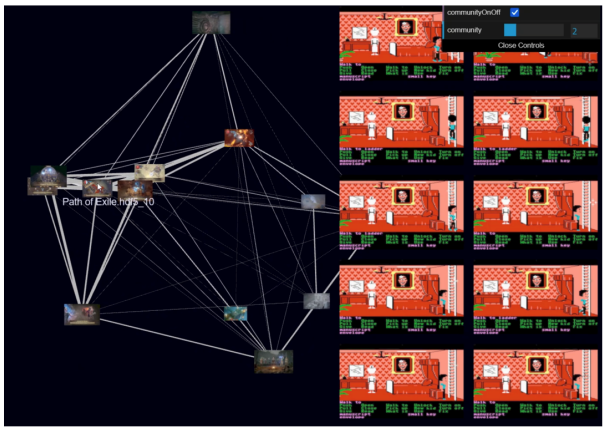


Abbildung 6: Interaktive Netzwerkanalyse (erstellt von Nicolas Ruth, Computational Humanities Group, Universität Leipzig) von 20 beispielhaften Computerspielen anhand deren Game Trailer (3-5 Minuten). In der Abbildung ist ein Teilgraph visuell ähnlicher Frame-Cluster zu sehen. Mit Klick auf einen beliebigen Knoten können die einzelnen Frames, die hier geclustert wurden, inspiziert werden.

Zusammenfassung

Dieser Artikel versucht sich an einem Mapping grundlegender DH-Teilbereiche auf das Forschungsfeld der Game Studies. Dabei fällt auf, dass insbesondere der Teilbereich der Computational Humanities in den Game Studies bislang stark unterrepräsentiert ist. Der Beitrag schlägt drei Annäherungsperspektiven für „Computational Game Studies“ vor, um so einen Dialog in der DH-Community zu starten und künftig weitere Studien in diesem Bereich zu befördern. Eine wesentliche Einschränkung ist dabei zunächst die Fokussierung auf nur drei von insgesamt fünf relevanten semiotischen Modi, wobei eine Formalisierung insbesondere des sog. *procedural modes* (Bogost 2007; Hawreliak, 2018) eine besondere Herausforderung darstellt. Wenngleich Komplexitätsreduktion bei den Analyseparametern und damit eine grundlegende Formalisierung und (Re-)Modellierung der Untersuchungsgegenstände ein wesentliches Merkmal der Computational Humanities ist, so bleibt für komplexe Kulturartefakte wie Computerspiele langfristig die Herausforderung, einer ganzheitlichen, integrierten Analyseperspektive, wie sie vereinzelt schon im Bereich der multimodalen Diskursanalyse vorgeschlagen wurde (Toh, 2018; Wildfeuer & Stamenković, 2022).

Fußnoten

1. Online verfügbar unter https://www.youtube.com/watch?v=ou4kR7FpmOM&ab_channel=Universit%C3%A4tLeipzig
2. https://gi.de/grand-challenges
3. https://gamestudies.org/
4. http://www.digra.org/
5. https://developer.spotify.com/documentation/web-api

6. https://www.hvsc.c64.org/

Bibliographie

Ackermann, Judith (Hg.) 2017. Phänomen Let´s Play-Video. Entstehung, Ästhetik, Aneignung und Faszination aufgezeichneten Computerspielhandelns. Reihe: Neue Perspektiven der Medienästhetik. Springer VS.

Aitkin, Alexander Lewis. 2004. „Playing at Reality: Exploring the Potential of the Digital Game as a Medium for Science Communication“. Application/pdf. PhD Thesis, Canberra: The Australian National University. https://openresearch-repository.anu.edu.au/handle/1885/46051 .

Arnold, Taylor., und Lauren Tilton (2019). Distant viewing: analyzing large visual corpora. Digital Scholarship in the Humanities, 34(Supplement_1), i3-i16.

Aycock, John. 2017. Game Studies at Scale: Towards Facilitating Exploration of Game Corpora. Loading... 10 (17). https://journals.sfu.ca/loading/index.php/loading/article/view/198 .

Backer, Eric, und Peter van Kranenburg. On musical stylometry—a pattern recognition approach. Pattern Recognition Letters 26, no. 3 (2005): 299-309.

Balela, Majed S., und Darren Mundy. 2015. Analysing Cultural Heritage and its Representation in Video Games. In Proceedings of the 2015 DiGRA International Conference. Lüneburg, Germany.

Bateman, John, Janina Wildfeuer, und Tuomo Hiippala. 2017. Multimodality: Foundations, research and analysis – A problem-oriented introduction. De Gruyter.

Bogost, Ian. 2007. Persuasive Games: The Expressive Power of Videogames. The MIT Press. https://doi.org/10.7551/mitpress/5334.001.0001 .

Burghardt, Manuel. 2020. Theorie und Digital Humanities – Eine Bestandsaufnahme. Digital Humanities Theorie (Blog). 15. Mai 2020. https://dhtheorien.hypotheses.org/680 .

Burghardt, Manuel, und Jochen Tiepmar. 2021. The Game Walkthrough Corpus (GWTC) – A Resource for the Analysis of Textual Game Descriptions, 7 (0): 14. https://doi.org/10.5334/johd.34 .

Coavoux, Samuel, Manuel Boutet, und Vinciane Zabban. What we know about games: A scientometric approach to game studies in the 2000s. Games and Culture 12, no. 6 (2017): 563-584.

Dunst, Alexander, und Rita Hartel. 2018. The Quantitative Analysis of Comics: Towards a Visual Stylometry of Graphic Narrative. In Alexander Dunst, Jochen Laubrock & Janina Wildfeuer (Hrsg.): Empirical Comics Research. Digital, Multimodal and Cognitive Methods, S. 43-61. Routledge.

Ensslin, Astrid. 2012. The Language of Gaming. Houndmills, Basingstoke, Hampshire, New York: Palgrave Macmillan.

Fullerton, Tracy. 2020. Surveying the Soul: Creating the World of Walden, a Game. In Mark J.P Wolf: World-Builders on World-Building, S. 93–109. Routledge. https://doi.org/10.4324/9780429242861-6 .

Grotkopp, Matthias, Thomas Scherer, Jasper Stratil, Henning Agt-Rickauer, Christian Hentschel, und Jan-Hendrik Bakels. 2019. Between Data Mining and Human Experience – Digital Approaches to Film, Television and Video Game Analysis. DataverseNL. https://doi.org/10.34894/M2NQMI .

Harkai, István. 2022. Preservation of video games and their role as cultural heritage. Journal of Intellectual Property Law & Practice 17 (10): 844–56. https://doi.org/10.1093/jiplp/jpac090 .

Hawreliak, Jason. 2018. Multimodal semiotics and rhetoric in videogames. Routledge.

Heritage, Frazer. 2020. Applying corpus linguistics to videogame data: Exploring the representation of gender in videogames at a lexical level. Game Studies 20 (3). https://gamestudies.org/2003/articles/heritage_frazer .

Heritage, Frazer. 2022. Magical Women: Representations of Female Characters in the Witcher Video Game Series. Discourse, Context & Media 49 (Oktober): 100627. https://doi.org/10.1016/j.dcm.2022.100627 .

Hutchinson, Rachael. 2021. Observant Play: Colonial Ideology in The Legend of Zelda: Breath of the Wild. The International Journal of Computer Game Research 21 (3). https://gamestudies.org/2103/articles/hutchinson .

Inderst, Rudolf Thomas. 2022. #Gamestudies: 20 Jahre Forschungsfantasie: von der Disziplinierung eines Mediums. Kritische Reflexionen, Band 8. Marburg: Büchner-Verlag, Wissenschaft und Kultur.

Kamp, Michiel, Tim Summers, und Mark Sweeney (Hrsg.). 2016. Ludomusicology: Approaches to Video Game Music. Genre, Music and Sound. https://www.equinoxpub.com/home/ludomusicology/ .

Lipscomb, Scott D., und Sean M. Zehnder. 2004. Immersion in the Virtual Environment: The Effect of a Musical Score on the Video Gaming Experience. Journal of Physiological Anthropology and Applied Human Science 23 (6): 337–43. https://doi.org/10.2114/jpa.23.337 .

McAlpine, K. B. (2018). Bits and pieces: A history of chiptunes. Oxford University Press, USA.

Meredith, David, Hrsg. 2016. Computational Music Analysis. https://link.springer.com/book/10.1007/978-3-319-25931-4 .

Minderer, M., A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, und Z. Shen. 2022. Simple open-vocabulary object detection with vision transformers. arXiv 2022. arXiv preprint arXiv:2205.06230 .

Müller, Meinard. 2021. Fundamentals of Music Processing: Using Python and Jupyter Notebooks. Vol. 2. Cham: Springer.

Ouariachi, Tania, María Dolores Olvera-Lobo, und José Gutiérrez-Pérez. 2017. „Analyzing Climate Change Communication Through Online Games: Development and Application of Validated Criteria“. Science Communication 39 (1): 10–44. https://doi.org/10.1177/1075547016687998 .

Qi, Hanchao, und Shannon Hughes. 2011. A new method for visual stylometry on impressionist paintings. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , S. 2036-2039.

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 International conference on machine learning, pp. 8748-8763. PMLR.

Raucher, Andreas. 2021. „Cineludische Synergien - Game Studies zwischen Spiel und Film“. Paidia. In Paidia – Zeitschrift für Computerspielforschung: https://paidia.de/cineludische-synergien-game-studies-zwischen-spiel-und-film/ .

Rettinghaus, Klaus. 2018. Sidology: Zur Geschichte und Technik des C64-Soundchips. In C. Hust (Hrsg): Digitale Spiele: Interdisziplinäre Perspektiven zu Diskursfeldern, Inszenierung und Musik, S. 269-280.

Roeder, Torsten, und Klaus Rettinghaus. 2020. „Game On! Digitale Archäologie und Edition zu(m) Spielen“. In DHd 2020: Book of Abstracts. Paderborn: Zenodo. https://doi.org/10.5281/zenodo.4621732 .

Roth, Camille. 2019. „Digital, Digitized, and Numerical Humanities“. Digital Scholarship in the Humanities 34 (3): 616–32. https://doi.org/10.1093/llc/fqy057 .

Ryan, J., Eric Kaltman, M. Mateas, und Noah Wardrip-Fruin. 2015. „What We Talk About When We Talk About Games: Bottom-Up Game Studies Using Natural Language Processing“. 2015. https://www.semanticscholar.org/paper/What-We-Talk-About-When-We-Talk-About-Games%3A-Game-Ryan-Kaltman/fa8dda2a6c54a34dc330f9587038b232eec62b8e .

Summers, Tim, und James Hannigan. 2016. Understanding Video Game Music. Cambridge University Press. https://doi.org/10.1017/CBO9781316337851 .

Terras, Melissa. 2011. Peering Inside the Big Tent: Digital Humanities and the Crisis of Inclusion. Melissa Terras: Adventures in Digital Cultural Heritage (blog). 26. Juli 2011. https://melissaterras.org/2011/07/26/peering-inside-the-big-tent-digital-humanities-and-the-crisis-of-inclusion/ .

Thach, Hibby. 2021. A Cross-Game Look at Transgender Representation in Video Games. Press Start 7 (1): 19–44.

Toh, Weimin. 2018. A multimodal approach to video games and the player experience. Routledge.

Unterhuber, Tobias, Hrsg. 2021. Deutschsprachige Game Studies 2011-2021: Eine Bilanz. Sonderausgabe. Paidia.

Velardo, Valerio, Mauro Vallati und Steven Jan. 2016. Symbolic melodic similarity: State of the art and future challenges. In Computer Music Journal 40(2), S. 70-83.

Vosinakis, Spyros, Nikos Avradinis, und Panayiotis Koutsabasis. 2018. Dissemination of Intangible Cultural Heritage Using a Multi-agent Virtual World. Advances in Digital Cultural Heritage, Januar, 197–207. https://doi.org/10.1007/978-3-319-75789-6_14 .

Wallner, Günter, Simone Kriglstein, und Anders Drachen. 2019. Tweeting Your Destiny: Profiling Users in the Twitter Landscape around an Online Game. 2019 IEEE Conference on Games (CoG). https://www.semanticscholar.org/paper/Tweeting-your-Destiny%3A-Profiling-Users-in-the-an-Wallner-Kriglstein/f7749c7c09668245046e9f266f78ce6ef3779532 .

Wang, Xiaohui, und Dion Hoe-Lian Goh. 2020. Components of Game Experience: An Automatic Text Analysis of Online Reviews. Entertainment Computing 33 (März): 100338. https://doi.org/10.1016/j.entcom.2019.100338 .

Wendler, Zoe Ann. 2014. “Who Am I?”: Rhetoric and Narrative Identity in the Portal Series. Games and Culture 9 (5). https://doi.org/10.1177/1555412014543517 .

Wildfeuer, Janina, und Dušan Stamenković. 2022. The discourse structure of video games: A multimodal discourse semantics approach to game tutorials. Language & Communication 82: 28-51.

Cross-Linguistic Data Formats (CLDF): D’où Venons Nous? Que Sommes Nous? Où Allons Nous?

Forkel, Robert

robert_forkel@eva.mpg.de
Max-Planck-Institut für Evolutionäre Anthropologie, Deutschland

List, Johann-Mattis

Mattis.List@uni-passau.de
Universität Passau; Max-Planck-Institut für Evolutionäre Anthropologie, Deutschland

Im Jahre 2024 blicken wir auf 10 bewegte Jahre zurück, in denen wir uns nun dafür eingesetzt haben sprachübergreifende Datenformate in der vergleichenden und typologischen Sprachforschung als einen Standard zu etablieren, der Kuratierung, Analyse, und das Teilen von Forschungsdaten im Bereich der digitalen Linguistik erleichtert. Diese Standardisierung erschließt große Datensätze nicht nur für die sprachwissenschaftliche orientierte Forschung im Speziellen, sondern auch für die digital orientierte geisteswissenschaftliche Forschung im Allgemeinen.

Etablierung von Standardformaten in den DH

Die “Cross-Linguistic Data Formats”-Initiative (CLDF, https://cldf.clld.org) – versucht, ein Standardformat für Daten in der Vergleichenden Sprachforschung zu etablieren, um die Entwicklung von Methoden im Bereich der sprachlichen Diversitätsforschung von der Pflege einzelner Datensätze zu entkoppeln. Wie in anderen Bereichen der Digital Humanities befindet man sich hier in der Zwickmühle zwischen “zu simpel” und “zu kompliziert”: generalisierbare “gut genug”-Lösungen zu finden erweist sich als schwierig. Zwar konnten Datensammlungen aus den Digital Humanities in den letzten Jahren oft durch immer bessere Webapplikationen deutlich an Sichtbarkeit gewinnen (im Bereich der Sprachtypologie prominent vertreten durch Projekte wie den World Atlas of Language Structures Online, Dryer and Haspelmath 2013, https://wals.info), zugleich bedeutete aber die Kopplung zwischen Daten und ihrer Präsentation im Web eine Abhängigkeit von den kurzen *Hype Cycles* der Webtechnologien, die (verstärkt durch die kurzen Beschäftigungszyklen im akademischen Bereich) einer Etablierung von Standards durch langfristige *Communities of Practice* entgegenwirkten.

Bereits 2014 entwickelte sich aus einem Publikationsprojekt für Datensammlungen die Idee, aus den applikationsspezifischen Datenbanken von Webapplikationen ein gemeinsames Datenmodell zu extrahieren, dass die Spezifikation einer Datenzugriffs- *API* erlaubt. Dank ausreichend langer Förderung durch die Max-Planck-Gesellschaft und den Europäischen Forschungsrat (im Rahmen mehrerer Projekte über 10 Jahre hinweg) konnte diese Idee Realität werden. Gesicherte Förderung bis mindestens Ende 2027 wird uns helfen, den CLDF Standard weiter zu etablieren und insbesondere ein geeignetes Modell für die weitere Pflege des Standards durch die Nutzergemeinde zu finden.

Woher kommen wir?

Auch für sprachübergreifende Datensätze galt lange Zeit: Die Webapplikation **ist** die Datenbank. Zugang zu Daten und Pflege von Daten erfolgte also durch die Applikation (“through the web”), und die Frage der Generalisierbarkeit von Lösungen wurde in erster Linie als Suche nach einem Content-Management-System verstanden, das idealerweise nicht nur das Kuratieren und Teilen von sprachübergreifenden Daten erleichtern, sondern auch der gleichzeitigen Publikation dienen sollte. Das im CLLD Projekt entwickelte clld-Framework stellte einen weiteren Schritt in diese Richtung dar. Zwar konnten mehrere bekannte Datensätze (etwa WALS, https://wals.info, APiCS, https://apics-online.info, und Glottolog, https://glottolog.org) durch CLLD-Applikationen im Web zugänglich gemacht werden, aber dennoch stellte sich im Laufe des Projekts heraus, dass – insbesondere etablierte Datensammlungen – typischerweise schon etablierte Kurationsworkflows aufwiesen, und sich die Aufgabe der Webapplikation damit auf Konsistenz-Prüfung und Datenpublikation reduzierte. Diese Funktionseinschränkung war zwar ein Dämpfer für die Hoffnung auf ein sprachübergreifendes CMS, erwies sich aber als essentiell für das Verstehen der Schnittstellen zwischen Datenpflege, Datenpublikation und Datenverwendung – also CLDF.

Während sprachübergreifende Daten auch in einer großen Bandbreite von Komplexität auftreten – von einfachen Wortlisten zu syntaktischen Dependency-Trees (vgl. Marneffe et al. 2021, https://universaldependencies.org) – gibt es doch mittlerweile einige Analyse-Verfahren, die klar umrissene, einfache Daten als Grundlage haben, wie z. B. phylogenetische Methoden, die auf kognatenkodierten Wortlisten basieren (Tresoldi et al. 2022), oder typologisch-quantitative Methoden, die mit Daten typologischer Surveys arbeiten (Blasi et al. 2019). Eine Fokussierung auf den *Input* etablierter Methoden – ohne auf volle Abdeckung des Spektrums an möglichen Daten abzuzielen – versprach deshalb einen gangbaren Weg zur Etablierung von standardisierten Datenformaten.

Bei der ersten DHd Konferenz im Jahr 2014 in Passau, wurde das CLLD Projekt mit einem Poster vorgestellt. Mit WALS und Glottolog gab es zu diesem Zeitpunkt schon prominente Datenbanken, die sprachübergreifende Daten als CLLD-Applikation zugänglich machten. Das Poster beschreibt CLLD-Applikationen als Plattform, um eine “emerging API between data publications and tools” zu “entdecken”. In den 10 Jahren seitdem konnte diese Chance erfreulicherweise genutzt werden. Mit CLDF konnte das CLLD-Datenmodell in einen plattformunabhängigen Standard überführt werden, der mittlerweile die computergestützte Forschung in vielen Bereichen der digital orientierten Vergleichenden Sprachwissenschaften erleichtert – oder erst möglich macht. Einige Pläne der ersten Jahre stellten sich aber auch als Irrwege heraus, die dennoch vielleicht als generalisierbare Lehren für Standardisierungsbemühungen in anderen Wissenschaftssparten dienen können.

Wer sind wir?

Nach fast 10 Jahren Arbeit an CLDF und 5 Jahren seit Publikation der ersten Version des Standards (Forkel et al. 2018) formt sich nach und nach das Bild eines Datenökosystems, in dem tatsächlich einige Probleme, die früher allgegenwärtig waren, verschwunden sind.

CLDF ist eine Spezifikation für ein Paketformat für sprachübergreifende Datensätze, die inzwischen eine Vielzahl von regelmäßig wiederkehrenden Datentypen abdeckt, die von Wortlisten (https://lexibank.clld.org, List et al. 2022) über tabellarische Sammlungen sprachtypologischer Merkmale (https://grambank.clld.org, Skirgård et al. 2023) bis hin zu interlinearglossiertem Text reicht (vgl. https://apics-online.org, Michaelis et al. 2013) und auch eine Integration verschiedener Datentypen ermöglicht (vgl. https://tppsr.clld.org, Geisler et al. 2021).

Da viele der quantitativ genutzten Daten auch in der Linguistik im Grunde tabular sind, bot es sich an, die W3C recommendation für "CSV on the Web (CSVW)" (https://csvw.org, Gower 2021) zur Grundlage zu nehmen, um ein relationales Datenmodell zu formulieren.

Wie oben erwähnt folgt CLDF dem Designprinzip, nur solche Aspekte der Daten zu standardisieren, die eine durch Analysemethoden klar spezifizierte Semantik haben. "Expressive adequacy" im (umfassenden) Sinn von Ide et al. (2003) wird also nicht angestrebt. Spielraum für erweiterte Anwendbarkeit und damit Raum zum Erproben von Kandidaten für weitere Standardisierung wird in CLDF durch den zugrundeliegenden CSVW Standard erreicht: Jeder CLDF Datensatz besteht aus tabularen Daten, die als CSV Dateien serialisiert sind, und deren Zusammenhang durch eine Metadatendatei CSVW-konform bescrieben wird. Damit können CLDF Datensätze transparent durch zusätzliche Tabellen und Spalten erweitert werden, für die weiterhin "technische" Metadaten per CSVW spezifiziert werden können, die aber (noch) keine Rolle im CLDF Datenmodell spielen. Eine zu myopische Entwicklung des Standards war dennoch nicht zu befürchten, da eine pragmatische Adäquatheit schon allein dadurch gewährleistet war, dass unter den Datensätzen, die als *proof-of-concept* in CLDF umgewandelt wurden, "Flaggschiffe" wie WALS und Glottolog waren – also genau solche Datensätze, die auch schon vor CLDF große Nachnutzung erfuhren.

Zwar adressiert ein Datenformat wie CLDF nur die I und R Komponente der FAIR Prinzipien nach Wilkinson et al. (2016), also *interoperability* und *reusability* (während *findability* und *accessibility* von anderen Aspekten abhängen, unter anderem von standardisierten **Meta**daten – die immerhin transparent in CSVW integriert werden können), aber wenn interoperable Daten erst einmal prinzipiell vorhanden sind (also erzeugt und in Publikationen verwendet werden), lassen sich F(indability) und A(ccessibility) wesentlich leichter als erstrebenswerte Ziele im wissenschaftlichen Betrieb verankern.

CLDF und Interoperability : CLDF addressiert das Interoperabilitätsproblem, indem das generische CSVW-Datenmodell mit linguistischer Semantik verknüpft wird. Da CSVW bereits eine Integration mit *Linked Data* vorsieht, kann das durch eine CLDF-Ontologie erfolgen. *Linked Data* erlaubt nicht nur das Einbinden von Ontologien, sondern auch das Referenzieren externer Objekte. Kuratierte Kataloge solcher Referenzobjekte bilden den zweiten Beitrag von CLDF zur Interoperabilität von Sprachdaten. So kann beispielsweise das Problem der Sprachidentifikation transparent gelöst werden, indem Datensätze auf Sprachvarietäten via Glottolog verweisen, den bisher größten Referenzkatalog für Sprachen, der mehr als 8000 historische und kontemporäre Sprachvarietäten auflistet und systematisch mit verschiedenen Daten, wie Referenzen und Klassifikationen verlinkt (Forkel and Hammarström 2022). Mit dieser Blaupause können aber auch weitere Referenzkataloge operationalisiert werden, wie zum Beispiel das *Concepticon* (https://concepticon.clld.org, List et al. 2023), als Katalog elizitierter Wortbedeutungen in Wortlisten, oder CLTS (Cross-Linguistic Transcription Systems, https://clts.clld.org, List et al. 2021) als Katalog möglicher Sprachlaute.

CLDF und Reusability : CLDF, wie auch schon CSVW, ermöglicht es, Metadaten mit Datensätzen zu verknüpfen. Wird dieser Mechanismus genutzt, um Lizenz oder Provenienz von Daten zu beschreiben, kann damit die Grundlage für die Weiterverwendung (*data re-use*) gelegt werden. CLDF erhöht aber die Weiterverwendbarkeit von Daten auch auf andere Weise: Standardformate wie CLDF erlauben insbesondere die Entkopplung von Datensamlung und -pflege auf der einen Seite und die Weiterentwicklung von Analysemethoden auf der anderen Seite. Das bedeutet, dass CLDF-Datensätze im Laufe der Zeit immer "besser" werden – ohne dass sich die Datensätze noch ändern. Diesen Effekt sieht man besonders einfach bei der Weiterentwicklung von Visualisierungs- und Clustermethoden. So können Daten in der *Database of Cross-Linguistic Colexifications* (CLICS, https://clics.clld.org, Rzymski et al. 2020), welche Konzeptnetzwerke aus sprachübergreifenden Daten extrahiert und interaktiv visualisiert, nicht nur mit jedem neuen Clusteralgorithmus neu visualisiert werden, sondern sie können auch trotz gleichbleibender Daten aufgrund neuer Verfahren zur Datenanalyse zur Beantwortung verschiedenster Fragestellungen verwendet werden. So könnte zum Beispiel die neue Methode zur Erstellung von Konzeptnetzerken aus Wörtern, die nur in Teilen Ähnlichkeiten aufweisen (List 2023), ohne Probleme auf die bestehenden CLICS-Daten angewendet werden, um diese Konzeptnetzwerke um eine weitere Dimension von konzeptueller Ähnlichkeit zu erweitern.

Ein weiteres Designprizip von CLDF besteht darin, einfacher Datennutzung einen höheren Stellenwert einzuräumen als einfacher Datenproduktion. Während von einigen Designentscheidungen beide Nutzungsaspekte profitieren – etwa von der Beschränkung auf zeilenbasierte Textformate, die eine effektive Versionskontrolle mithilfe von tools wie Git (https://git-scm.com/) erlauben – wurden andere Entscheidungen klar zu Lasten zukünftiger Datenproduzenten getroffen. Die Erfordernis von referenzieller Integrität innerhalb eines Datensatzes etwa bietet für die Datennachnutzung eine zusätzliche Garantie, während ma in der Datenproduktion auf das Vermeiden von "broken links" – wie sie etwa in HTML erlaubt sind – achten muss.

Dieser Nachteil wiegt allerdings weniger schwer, da – wie oben breits ausgeführt – viele relevante Datensammlungen sowieso über eigene Kurationsworkflows verfügen, d.h. eine Pflege der Daten in CLDF kaum stattfindet und eine Bereitstellung der Daten als CLDF in jedem Fall eine Konvertierung erfordert. Umgekehrt wird es durch ein spezifiziertes Ziel-format wie CLDF möglich, diese Konvertierung mit tools zu unterstützen.

Von den mittlerweile mehr als 200 CLDF Datensätzen, die auf Zenodo archiviert wurden (siehe https://zenodo.org/search?q=keywords%3A%2Fcldf%3A%28Wordlist%7CStructureDataset%7CDictionary%7CGeneric%29%2F), werden denn auch viele per

CldfBench (https://pypi.org/project/cldfbench, Forkel and List 2020) gepflegt. Dieser zusätzliche Aufwand bei der Datenpublikation wird aber spätestens dann wettgemacht, wenn ein Datensatz über längere Zeit gepflegt wird. Sobald sich nämlich nicht nur der Datensatz selbst ändert, sondern auch die Referenzkataloge, erweist sich die Konvertierung in CLDF als hilfreich, um möglichen Synchronisierungsbedarf zu signalisieren, oder im besten Fall eine automatische Synchronisierung durchzuführen.

Wo wollen wir hin?

Die oben beschriebene Kombination aus spezifiziertem Zielformat und flexiblen Konvertierungswerkzeugen eröffnet ein grosses Potential für "neue" Datensätze via "retro-digitalisierung". Nachdem etwa mit dem Linguistic Survey of India bereits nachgewiesen wurde, dass der Weg von der Buchseite zu CLDF gangbar ist, liest sich die mehr als 8000 Einträge umfassende Liste von bibliographischen Referenzen mit dem Schlagwort "wordlist"=" auf Glottolog (siehe https://glottolog.org/langdoc?sEcho=2&sSearch_7=wordlist) wie eine operationalisierbare Aufgabenliste.

Aber auch für noch nicht gesammelte Daten verbessert CLDF die Ausgangslage. So kann wesentlich einfacher ermittelt werden, welche Daten zu welchen Sprachen potentiell von Interesse sind. Eine effektive Katalogisierung der vorhandenen Datensätze erlaubt dann wesentlich gezielteren Einsatz von Mitteln (siehe etwa "Glottobank – ELDP small grants" https://www.eva.mpg.de/linguistic-and-cultural-evolution/events/2023-glottobank-eldp-small-grants/).

Eher ein zufälliger Nebeneffekt der Technologiewahl im CLDF-Standard könnte das größte Potential für die weitere Etablierung von CLDF bieten: Da CSVW – der CLDF zugrundeliegende W3C Standard – im Grunde ein Serialisierungsformat für relationale Daten darstellt, lassen sich alle CSVW-Datensätze (also auch alle CLDF-Datensätze) algorithmisch und klar definiert in relationale Datenbanksysteme importieren. Insbesondere SQLite (https://sqlite.org) – mit seiner Kombination aus Einfachheit, ubiquitärer Verfügbarkeit und Performanz – macht diese Eigenschaft zu einem bisher grob unterschätzten "Killer Feature" für zukünftige Anwendungen. Per SQL kann der Zugriff auf große Datenmengen oft um ein Vielfaches beschleunigt werden, wodurch sowohl explorative Datenanalyse als auch die Replizierbarkeit von publizierten Analysen wesentlich vereinfacht werden. Dies wird beispielsweise auch in der auf dem Lexibank-Repositorium (die bisher größte Sammlung sprachübergreifender Wortlisten, vgl. List et al. 2022) basierenden Webapplikation deutlich (https://lexibank.clld.org). Aufgrund von vorberechneten abstrakten semantischen und phonetischen Repräsentationen, die alle vom CLDF-Standard abgedeckt werden, wird es möglich, in sekundenschnelle Hunderttausende von Wörtern auf phonetische und semantische Ähnlichkeit zu vergleichen (List et al. 2023). CLDF liefert hier die Vergleichbarkeit, SQL die Ausdrucksmöglichkeiten für den Vergleichsalgorithmus und PostgreSQL die performante Implementierung.

Im Idealfall wären bereits ausreichende Nutzbarkeit und Nutzung Garant für die Zukunft von CLDF. Um in der Forschung relevant zu bleiben, muss realistischerweise aber auch Anpassbarkeit gewährleistet sein. Aus technologischer Sicht sind die Voraussetzungen dafür gegeben. Aus organisatorischer Sicht muss sich aber noch erweisen, ob das Modell von *open source software development* auf Dauer geeignet ist, die Entwicklung eines Standards zu steuern. Die nächsten 4 Jahre, in denen wir noch auf gesicherte Förderung zurückgreifen können, sollten aber hoffentlich ausreichen, ein Steuerungsmodell für die weitere Zukunft auf den Weg zu bringen.

Schlussbetrachtung

Mit unserer Arbeit an CLDF schauen wir auf nunmehr fast 10 Jahre zurück, in denen wir uns dafür eingesetzt haben, die Welt der Daten in der digital orientierten Vergleichenden Sprachforschung ein wenig vergleichbarer zu machen. Auch wenn wir auf dieser Reise viele ursprünglich anvisierte Meilensteine passieren konnten, sind wir längst noch nicht da angekommen, wo wir ursprünglich hinwollten. Wir denken jedoch, dass die Geschichte dieses sprachübergreifenden Standards, den wir nun schon so lange pflegen und ständig weiter modifizieren und ausbauen, nicht nur für einen linguistischen Forschungskreis interessant ist, sondern dass es sich auch lohnt, sie im größeren Kontext der Digitalen Geisteswissenschaften zu erzählen.

Ob die Entwicklung von "kleinen" Standardisierungsbemühungen wie CLDF letztendlich Teil einer größeren Lösung wird, die auch komplexere Datentypen und größere Wissenschaftsbereiche umfasst (vgl. etwa die Bemühungen des Data for History Consortiums bzgl. *geo-historical data*, http://dataforhistory.org) oder vielleicht womöglich größere Lösungen erschwert, kann nur die Zukunft zeigen. Aber auch wenn sich der momentane Höhenflug von CSV (vgl. https://csvconf.com) nur als letztes Aufflackern eines verlöschenden Lichts herausstellen sollte, bleibt uns noch die Integration in die große Idee des Semantic Web per CSVW.

Bibliographie

Blasi, Damián E., Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu, and Balthasar Bickel. 2019. "Human Sound Systems Are Shaped by Post-Neolithic Changes in Bite Configuration." *Science* 363 (1192): 1–10. https://doi.org/10.1126/science.aav3218.

Dryer, Matthew S. and Martin Haspelmath. 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Forkel, Robert and Harald Hammarström. 2022. "Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information."

Semantic Web 13 (6): 917–24. https://doi.org/10.3233/sw-212843.

Forkel, Robert and Johann-Mattis List. 2020. "CLDFBench: Give Your Cross-Linguistic Data a Lift." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6995–7002. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. "Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics." *Scientific Data* 5 (180205): 1–10. https://doi.org/10.1038/sdata.2018.205.

Geisler, Hans, Robert Forkel, and Johann-Mattis List. 2021. "A Digital, Retro-Standardized Edition of the Tableaux Phonétiques Des Patois Suisses Romands (TPPSR)." In *Nouveaux Regards Sur La Variation Dialectale*, edited by M. Avanzi, N. LoVecchio, A. Millour, and A. Thibault, 13–36. Strasbourg: Éditions de Linguistique et de Philologie. https://tppsr.clld.org.

Gower, Robin. 2021. *CSV on the Web*. Stirling: Swirrl. https://csvw.org.

Ide, Nancy, Laurent Romary, and Eric de la Clergerie. 2003. "International standard for a linguistic annotation framework." In *Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language technology systems*. 25–30. https://doi.org/10.3115/1119226.1119230.

Jackson, Joshua Conrad, Joseph Watts, Teague R. Henry, Johann-Mattis List, Peter J. Mucha, Robert Forkel, Simon J. Greenhill, Russell D. Gray, and Kristen Lindquist. 2019. "Emotion Semantics Show Both Cultural Variation and Universal Structure." *Science* 366 (6472): 1517–22. https://doi.org/10.1126/science.aaw8160.

List, Johann-Mattis. 2023. "Inference of Partial Colexifications from Multilingual Wordlists." *Frontiers in Psychology* 14 (1156540): 1–10. https://doi.org/10.3389/fpsyg.2023.1156540.

List, Johann-Mattis, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems* [Dataset, Version 2.2.0]. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.3515744.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell Gray. 2023. *Lexibank Analysed* [Dataset, Version 1.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.7836668.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. "Lexibank, a Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features." *Scientific Data* 9 (316): 1–31. https://doi.org/10.1038/s41597-022-01432-0.

List, Johann-Mattis, Annika Tjuka, Mathilda van Zantwijk, Frederic Blum, Carlos Barrientos Ugarte, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel. 2023. *CLLD Concepticon* [Dataset, Version 3.1.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org/.

Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. "Universal Dependencies." *Computational Linguistics* 47 (2): 255–308. https://doi.org/10.1162/coli_a_00402.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath, and Magnus Huber, 2013. *APiCS Online*. Jena: Max Planck Institute for the Science of Human History. https://apics-online.info/.

Rzymski, Christoph, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2020. "The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross- Linguistic Polysemies." *Scientific Data* 7 (13): 1–12. https://doi.org/10.1038/s41597-019-0341-x.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G.~A. Kowalik, Olga Krasnoukhova, Nora L.~M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.~C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera,

Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. "Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss." *Science Advances* 9 (16). https://doi.org/10.1126/sciadv.adg6175.

Tresoldi, Tiago, Christoph Rzymski, Robert Forkel, Simon Greenhill, Johann-Mattis List, and Russell D. Gray. 2022. "Managing Historical Linguistic Data for Computational Phylogenetics and Computer-Assisted Language Comparison [with Accompanying Tutorial]." In *Open Handbook of Linguistic Data Management*, edited by Andrea Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, 345–54. Massachusetts: MIT Press. https://doi.org/10.7551/mitpress/12200.001.0001.

Wilkinson, Mark D., Michel Dumontier, Ilsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz B. da Silva Santos, Philip E. Bourne. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 1–9. https://doi.org/10.1038/sdata.2016.18.

Cultura Ibi Vadis! Zur Rekontextualisierung und Visualisierung kultureller Informationen in InTaVia

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

Schlögl, Matthias
matthias.schloegl@oeaw.ac.at
ACDH-CH, Österreichische Akademie der Wissenschaften, Österreich

Windhager, Florian
florian.windhager@donau-uni.ac.at
Universität für Weiterbildung Krems, Österreich
ORCID: 0000-0002-5170-2243

Einleitung

In den letzten Jahrzehnten wurde die Digitalisierung kultureller Informationen in Europa deutlich vorangetrieben. So wurden die Digitalisate kultureller Objekte aus Museen, Archiven und Bibliotheken auf lokalen, nationalen und europäischen Plattformen einer breiten Öffentlichkeit und der Wissenschaft zugänglich gemacht. Gleichzeitig wurde davon unabhängig kulturelle Informationen digital erfasst und in Datenbanken zur Verfügung gestellt; z.B. Wissen über bedeutende Persönlichkeiten liegt in strukturierter Form in nationalen prosopographischen Datenbanken vor. Diese Entwicklungen bieten eine gute Basis, um den Blick auf digitale kulturelle Informationen auszuweiten und so ein besseres Verständnis unseres kulturellen Erbes zu ermöglichen. Jedoch erschweren fehlende Verknüpfungen (zwischen biographischen und Objektdatenbanken) und Standardisierungen (zwischen verschiedenen Datenbanken) sowie mangelnde (Maschinen-)Lesbarkeit und Sichtbarkeit lokaler Datensammlungen deren Nutzung – für die wissenschaftliche Tätigkeit von Expert:innen ebenso wie für eine intuitive Exploration und ein besseres Verständnis von kulturhistorischen Themen durch die interessierte Öffentlichkeit.

Das H2020-Projekt InTaVia (*In/Tangible Cultural Heritage: Visual Analysis, Communication and Curation*, https://intavia.eu, 2020-2023) verfolgte das Ziel, digitalisierte kulturelle Objekte und biographische Informationen zu verknüpfen und so eine synoptische Betrachtung von Leben und Werken der europäischen Kulturgeschichte zu ermöglichen. Das Konsortium harmonisierte und integrierte zu diesem Zweck nationale biographische Datenbanken aus Slowenien, Österreich, den Niederlanden und Finnland und verknüpfte diese mit kulturellen Objekten aus Europeana und Wikidata, die mit diesen Akteuren in Zusammenhang stehen, in einem größeren kulturellen Wissensgraphen. Um den Zugang zu diesen reichhaltigen Informationen zu ermöglichen, entwickelte das Projekt ein prototypisches Informationsportal für die visuelle Analyse, Kuratierung und Kommunikation von diesen Kulturdaten auf multiplen Ebenen der Aggregation (vgl. Abbildung 1).

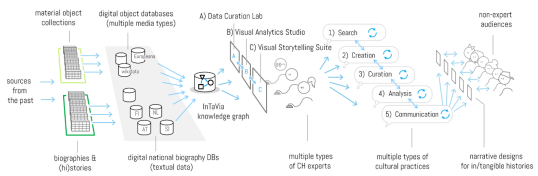


Abbildung 1. Architektur des Projektes InTaVia: digitalisierte Objekt- und Personendaten werden im InTaVia Wissensgraphen integriert, den Kulturerbe-Expert:innen und -Forscher:innen in einem Interface für die Suche, Kuratierung, visuelle Analyse und Kommunikation an die Öffentlichkeit nutzen können.

Dieser Beitrag stellt die Methoden und Ergebnisse dieses europäischen Projektes vor und reflektiert die Implikationen für die Digitalisierung, Verknüpfung und Zugänglich-

machung von (im)materiellem Kulturerbe. Dazu stellen wir zunächst den InTaVia Knowledge Graph vor (Abschnitt 2), bevor wir auf die Module zur visuellen Analyse (Abschnitt 3) und Storytelling (Abschnitt 4) eingehen. Abschließend diskutieren wir in Abschnitt 5 Herausforderungen und Implikationen für den Umgang mit digitalen Kulturerbedaten.

Der InTaVia Knowledge Graph

Als zentrale Wissensbasis wurde der InTaVia Knowledge Graph (IKG) aufgebaut, in dem die Daten der unterschiedlichen Biographie- und Objektdatenbanken zusammengeführt werden und dabei die Provenienz der Daten gewahrt bleiben. Daneben erlaubt der IKG aber auch in technisch relativ einfacher und verständlicher Form die Abfrage, automatische Anreicherung, Harmonisierung und Verlinkung der Daten.

Für möglichst große Flexibilität wurde ein Triplestore als Datenbanklösung eingesetzt. Als Ontologie entwickelten wir IDM-RDF (https://github.com/InTaVia/idm-rdf), das auf CIDOC CRM (Doerrl, 2003) in der Version 7.1.1 und der BioCRM Extension (Tuominen et al., 2017) aufbaut und für das Verlinken der einzelnen Entitäten eine adaptierte Version des Proxy Modells aus der Object Reuse and Exchange Ontologie (http://www.openarchives.org/ore/1.0/datamodel) implementiert. Für einen vereinfachten Zugang zu den Daten stellt der IKG eine RestAPI zur Verfügung (https://intavia-backend.acdh-dev.oeaw.ac.at/v2/docs). Für die Anreicherung, Harmonisierung und Verlinkung der Daten setzt der IKG auf ETL pipelines, die in einem Kubernetes Cluster laufen. Verschiedene Prefect v1 (https://docs-v1.prefect.io/) Pipelines erlauben den Import und die Aktualisierung von Personen-Entitäten, sowie deren Konvertierung und Anreicherung mit Objekt-Entitäten. Zur Verbesserung dieses Workflows sollen Daten zukünftig (mittels SHACL) validiert werden, bevor diese in den Triplestore geladen werden.

Im IKG finden sich zum Projektende ca. 25 Millionen Triples, die mehr als 250.000 Personen-Proxies (es kann mehrere Proxies geben, die dieselbe Person beschreiben), knapp 35.000 Orts-Proxies und rund 390.000 Proxies zu Objekten des kulturellen Erbes beschreiben und zueinander in Beziehung setzen (vgl. Abbildung 2).

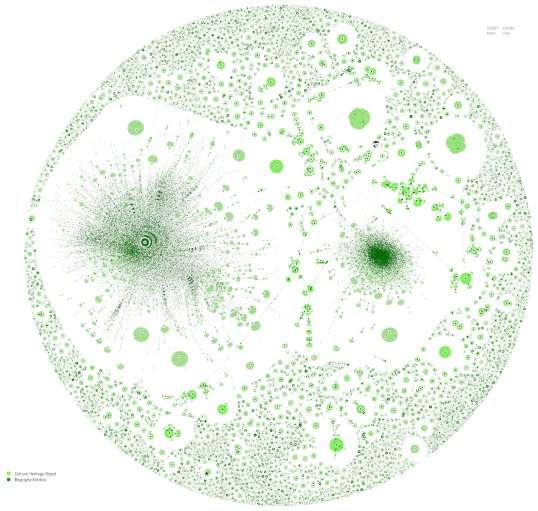


Abbildung 2. InTaVia Knowledge Graph (IKG) mit kulturellen Objekten (hellgrün) und Personen (dunkelgrün), erstellt mit cosmograph.

Visuelle Analyse von kulturellen Informationen

Die im IKG gespeicherten Daten zu kulturellen Objekten und Biografien haben eine Vielzahl von Facetten und Dimensionen, die für Historiker:innen und Kulturwissenschaftler:innen von Interesse sein können und mittels visueller Analyse in neuer Form erschlossen werden können (Windhager et al., 2018, 2022): die geografische Position von biografischen oder künstlerischen Ereignissen, diverse Ereignisse oder kulturelle Entitäten (Objekte oder Personen), Relationen zwischen Personen und/oder Objekten sowie chronologische Abfolgen von Ereignissen. Diese Aspekte können auf verschiedenen Ebenen der Aggregation - von historischen Individuen bis hin zu diversen Gruppierungen - für verschiedene Fragestellungen von Relevanz sein. Das Visualisierungsmodul im InTaVia Frontend (Mayr et al., 2022) ermöglicht close und distant reading von Objekt- und Personendaten aus dem IKG. Abbildung 3 zeigt das multiperspektivische Interface zur visuellen Analyse von Leben und Werk anhand eines Beispiels zur niederländischen Reise Albrecht Dürers (Grebe & Großmann, 2013; Windhager et al., 2023): Es enthält eine Liste von Ereignissen (links), zwei Karten mit zeitlich enkodierter Trajektorie (oben links) und mit zu Orten zugeordneten Objekten (oben rechts), sowie eine Zeitleiste, die einen Überblick über Personen- und Objektereignisse im Leben von Dürer bietet (unten). Bewegt ein:e Benutzer:in die Maus über ein Ereignis, wird dieses auch in den anderen Ansichten hervorgehoben und eine Detailvorschau angezeigt. Ein Klick auf das Objekt öffnet eine Detailansicht mit weiteren Informationen und Visualisierungen.

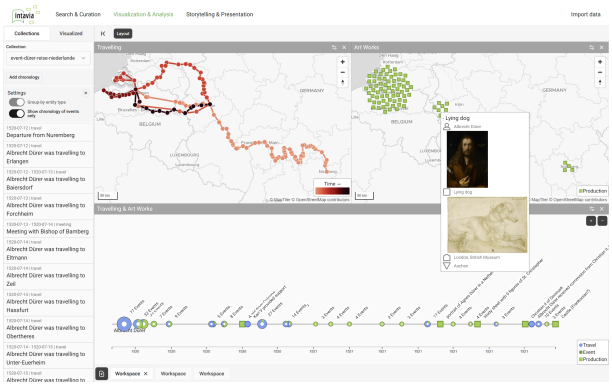
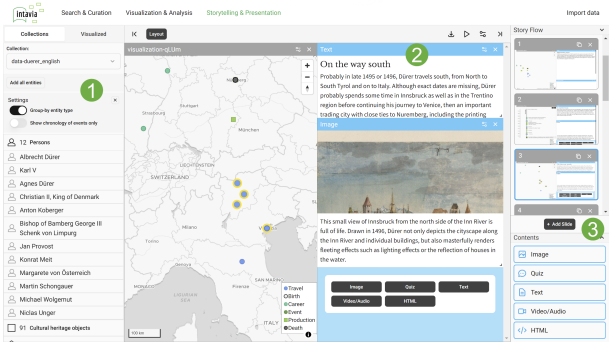


Abbildung 3. Multiperspektivisches Interface zur visuellen Analyse von Leben und Werk.

Storytelling mit kultureller Information

Narrative Techniken des visualisierungsbasierten Storytellings eröffnen neue, niederschwelligere Zugänge zu Kulturerbedaten - im Speziellen für interessierte Laien und die breite Öffentlichkeit (Kusnick et al., 2021). Jedoch ist die Erstellung solcher daten- und visualisierungsbasierten Geschichten ohne geeignete Werkzeuge sehr aufwändig. Die InTaVia Plattform bietet daher die Möglichkeit ausgewählte Daten aus dem IKG und Visualisierungen aus dem Visualisierungsmodul mit erläuternden Texten, Medien und interaktiven Elementen anzureichern und narrativ zu verknüpfen (Liem et al., 2023; vgl. Abbildung 4). Beim Gestalten der Geschichte entscheiden die Expert:innen, welche Visualisierungen genutzt werden sollen, welche Ereignisdaten darin hervorgehoben werden sollen und welche somit vom Zielpublikum interaktiv exploriert werden können. Übergänge zwischen einzelnen Stationen der Geschichte auf Karten oder Zeitleisten werden animiert und so intuitiv erfahrbar gemacht. Die fertiggestellten Geschichten können über einen URL geteilt und am Browser, sowie auf mobilen Endgeräten betrachtet werden.



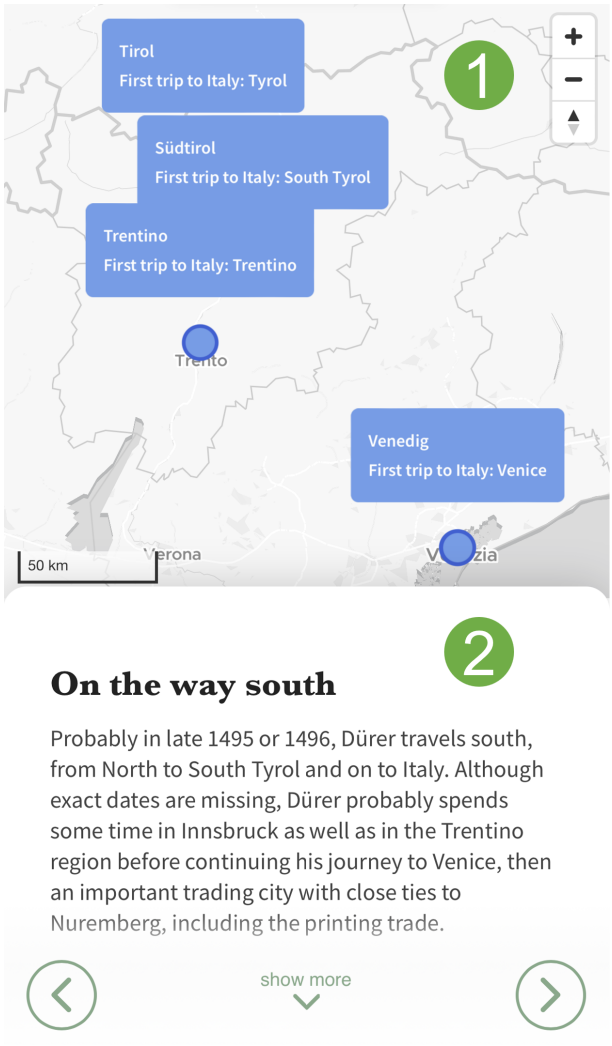


Abbildung 4. Erstellung einer Geschichte mit Karte, Text und Bild (links: (1) Entitäten, (2) Vorschau einer Station in der Geschichte, (3) Übersicht über die Stationen und möglichen Inhalte) und mobile Ansicht der fertigen Story (rechts: (1) Karte, (2) Text).

Digitales Kulturerbe - Quo Vadis?

Was können wir aus InTaVia für digitales Kulturerbe lernen, was sind die Herausforderungen und nächsten Schritte? Die Vision von InTaVia war es, durch die Verknüpfung von verschiedenen digitalisierten kulturellen Informationen ein reichhaltigeres Bild unseres kulturellen Erbes zu eröffnen. In der Umsetzung hat sich jedoch gezeigt, dass diese Idee wohl noch länger eine Vision bleiben wird, da zuvor noch einige Herausforderungen bewältigt werden müssen:

(1) Viele kulturelle Datenbanken sind (noch nicht) bereit für diese Ansätze: Sowohl bei kulturellen Objekten als auch bei Biographien sind diese oft noch zu wenig strukturiert, um diese untereinander zu verknüpfen und zu visualisieren. In Studien haben Expert:innen klar das Potenzial der Visualisierung von vernetzten Kulturdaten für den Erkenntnisgewinn hervorgehoben. Einige haben jedoch auch

einschränkend festgestellt, dass diese Verknüpfungen und die Ereignisdichte im bestehenden IKG oft noch zu gering sind, um damit wissenschaftliche Fragestellungen zu untersuchen.

(2) Eine weitere Hürde stellen Schutzrechte und die freie Zugänglichkeit zu den zugrundeliegenden Daten dar: Während in manchen Bereichen und Nationen freier Zugang zum kulturellen Erbe eine Selbstverständlichkeit ist, sind viele kulturelle Informationen rechtlich (noch?) nicht frei zugänglich. So sind in einigen biographischen Datenbanken die strukturierten Daten zugänglich, nicht jedoch die Volltexte - selbst in Nationen wie den Niederlanden, wo der freie Zugang zu kulturellen Objekten zur nationalen Strategie erklärt wurde. Für die Visualisierung (nicht nur) von kulturellen Daten ist die Verknüpfung von close und distant reading / viewing ein zentrales Desideratum, das jedoch den freien Zugang zu den kulturellen Objekten und Texten für eine vertiefte Auseinandersetzung voraussetzt.

(3) Ein großes Potenzial eröffnet sich für die Erstellung eines kulturellen Wissensgraphen durch den Einsatz von AI und NLP Methoden. Diese können dabei helfen strukturierte Daten zu textlichen Beschreibungen zu erzeugen (Fokkens et al., 2017), aber auch z.B. bei der Konstruktion von Geschichten unterschützen (Bartalesi et al., 2023). Methoden des Semantic Web erlauben die stärkere Verbindung und Vernetzung von Daten. Werden diese Methoden mit AI kombiniert, können deutlich dichtere Datensätze erstellt werden. AI kann z.B. ähnliche Knoten für eine owl:sameAs Relation vorschlagen, oder Konzepte aus kontrollierten Vokabularien automatisch übersetzen und anschließend ebenfalls owl:sameAs Relationen vorschlagen (Hyvönen et al., 2023). In diesem Zusammenhang müssen jedoch Fragen der Reliabilität und Qualität der Daten geklärt werden. Für Wissenschaftlerinnen ist ein wichtiges Desideratum in diesem Zusammenhang zumindest die Transparenz, wie die Daten erzeugt wurden und die Visualisierung damit einhergehender Unsicherheiten (Windhager et al., 2019).

(4) Die Einbeziehung der Benutzer:innen in InTaVia hat klar gezeigt, dass es offener und flexibler Ansätze bedarf, um deren heterogener Bedürfnisse und Anforderungen gerecht zu werden: Viele Benutzer:innen würden gerne eigene Daten gemeinsam mit dem Wissensgraphen nutzen (oft aber ohne diese in den Wissensgraphen einzuspeisen - aus rechtlichen Gründen und Gründen der unsicheren Datenqualität). Einige Expert:innen würden gerne Daten, aber auch Visualisierungen flexibel exportieren, um diese mit andere Werkzeugen weiterzuarbeiten. Generell sind daher rechtliche Möglichkeiten für die offene Nutzung und Weiterentwicklung der Daten und des Programmiercodes (im Sinne von FAIR data) in Forschungsprojekten ein Desideratum, das wir auch in InTaVia verfolgt haben.

Zusammenfassend zeigt das Projekt InTaVia vor allem das Potenzial eines Wissensgraphen für die (visuelle), kontextualisierte Exploration von kulturellen Informationen, aber auch die Grenzen, aufgrund weniger Verknüpfungen zwischen den genutzten Datenbanken. Die Einbindung weiterer kultureller Datenbanken in den Wissensgraphen kann die Dichte der Verknüpfungen und damit das Potenzial für Exploration erhöhen, z.B. die Integration der Neuen Deutschen Biographie als Brücke zwischen Österreich und den Niederlanden, aber auch die Ergänzung weiterer Objekt-Personen-Relationen, wie beispielsweise Ausstellungen, Auktionen oder Korrespondenzen.

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

Bartalesi, Valentina, Coro, Gianpaolo, Lenzi, Emanuele, Pagano, Pasquale, und Pratelli, Nicolò. 2023. "From unstructured texts to semantic story maps." *International Journal of Digital Earth* 16: 234-250. 10.1080/17538947.2023.2168774

Doerr, Martin. 2003. “The CIDOC Conceptual Reference Module - An Ontological Approach to Semantic Interoperability of Metadata.” *AI Magazine*, *24*(3), 75. 10.1609/aimag.v24i3.1720

Fokkens, Antske, ter Braake, Serge, Ockeloen, Niels, Vossen, Piek, Legêne, Susan, Schreiber, Guus, und de Boer, Victor. 2017. „BiographyNet: Extracting Relations between People and Events.“ In *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik*, hg. Ágoston Z. Bernád, Christine Gruber, und Matthias Schlögl 193-224. Wien: new academic press. https://arxiv.org/pdf/1801.07073.pdf

Grebe, Anja und Großmann, Ulrich. 2013. *Albrecht Dürer. Niederländische Reise. Tagebuch und Kommentar*. Imhof Verlag.

Hyvönen, Eero, Rantala, Heikki, Leskinen, Petri, und Peura, Lilli. 2023. *Discovering Interesting Relations in Knowledge Graphs Based on Tangible and Intangible Cultural Heritage*. White Paper within the H2020 project InTaVia.

Kusnick, Jakob, Jänicke, Stefan, Doppler, Carina, Seirafi, Kasra, Liem, Johannes, Windhager, Florian und Mayr, Eva. 2021. *Report on narrative visualization techniques for OPDB data*. Technical report, InTaVia project, 2021. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e47d9524&appId=PPGMS (zugegriffen 10. Juli 2023)

Liem, Johannes, Kusnick, Jakob, Beck, Samuel, Windhager, Florian und Mayr, Eva. 2023. „A Workflow Approach to Visualization-Based Storytelling with Cultural Heritage Data”. In *2023 IEEE 8th Workshop on Visualization for the Digital Humanities (VIS4DH)*. IEEE. https://vis4dh.dbvis.de/papers/2023/vis4dh2023-preprint.pdf (zugegriffen 28. November 2023)

Mayr, Eva, Windhager, Florian, Liem, Johannes, Beck, Samuel, Koch, Steffen, Kusnick, Jakob und Jänicke, Stefan. 2022. „The multiple faces of cultural heritage: Towards an integrated visualization platform for tangible and intangible cultural assets”.

In *2022 IEEE 7th Workshop on Visualization for the Digital Humanities (VIS4DH)* 13–18. IEEE. 10.1109/VIS4DH57440.2022.00008

Tuominen, Jouni, Hyvonen, Eero, und Leskinen, Petri. 2017. „Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research". In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*. https://ceur-ws.org/Vol-2119/paper10.pdf (zugergriffen 10. Juli 2023)

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* 137–150. Walter de Gruyter GmbH & Co KG. 10.1515/9783110757101-008

Windhager, Florian, Federico, Paolo, Schreder, Günther, Glinka, Katrin, Dörk, Marian, Miksch, Silvia, und 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. 10.1109/TVCG.2018.2830759

Windhager, Florian, Salisu, Saminu, und Mayr, Eva. 2019. „Exhibiting uncertainty: Visualizing data quality indicators for cultural collections." *Informatics*, 6: 29. 10.3390/informatics6030029

Das »ureigenste theatralische Element« – Automatische Extraktion von Requisiten aus deutschsprachigen Dramentexten

Lubin, Jonah
jonah.lubin@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0009-0000-7900-1994

Detken, Anke
anke.detken@phil.uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Fischer, Frank
fr.fischer@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0003-2419-6629

Forschungsstand und Vorhaben

Die statuierte Untererforschung von Regieanweisungen im Drama (vgl. Aston/Savona 1991, S. 11 und S. 71, sowie Rasmussen 2003, S. 226) ist inzwischen einigermaßen adressiert worden. Kleinere und größere Studien zum »Nebenraum« dramatischer Texte (Detken 2009), meist entwickelt anhand weniger konkreter Dramentexte, wurden durch erste quantitative Ansätze ergänzt (Sperantov 1998), die zuletzt auch um die Methoden der Computational Literary Studies erweitert wurden (Maximova et al. 2018, Trilcke et al. 2020).

Der vorliegende Beitrag fokussiert auf einen wichtigen Teilaspekt, nämlich die Requisiten, die in Nebenbemerkungen Erwähnung finden (vgl. die grundlegende Studie von Sofer 2003). Das Requisit ist »das ureigenste dramatische Element«, das »den Übergang von der sprachlichen zur theatralischen Ebene« markiert. Es ist »zur Kompetenz des Dichters« zu rechnen und darf »bei der Untersuchung eines dramatischen Werkes nicht übergangen werden« (Schwarz 1974, S. 12).

Die bisherige Forschung konzentrierte sich meist auf einzelne herausgehobene Requisiten, etwa das Attributsrequisit, das zur Verdeutlichung einer dramatischen Figur dient, und das Emblemrequisit, das unabhängig von den *dramatis personae* eine weltanschaulich-existenzielle Funktion hat, etwa Königsinsignien wie die Krone oder der Totenschädel in Shakespeares »Hamlet« (vgl. Schwarz 1974, S. 18f. und S. 33). Dem Attributsrequisit, das zum Ausführen alltäglicher Tätigkeiten dient und das gewissermaßen zur Figur und ihrer Rolle gehört, stehen Requisiten gegenüber, denen eine eigene besondere Bedeutung für den Handlungsverlauf zukommt. Ist Letzteres der Fall, dann werden sie zu handlungsbestimmenden Gegenständen, etwa den »fatalen Requisiten« (vgl. Benjamin 1963, S. 141f.).

Ziel dieses Beitrags ist es, den Blick über diese speziellen Requisiten hinaus tendenziell auf die Gesamtheit der Requisiten zu richten. Dabei greifen wir auf folgende Definition zurück, ohne diese jedoch beim derzeitigen Stand vollständig operationalisieren zu können, besonders was den Interaktionsaspekt betrifft: »Im Unterschied zu den Dingen, die der dekorationsmäßigen Ausgestaltung des Bühnenraums dienen, sind Requisiten im engeren Sinne ›Gegenstände, mit denen der Schauspieler bei der Aufführung von Bühnenstücken agiert‹.« (Schwarz 1974, S. 18)

Anhand des German Drama Corpus (GerDraCor; Fischer et al. 2019), das als »living corpus« beständig wächst und derzeit über 600 deutschsprachige Dramen vom 16. bis zum 20. Jahrhundert im Volltext enthält, soll die Verteilung dramatischer Requisiten quantifiziert werden, sowohl chronologisch als auch genrebezogen. Durch die im Vergleich zu bisherigen Arbeiten größere Anzahl an Texten geraten auch nicht-kanonische Texte mit ins Bild und ermöglichen einen repräsentativeren Blick auf die Dramenproduktion des betrachteten Zeitraums.

Bezogen auf den Dramentext unterscheidet Roman Ingarden in seiner formalen Betrachtungsweise die Requisiten

von nur im Haupttext, also in der Figurenrede erwähnten Gegenständen. Allein die Nennung im Haupttext macht ein Ding also noch nicht zwingend zu einem Requisit. Dies ist erst der Fall, wenn der Gegenstand zusätzlich in den Regiebemerkungen (oder auch nur dort) erwähnt wird (vgl. Ingarden 1965, S. 405). Wir folgen Ingarden hierin und betrachten nur Requisiten, die in Regiebemerkungen genannt werden.

Im Folgenden wird zunächst der Workflow für die Extraktion der Requisiten vorgestellt, gefolgt von einigen exemplarischen Analysen.

Workflow

Die auf der *Drama Corpora*-Plattform (DraCor) versammelten Theaterstücke sind im TEI-Format kodiert. Der Standard sieht vor, dass Regieanweisungen (engl. *stage directions*) mit dem Element ›stage‹ ausgezeichnet werden. Sind sie entsprechend markiert, können sie zielgerichtet aus den Dokumenten extrahiert werden. Die DraCor-API erleichtert das noch insofern, als Regieanweisungen zu allen Stücken direkt heruntergeladen werden können, was wir für die zum Zeitpunkt des Datenbezugs insgesamt 609 deutschsprachigen Stücke getan haben, die zwischen 1730 und 1950 erschienen sind.

Aufgrund der verwendeten historischen Orthografie in einer Vielzahl der in GerDraCor enthaltenen digitalisierten Editionen wurde ein Normalisierungsschritt nötig, um die Lemmata zu vereinheitlichen (»Schwerd« bzw. »Schwerdt« wird zu »Schwert«). Dies geschah mithilfe des DTA::CAB Web Service (vgl. Jurish 2012). Die resultierende Liste modifizierter Wörter haben wir manuell moderiert, um die Ergebnisse zu optimieren.

Um im nächsten Schritt die Requisiten zu extrahieren, wurden die Regiebemerkungen zunächst mit ›spaCy‹ POS-getaggt, um die Suche auf Substantive bzw. *noun phrases* beschränken zu können, wobei auch Eigennamen ausgelassen wurden.

Mithilfe von GermaLemma (https://github.com/WZBSocialScienceCenter/germalemma) wurden die Substantive morphologisch auf ihre Grundformen reduziert. Komposita wurden mit CharSplit (https://github.com/dtuggener/CharSplit) in ihre Bestandteile zerlegt. GermaNet (Hamp/Feldweg 1997, Henrich/Hinrichs 2010) in der aktuellen Version 18.0 wurde benutzt, um die resultierenden Lemmata gegebenenfalls als potenzielle Requisiten zu klassifizieren – dies geschieht immer dann, wenn ein Lemma in GermaNet in mindestens einer Bedeutung als ›Artefakt‹, ›Pflanze‹ oder ›Nahrung‹ klassifiziert wird.

Als Ansatz für die Disambiguierung haben wir dann einen Simplified Lesk Algorithmus mit Glossen von Wiktionary und lexikalischen Feldern von GermaNet verwendet (wie beschrieben in Henrich/Hinrichs 2012). Konnte diese Methode keine definitive Word Sense Disambiguation liefern, haben wir den ranghöchsten Sinn von Wiktionary übernommen.

Obwohl wir dank GerDraCor zwar sehr viel Text, also auch Regieanweisungen haben, sind die Ergebnislisten doch überschaubar und taugen für einen explorativen *mixed methods*-Ansatz. Beispielhaft seien die 18 Requisiten aufgeführt, die aus Luise Gottscheds Lustspiel »Der Witzling« (1745) extrahiert wurden: »Tische«, »einen versiegelten Brief«, »den Brief«, »den Brief«, »den Tisch«, »dem Geräte«, »ein Schälchen Kaffee«, »ein Stück Zucker«, »die Degen«, »einem jeden Kaffee«, »den Brief«, »Tische«, »seinen Degen«, »den Brief«, »den Brief«, »den Stuhl«, »einen versiegelten Brief«, »Degen«. Die in den Regieanweisungen vorkommenden »Hüte« wurden entsprechend unserer Operationalisierung nicht extrahiert, da es sich um ›Kleidung‹ handelt. Insgesamt schärfen die so herausgelösten Requisiten den Blick auf ihre Rolle im Stück; drei von ihnen seien in ihrem Bezug zur Handlung kurz erläutert: Dem Degen kommt hier als Kavaliersdegen nur die Funktion eines Attributsrequisits zu, der (falsch adressierte) Brief hingegen ist typisch für die Komödie, und der Kaffee – das Getränk, das die Kaffeehauskultur der Aufklärung prägte – erkaltet symbolisch über der Diskussion der drei vorgeführten Möchtegerns und wird schließlich unverrichteter Dinge abgeräumt.

Analyse 1: Requisiten chronologisch

Analog zu den chronologischen Verlaufskurven in Trilcke et al. 2020, die gezeigt haben, dass Regieanweisungen entlang der Zeitachse systematisch umfangreicher und lebendiger werden (mehr Adjektive, mehr Verben), lässt sich nachweisen, dass auch die Erwähnung von Requisiten systematisch ansteigt. Abbildung 1 scheint zunächst auch genau dies zu zeigen. Allerdings korreliert in dieser Darstellung die Anzahl von Requisiten mit dem Umfang der Dramen: In einem formvollendeten Fünfakter ist durchschnittlich mehr Raum für Requisiten als in einem Einakter. Die drei Werke mit den meisten Nennungen von Requisiten – Arno Holz' »Ignorabimus« (1914) und »Sonnenfinsternis« (1908) sowie Karl Kraus' »Die letzten Tage der Menschheit« (1919) – gehören außerdem zu den umfangreichsten Dramen.

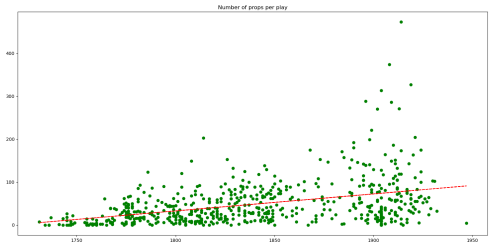


Abbildung 1: Erwähnungen von Requisiten im Korpus pro Drama, chronologisch sortiert.

Um die Daten zumindest ansatzweise zu normalisieren, haben wir die Anzahl der Requisiten durch die Anzahl von Segmenten pro Stück geteilt. Ein Segment ist eine Szene oder (in Dramen ohne Szeneneinteilung) ein Akt. Die Berechnung sei in Tabelle 1 anhand von drei Beispielen demonstriert.

Tabelle 1: Berechnung der durchschnittlichen Erwähnungen von Requisiten pro Segment (Akt, Szene oder eine andere Unterteilung) eines Dramas.

Drama	Anzahl Requisiten	Anzahl Segmente	Requisiten/Segment
Luise Gottsched: »Der Witzling« (1745)	18	9	2
G. E. Lessing: »Emilia Galotti« (1772)	24	43	0,56
Karl Kraus: »Die letzten Tage der Menschheit« (1919)	327	220	1,49

Abbildung 2 zeigt die Zahlen für das gesamte Korpus in chronologischer Darstellung. In dieser Darstellung finden wir bestätigt, dass sich die relative Häufigkeit von Requisiten in unseren Daten ab Ende des 19. Jahrhunderts ändert, parallel zur viel beschriebenen Krise des Dramas, die dann unter anderem Epifizierungstendenzen zeitigte (vgl. Weber 2017, S. 216). Allerdings sind die Outlier nach oben Stücke ohne Szeneneinteilung, die Segmente entsprechen hier umfangreichen Akten: Hermann Bahrs »Das Konzert« (1909) und »Das Phantom« (1913) sowie wiederum »Ignorabimus« von Arno Holz.

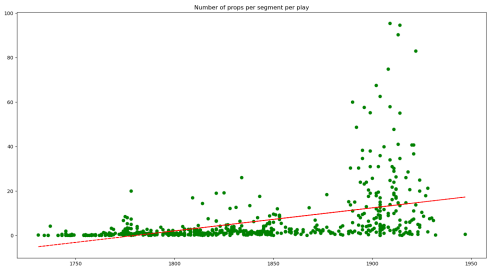


Abbildung 2: Durchschnittliche Erwähnungen von Requisiten im Korpus pro Segment (Akt, Szene oder eine andere Unterteilung) pro Jahr der Erstveröffentlichung.

Daher haben wir in Abbildung 3 noch einmal nur mit Akten als Segment gearbeitet (bei einem Minimum von einem Akt pro Stück, d. h., bei Dramen ohne explizite Akt-Angabe wurde die Aktzahl auf 1 gesetzt). Diese Normalisierung der Daten funktioniert am besten, das Feld fächert sich nach oben auf, ohne dass sich allerdings die Tendenz ändert.

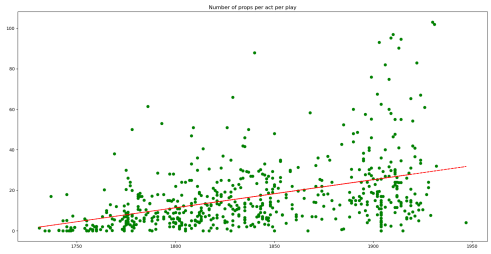


Abbildung 3: Durchschnittliche Erwähnungen von Requisiten im Korpus pro Akt pro Jahr der Erstveröffentlichung.

Analyse 2: Verteilung einzelner Typen von Requisiten nach Genre und formalen Aspekten

Die Verwendung von Requisiten gestaltet sich sowohl genreabhängig verschieden als auch im Hinblick auf formale Aspekte. Wiederum mithilfe von GermaNet haben wir uns beispielhaft einige Typen von Requisiten genauer angesehen: Waffen allgemein sowie die Unterklasse Feuerwaffen, dann speziell den Dolch, außerdem das Vorkommen von Kaffee als Requisit (Tab. 2 und 3).

Tabelle 2: Anteil aller Komödien bzw. Tragödien, in denen ein bestimmtes Requisit bzw. Typ von Requisit vorkommt.

Requisit bzw. Typ von Requisit	Komödie (N = 184)	Tragödie (N = 130)
Waffen	49,5%	86,2%
Feuerwaffen	12,5%	16,2%
Dolch	6,5%	38,5%
Kaffee	10,9%	6,9%

Die berechneten Prozentzahlen geben an, in wie vielen Komödien und Tragödien bzw. Prosa- und Versdramen ein bestimmtes Requisit bzw. ein bestimmter Typ von Requisit mindestens einmal vorkommt. Die Unterteilung in Vers- bzw. Prosadramen haben wir auf einfache Weise getroffen: Enthält ein Werk mehr Verszeilen (in TEI kodiert mit dem Element ›l‹) als Prosaabsätze (TEI-Element ›p‹), dann gilt es als Versdrama, ist das Verhältnis umgekehrt, dann klassifizieren wir es als Prosadrama.

Tabelle 3: Anteil aller Prosa- bzw. Versdramen, in denen ein bestimmtes Requisit bzw. Typ von Requisit vorkommt.

Requisit bzw. Typ von Requisit	Prosadrama (N = 378)	Versdrama (N = 231)
Waffen	60,1%	75,3%
Feuerwaffen	18,8%	7,8%
Dolch	10,6%	26,0%
Kaffee	13,8%	1,3%

Kaffee lässt sich also als Element der Prosakomödie bezeichnen, der Dolch ist ein Requisit der Verstragödie. Diese zwar nicht überraschenden, aber nun bezifferbaren Tenden-

zen mögen als Fingerzeig dafür dienen, wie sich die Quantifizierung von Requisiten sinnvoll einsetzen lässt.

Neben der Fokussierung auf einzelne Requisiten in größeren Korpora gerät auch die Breite des gesamten Arsenals in den Blick, was am Beispiel von erwähnten Waffen in Regieanweisungen erfolgen soll. So lassen sich mindestens 38 individuelle Waffentypen (bzw. Munition) ausmachen, inklusive spezifizierenden Komposita: Armbrust, Bajonett, Bogen, Bombe, Büchse, Degen, Dienstgewehr, Dolch, Doppelflinte, Dreizack, Fangmesser, Flinte, Florett, Geißel, Gewehr, Hellebarde, Hetzpeitsche, Kanone, Keule, Klinge, Knüppel, Knüttel, Lanze, Messer, Muskete, Peitsche, Pfeil, Pistole, Rasiermesser, Revolver, Schild, Schläger, Schnitzmesser, Schwert, Seitengewehr, Speer, Spieß, und Säbel.

Waffen im Drama sind immer auch chronologisch kodiert und verorten ein Stück in der Zeit der Handlung. Unter Rückgriff auf die Terminologie des Militärhistorikers Trevor N. Dupuy stammen die meisten Waffentypen im Korpus aus dem »Age of Muscle«; dem »Age of Gunpowder« sind nur ein knappes Dutzend zuzuordnen (Dupuy 1980, S. 288f.).

Kritik und Ausblick

Bei dieser quantitativen Sicht auf Requisiten ist freilich noch nichts über die Qualität der Erwähnungen gesagt. Im Bezug auf das Beispiel ›Dolch‹ und die Arbeit »Man and Object in the Theater« des Prager Strukturalisten Jiří Veltruskýs schreibt Sofer: »a stage dagger might move from being a passive emblem of the wearer's status to participating in the action as an instrument of murder, and thence to a final independent association with the concept ›murder.‹« (Sofer 2003, S. 9) Diese qualitativen Einordnungen können nun von Aspekten der Distribution eines Requisits innerhalb größerer Korpora begleitet werden.

Der hier präsentierte quantitative Ansatz soll einen Eindruck von der Häufigkeit und Distribution von Requisiten innerhalb deutschsprachiger Dramentexte vermitteln und es ermöglichen, auch die Rolle bisher wenig beachteter Typen von Requisiten zu erforschen. Durch Anpassung der verwendeten computerlinguistischen Tools ließe er sich auch auf Dramenkorpora in anderen Sprachen übertragen.

Insgesamt funktioniert die Extraktion von Requisiten recht zuverlässig. Die beobachteten *false negatives* gehen vor allem auf historische Grammatik zurück, etwa das Dativ-›e‹ (zum Beispiel in der Regieanweisung »Sich nach dem Porträte umsehend.« in Lessings »Emilia Galotti«), sowie auf Probleme bei der Disambiguierung mittels Simplified Lesk Algorithmus, die dazu führen können, dass ein »Tisch« einmal erfolgreich extrahiert wird, ein andermal nicht.

Es wäre wünschenswert, wenn die Befunde dieser kleinen Studie mittelfristig dazu führen würden, dass ein größeres deutschsprachiges Dramenkorpus hinsichtlich vorhandener Requisiten mit entsprechendem Markup versehen wäre, das als Evaluierungsbasis wie auch als Trainingsdatenset dienen könnte.

Bibliographie

Aston, Elaine und George Savona. 1991. Theatre as Sign System. A Semiotics of Text and Performance. London, New York: Routledge.

Benjamin, Walter. 1963. Ursprung des deutschen Trauerspiels. Frankfurt/M.: Suhrkamp.

Detken, Anke. 2009. Im Nebenraum des Textes. Regiebemerkungen in Dramen des 18. Jahrhunderts. Tübingen: Niemeyer.

Dupuy, Trevor N. 1980. The Evolution of Weapons and Warfare. New York: Bobbs-Merrill.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling und Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In: Proceedings of DH2019: »Complexities«, Utrecht University, https://doi.org/10.5281/zenodo.4284002.

Hamp, Birgit und Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid.

Henrich, Verena und Erhard Hinrichs. 2010. GernEdiT – The GermaNet Editing Tool. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta. S. 2228–2235.

Henrich, Verena und Erhard Hinrichs. 2012. A Comparative Evaluation of Word Sense Disambiguation Algorithms for German. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul. S. 576–583. http://www.lrec-conf.org/proceedings/lrec2012/pdf/164_Paper.pdf.

Ingarden, Roman. 1965. Das literarische Kunstwerk. Mit einem Anhang von den Funktionen der Sprache im Theaterschauspiel. Dritte, durchgesehene Auflage. Tübingen: Niemeyer.

Jurish, Bryan. 2012. Finite-State Canonicalization Techniques for Historical German. PhD thesis, Universität Potsdam, https://nbn-resolving.org/urn:nbn:de:kobv:517-opus-55789.

Maximova, Daria, Frank Fischer und Daniil Skorinkin. 2018. A Quantitative Study of Stage Directions in Russian Drama. In: EADH2018: »Data in Digital Humanities«. 7.–9. Dezember 2018. National University of Ireland, Galway, https://eadh2018.exordo.com/files/papers/79/final_draft/Stage_Directions_for_EADH_Conference.pdf.

Rasmussen, Eric. 2003. Afterword. In: Hardin L. Aasand (Hg.): Stage Directions in Hamlet. New Essays and New Directions. Madison und Teaneck: Fairleigh Dickinson

University Press; London: Associated University Presses, S. 226–227.

Schwarz, Hans-Günther. 1974. Das stumme Zeichen. Der symbolische Gebrauch von Requisiten. Bonn: Bouvier.

Sperantov, V. V. 1998. Poetika remarki v russkoy tragedii XVIII – nachala XIX vv. (K tipologii literaturnykh napravleniy) [Poetik der Regieanweisungen in der russischen Tragödie des 18. und frühen 19. Jahrhunderts (Zur Typologie literarischer Strömungen)]. In: Philologica. Vol. 5, Nr. 11/13, S. 9–48, https://rvb.ru/philologica/05/05sperantov.htm.

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: DH2020: »carrefours/intersections«. 22.–24. Juli 2020. Book of Abstracts. University of Ottawa, https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html.

Weber, Alexander. 2017. Episierung im Drama: Ein Beitrag zur transgenerischen Narratologie. Berlin, Boston: De Gruyter 2017, https://doi.org/10.1515/9783110488159.

DHd Chronicles Anreicherung und Analyse der Beiträge zu den Jahrestagungen der Digital Humanities im deutschsprachigen Raum 2014-2023

Cremer, Fabian
cremer@ieg-mainz.de
Leibniz-Institut für Europäische Geschichte (IEG), Mainz, Deutschland
ORCID: 0000-0001-8251-9727

Blessing, André
andre.blessing@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland
ORCID: 0000-0001-7573-578X

Helling, Patrick
patrick.helling@uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland
ORCID: 0000-0003-4043-165X

Henny-Krahmer, Ulrike
ulrike.henny-krahmer@uni-rostock.de
Institut für Germanistik, Universität Rostock, Deutschland
ORCID: 0000-0003-2852-065X

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland
ORCID: 0000-0002-9548-8461

Reiter, Nils
nils.reiter@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln, Deutschland
ORCID: 0000-0003-3193-6170

Einleitung

Die seit 2014 jährlich stattfindenden Tagungen[1] des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd) stellen mit ca. 130 angenommenen Beiträgen pro Tagung ein regelmäßiges Schaufenster der digital arbeitenden Geisteswissenschaften (Schöch 2020:V) dar. Sie folgen dem Anspruch, die Entwicklungen, Ergebnisse und relevanten Fragestellungen der Digital Humanities (DH) im deutschsprachigen Raum zu präsentieren. Mit der zehnten Jahrestagung blickt die folgende Analyse auf die vergangenen Beiträge und untersucht, inwieweit sich Entwicklungen, Konstanten, Muster und Phänomene erkennen und analysieren lassen. Dieser Beitrag soll 1) die angereicherten Datensätze, basierend auf den angenommenen Beiträgen der DHd-Jahrestagungen, vorstellen und kritisch beleuchten. Darüber hinaus sollen die 2) Entwicklung der DHd-Jahrestagungen auf einer strukturellen und inhaltlichen Ebene untersucht und 3) neue Impulse für die konzeptionelle Weiterentwicklung der Tagungen und der Auswertung von Tagungsbeiträgen in den DH gegeben werden.

Der Datensatz zu den DHd-Jahrestagungen

Quelldaten und Provenienz

Seit 2015 gehört das Book of Abstracts mit allen angenommenen Tagungsbeiträgen zu einer zentralen Veröffentlichung der deutschsprachigen DH (Stiegler 2015; Burr, 2017; Stolz 2017; Vogeler 2018; Sahle 2019; Schöch 2020; Geierhos 2022; Busch und Trilcke 2023).[2] Im Zuge des Einreichungsprozesses (Helling et al. 2022a) werden seit 2016 durch den DHConvalidator[3] TEI-XML-Dateien der Beiträge erzeugt. Darüber hinaus wurden durch den DHd Data

Steward seit 2020 alle Konferenzbeiträge auch als einzelne PDF-Publikationen durch einen automatisierten Workflow (Borges et al. 2022) persistent und zitierbar veröffentlicht (Helling et al. 2022b). Insgesamt konnten so 1.207 Konferenzbeiträge sowie 109 Posterpräsentationen der DHd-Jahrestagungen 2014—2023 publiziert werden. Zusätzlich wurden für die Jahrestagungen 2016—2023 die TEI-XML-Dateien sowie ihre Metadaten in GitHub-Repositorien zur Nachnutzung verfügbar gemacht.[4] Für die Jahrestagungen 2014 und 2015 stehen die veröffentlichten PDF-Dateien und die Metadaten zur Verfügung.

Erstellung und Anreicherung

Auf der Basis von ConfTool-Daten wurden zu einigen DHd-Jahrestagungen bereits Analysen durchgeführt, um ein inhaltliches und strukturelles Bild der jeweiligen Tagung zu zeichnen (siehe bspw. Calvo Tello 2016; Henny-Krahmer und Sahle 2018; Kiefer 2019; Hoenen 2019). Einige experimentelle und exemplarische Ansätze haben zusätzlich die Potentiale und Herausforderungen einer Anreicherung aller Beitragsdaten zur quantitativen Analyse gezeigt (Andorfer et al. 2020; Andorfer et al. 2021; Busch et al. 2022). Die Qualität der Metadaten in den GitHub-Repositorien ist trotz der bereits erfolgten Vereinheitlichungen noch nicht für alle Analysen ausreichend. In einem iterativen Prozess, der sowohl automatische Verfahren (String-Similarities, Geonames[5]-Abfragen) als auch manuelle Bearbeitungen mittels OpenRefine[6] enthält, konnten im Rahmen dieses Analysevorhabens die verzeichneten Namen der Autor:innen verbessert (5,9 % waren fehlerhaft) und alle Affiliationen auf Städte und Länder, wo vorhanden auch auf eine (übergeordnete) ROR-ID,[7] abgebildet werden. Außerdem wurden alle Tagungsbeiträge 2014—2023 als plain text extrahiert und via GitHub veröffentlicht.[8] Auf diese Weise konnte eine Datenbasis generiert werden, um die DH-Landschaft anhand der DHd-Jahrestagungen über die Zeit hinweg auf struktureller und inhaltlicher Ebene zu analysieren.[9]

Analyse der Konferenzbeiträge

Strukturelle Entwicklungen

Anzahl der Beiträge und Formate

Die Beiträge werden in den Formaten Vortrag, Workshop, Panel (ehemals Sektionen), Poster und (seit 2020) Doctoral Consortium präsentiert. Die Anzahl der Beiträge pro Tagung beträgt im Durchschnitt 134 und der Anteil der Beitragsformate bleibt über die Zeit relativ konstant (siehe Abb. 1).

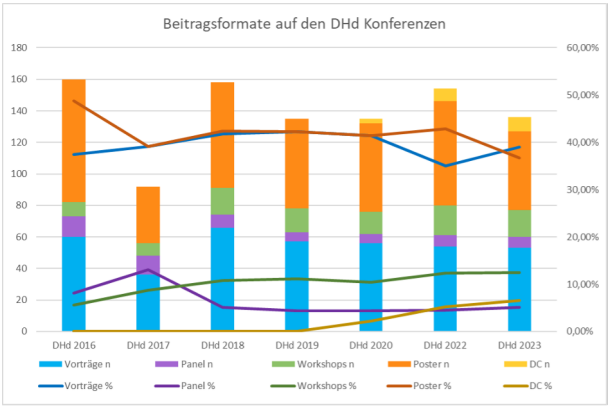


Abb. 1: Angenommene Beitragsformate.

Kollaborative Autor:innenschaft

Die DH definieren sich als interdisziplinäre und kollaborativ forschende Community, was sich u. a. in der Multiautor:innenschaft der Veröffentlichungen widerspiegeln sollte, aber weder in peer-reviewed Zeitschriftenartikeln (Nyhan und Duke-Williams 2014) noch in den Konferenzbeiträgen (Weingart 2015) der Fall ist. Auf den DHd-Jahrestagungen sind Einzelbeiträge jedoch seit Beginn seltener als im internationalen Vergleich und stetig rückläufig. Duos und Kleingruppen (3—5 Personen) präsentieren konstant die Mehrzahl der Beiträge. Große Gruppen (>5 Personen) sind seit 2016 mindestens in zweistelliger Zahl vertreten, bilden die zuletzt am stärksten wachsende Kooperationsform und übersteigen damit Einzelbeiträge und Duos (siehe Abb. 2).

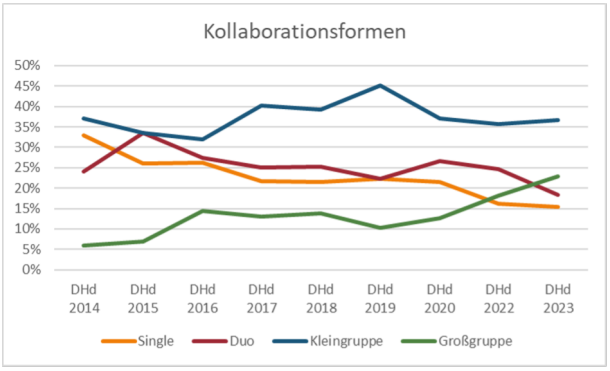


Abb. 2: Kollaborationsformen von Autor:innen.

Das Beitragsformat hat dabei zwar Auswirkungen auf die Anzahl der Autor:innenschaft, allerdings sind die unterschiedlichen Verteilungen hier vor allem der Präsentationsform geschuldet – so setzen etwa Panel und Workshop oft mehrere Beteiligte voraus. Es lassen sich keine Entwicklungen oder Muster über die Jahre verzeichnen (siehe Abb. 3).

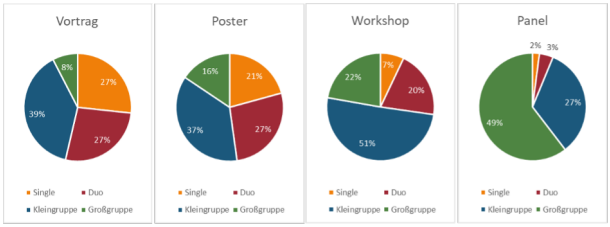


Abb. 3: Autor:innenschaft im Verhältnis zu Beitragsformaten.

An der Zusammenarbeit in Teams lassen sich neben der interdisziplinären und kooperativen Ausrichtung auch strukturelle Merkmale bei der institutionellen Verortung der DH finden. So sind institutionsübergreifende Teams bei Kleingruppen die Regel (95 %), während Duos häufiger aus einer Einrichtung kommen (38 % institutionsübergreifend). Großgruppen sind bereits größenbedingt eher auf mehrere Institutionen verteilt (78 %), jedoch lässt sich ein stabiler Anteil an institutionellen Großgruppen (z. B. Abteilungen und Zentren) verzeichnen.

Länder und länderübergreifende Kooperationen

Die meisten Einreichungen zu den Tagungen stammen aus Institutionen in Deutschland, gefolgt von Österreich. Aus der Schweiz stammen mit Ausnahme von 2017, als die Tagung in Bern stattfand, weniger Beiträge. Die Zahl der Kooperationen bei Beiträgen zwischen Autor:innen aus DACH-Ländern und weiteren Ländern steigt im Laufe der Tagungen geringfügig, bleibt jedoch immer unter einem Anteil von 10 %. Beiträge, die ausschließlich aus Ländern außerhalb des DACH-Raums stammen, nehmen einen sehr geringen Anteil ein (siehe Abb. 4). Ein ähnliches Bild ergibt die Auswertung der Autor:innen-Gesamtanzahl (siehe Abb. 5).

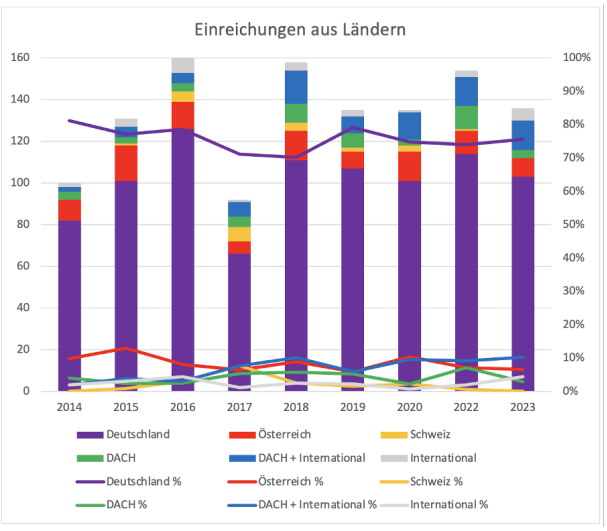


Abb. 4: Beiträge aus Ländern in absoluten Zahlen.

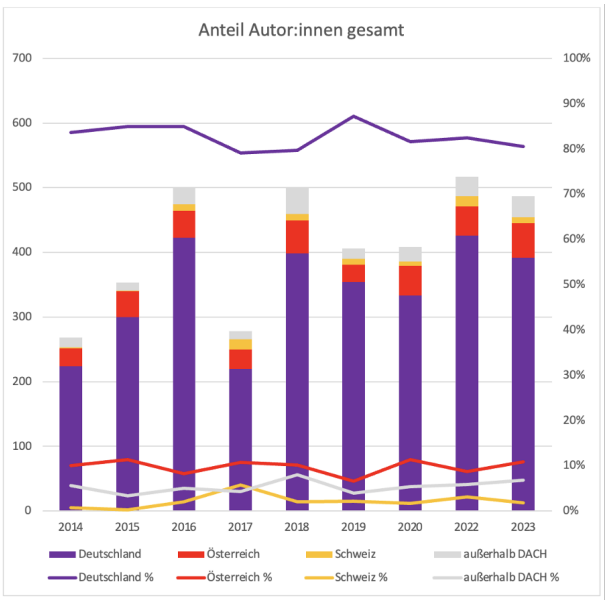


Abb. 5: Gesamtzahl der Autor:innen aufgeteilt auf die DACH-Länder sowie außerhalb.

Offenheit und Community

Offenheit und Kollegialität bilden zwei zentrale Werte der DH (Spiro 2012). Zwar können Beitragsstatistiken nur wenig über die gelebte Praxis dieser Werte auf den Tagungen aussagen, jedoch sollte das Verhältnis von Personen, die erstmals auf einer DHd einen Beitrag vorstellen und den Personen, die bereits schon einmal präsentiert haben, die Zusammensetzung der Tagung charakterisieren können. Nach der Aufbauphase präsentieren auf den DHd-Jahrestagungen konstant sowohl neue als auch wiederkehrende Autor:innen (siehe Abb. 6).

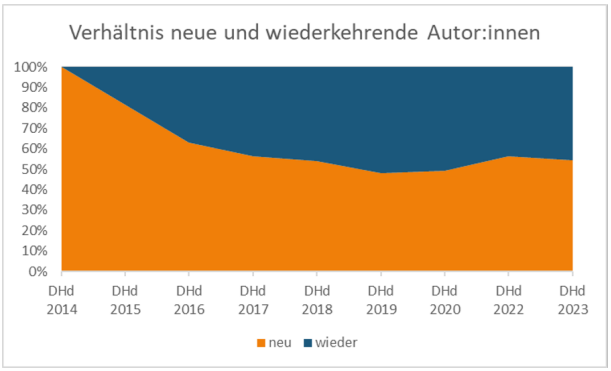


Abb. 6: Verhältnis zwischen neuen und wiederkehrenden Autor:innen.

Geschlechterverhältnisse und Gleichstellung

Die Repräsentation der Diversität der DH-Community auf den Tagungen hat sich in der Vergangenheit bei Untersuchungen als problematisch herausgestellt, insbesondere waren bei den Beiträgen früher weibliche Autor:innen un-

terrepräsentiert (Weingart und Eichmann-Kalwara 2017). Um die Beiträge zu den DHd-Tagungen unter diesem Aspekt untersuchen zu können, wurden die Vornamen der Autor:innen auf Basis der genderize.io API[10] hinsichtlich eines weiblichen oder männlichen Geschlechts vorhergesagt.[11]

Die Analyse der DHd-Daten zeigt für die ersten Jahre ein ähnliches Bild, das auch bei den internationalen ADHO-Konferenzen 2004—2013 zu sehen ist (siehe Abb. 7): Der Anteil weiblich erkannter Autor:innen verläuft bis 2019 um 30 %. Erst in den Jahren 2020—2023 steigt dieser Anteil bis 44 %. Die Lücke in der Beteiligung an den Beiträgen insgesamt (2014: 86 % männl. 47 % weibl.) verringert sich kontinuierlich und schließt sich 2023, wobei der kontinuierlich steigende Anteil an Einzelbeiträgen weiblich erkannter Personen seit 2020 hervortritt. Damit lässt sich zwar insgesamt eine Entwicklung hin zu einer verbesserten Repräsentation von weiblich gelesenen Personen ablesen, allerdings kann diese Entwicklung noch nicht als verlässlich hin zu einer Gleichstellung eingestuft werden.

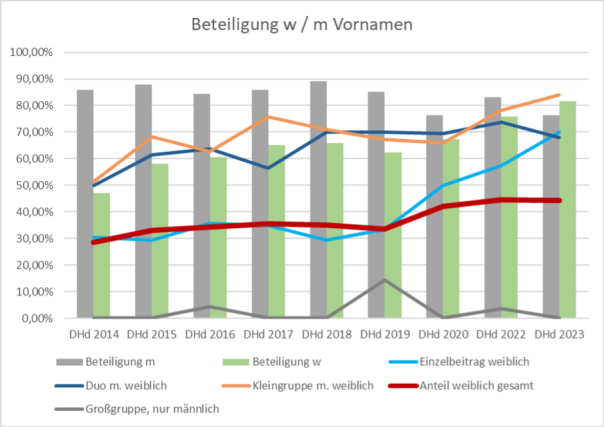


Abb. 7: Verhältnisse weiblich und männlich erkannter Autor:innen mit Beteiligungen an Beiträgen und Beitragsformaten.

Inhaltliche Entwicklungen

Um zu untersuchen, wie sich die DHd-Tagungsbeiträge über die Zeit inhaltlich entwickelt haben, wurden Topic-Modelle für alle Beiträge erstellt. Insgesamt wurden 1.196 Abstracts aus den Jahren 2014—2023 untersucht. Für die Auswertung wurde ein Modell mit 40 Topics ausgewählt.[12] Es gibt einige allgemeine Topics zu DH, die im Korpus auch stark vertreten sind (z. B. T15), methodenspezifische Topics (z. B. T6), Topics zu untersuchten Gegenständen (wie T26) und auch auf die Beiträge bezogene Topics (etwa T11, siehe Abb. 8).

Abb. 8: Typische Arten von Topics im Abstract-Topic Model.

Im Folgenden wird die Entwicklung von Gruppen von Topics über die Jahre in den Blick genommen. Werden bestimmte Methoden über die Zeit unterschiedlich stark eingesetzt und folgt diese Entwicklung allgemeinen technologischen Trends? Welche Gegenstände standen wann im Vordergrund? Abb. 9 zeigt die Entwicklung mehrerer ausgewählter Methoden-Topics über die Jahre.[13]

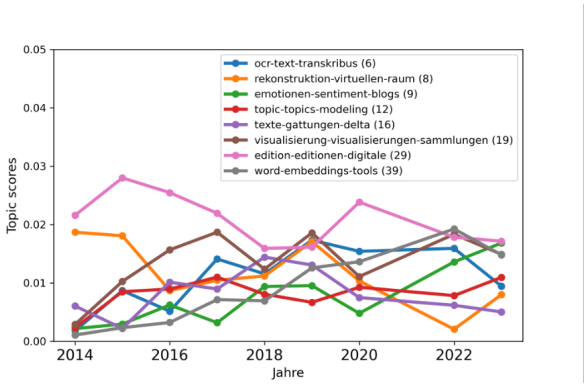


Abb. 9: Methoden-Topics über die Jahre.

Topics, die 2014 vergleichsweise hohe Wahrscheinlichkeiten hatten, sind z. B. “digitale Editionen/XML/TEI” (T29) und “Virtual Reality & Rekonstruktion” (T8). Beide zeigen leicht abnehmende Verläufe; das Editions-Topic bleibt aber auch 2023 eines der stärksten Methoden-Topics in der gezeigten Auswahl. Das “Stilometrie”-Topic (T16) erreichte den Höhepunkt 2018 und ist seitdem rückläufig. “Topic Modeling” selbst (T12) hält sich seit 2015 im unteren Mittelfeld relativ stabil. “Visualisierung” (T19) ist seit 2017 auf einem relativ hohen Niveau. Aufschwünge sind außerdem bei den Topics “Word Embeddings & Machine Learning” (T39) und “Sentiment Analysis” (T9) zu erkennen. Das Topic “OCR & Texterkennung” (T6) nahm bis 2019 zu und ging 2023 zurück. Der Verlauf der ausgewählten Methoden-Topics über die Jahre zeigt, dass es für

einzelne Methoden verschiedene Trends und im Vergleich zu 2014 insgesamt eine ausgewogenere Methodenvielfalt als vorher gibt. Abb. 10 zeigt eine Auswahl von Topics zu Forschungsgegenständen.

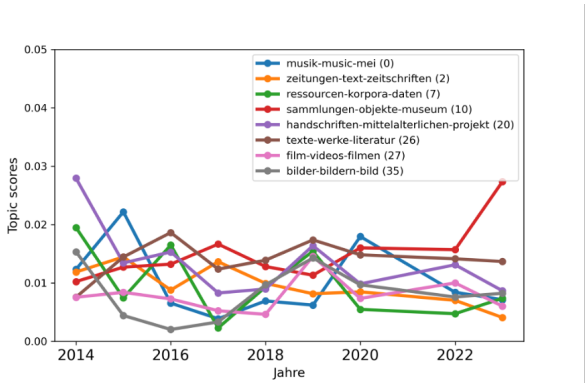


Abb. 10: Forschungsgegenstände über die Jahre.

Eine Zunahme über die Jahre ist beim Topic “Sammlungen, Objekte, Museum” (T10) zu erkennen, vor allem im Jahr 2023. Tendenziell abnehmend sind die Topics zu “mittelalterlichen Handschriften” (T20), “Sprachressourcen und Korpora” (T7) und “Zeitungen und Zeitschriften” (T2). “Literatur” (T26) als Gegenstand ist nach 2014 wichtiger geworden und entwickelt sich stabil. Für “Video und Film” (T27) und “Bilder” (T35) gab es 2019 Höhepunkte. Dies hängt vermutlich mit dem Thema der DHd-Tagung 2019 zusammen (“multimedial & multimodal”), was klar zeigt, dass die Wahl der Konferenzthemen auch die Forschung bzw. ihre Präsentation auf der Tagung beeinflusst. Das Thema “Musik” (T0) war besonders 2015 und 2020 präsent. 2020 fand die DHd-Tagung in Paderborn statt, wo u. a. das Zentrum “Musik - Edition - Medien” beheimatet ist. Demnach kann auch der Ort der jeweiligen Konferenz Einfluss auf die Themen haben.

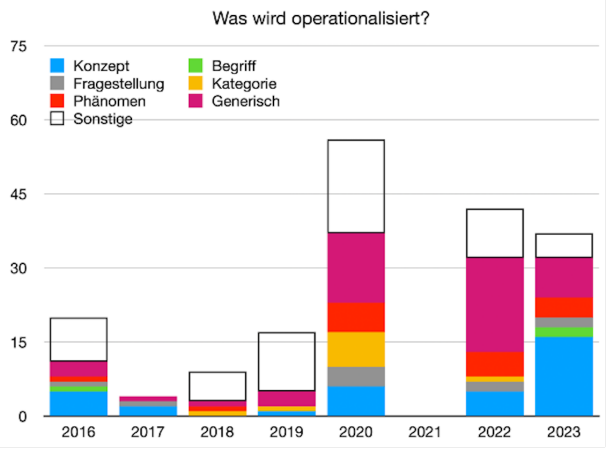


Abb. 11: Semantische Argumente verschiedener Wortformen von “Operationalisierung”.

In einem letzten Schritt haben wir mit der Operationalisierung einen häufig genannten und grundlegenden Tätigkeitsbereich in DH-Arbeiten (vgl. Pichler und Reiter 2022) in den Blick genommen (siehe Abb. 11). Eine ähnliche Analyse (mit deutlich mehr Fundstellen) wäre etwa zu den Begriffen der Modellierung oder Implementierung denkbar. In allen untersuchten Jahrgängen taucht der Begriff in verschiedenen sprachlichen Formen auf, häufig als Verb oder Nomen. Im Jahr 2020 steigt die Zahl der Vorkommen sprunghaft an und hält sich seitdem auf einem ähnlichen Niveau.[14] Unser Interesse galt in diesem Zusammenhang auch der Frage, was eigentlich operationalisiert wird. Dabei konnten sechs große Kategorien identifiziert werden: Neben Konzepten sind dies Fragestellungen, Phänomene, Begriffe und Kategorien.[15] Eine gewisse Menge an Vorkommen taucht ohne semantisches Argument auf, insbesondere wenn sich Beiträge mit theoretischen oder praktischen Folgen und Voraussetzungen von Operationalisierung befassen. Auffallend ist, dass diese Befassung mit ‘Operationalisierung als solcher’ erst ab 2020 in substanziellem Ausmaß stattfindet.

Als ein zentraler Output des Beitrags wurde schließlich mit dem webbasierten Visualisierungsframework Keshif (Yalçın et al. 2016) eine interaktive Visualisierung[16] der aufbereiteten Metadaten erstellt (siehe Abb. 12), die zur Exploration zur Verfügung steht.[17]

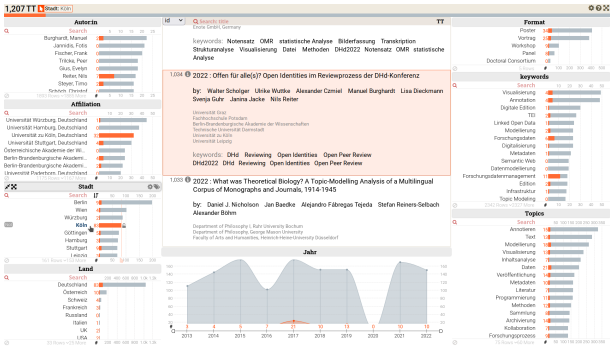


Abbildung 12: Visualisierung der iterativ bearbeiteten Metadaten mit Keshif.

Fazit

Die Anreicherung und Verwendung der Datensätze demonstrieren, dass die Beiträge der DHd-Jahrestagungen inzwischen quantitativ und qualitativ ein geeignetes Korpus zur Untersuchung struktureller und inhaltlicher Entwicklungen in den DH bilden. Dabei zeigt der Ausschnitt auf die zehn Jahre auch, dass “Trends” zwar erst über längere Zeiträume sichtbar werden können, aber einzelne Phänomene und Ansätze von Entwicklungen, die es zu beobachten oder näher zu betrachten gilt, auch kurzfristig identifizierbar werden. Der Ansatz, möglichst viele inhaltliche Attribute der Beiträge mit strukturellen und organisationsbezogenen Aspekten zu kombinieren, lässt die verschiedenen

Einflussfaktoren für inhaltliche wie auch strukturelle Entwicklungen hervortreten. Damit können die Beitragsdaten Grundlage vielfältiger Untersuchungen werden, etwa für ein kontinuierliches Monitoring zu ausgewogener Repräsentation auf den Jahrestagungen im DHd-Verband oder für die wissenschaftsgeschichtliche und selbstreflexive Auseinandersetzung mit den Beitragsthemen in der DH-Community und nicht zuletzt auch für die weiterhin notwendigen Anstrengungen zur Normalisierung und Verbesserung der Datenqualität des stets wachsenden und reicher werdenden Datenbestandes.

Die in diesem Beitrag vorgestellten Ergebnisse sollen in einem Vortrag auf der DHd-Jahrestagung 2024 detaillierter besprochen und durch weitere Ergebnisse vervollständigt werden, um ein umfassendes Bild der DHd-Jahrestagungen zu zeichnen und gleichzeitig für eine datenzentriertere Perspektive auf DHd-Tagungsbeiträge zu argumentieren.

Fußnoten

1. Mit einer Pandemie-bedingten Ausnahme 2021.
2. Zur ersten DHd-Jahrestagung 2014 wurde noch kein Book of Abstracts publiziert.
3. https://github.com/ADHO/dhconvalidator.
4. https://github.com/DHd-Verband .
5. https://www.geonames.org .
6. https://openrefine.org/ .
7. https://ror.org/
8. https://github.com/hennyu/dhd-chronicles .
9. Der diesem Beitrag zugrunde liegende Datensatz ist dennoch nicht vollständig, da nicht für alle DHd-Jahrestagungen alle Daten vorliegen. So gibt es bspw. keine Daten über die Einreichungsformate oder verwendete Schlagworte zu den Beiträgen zu den Tagungen 2014 und 2015.
10. https://genderize.io/ .
11. Die Methode ist in vielerlei Hinsicht limitiert: binäres Geschlechtersystem; Doppelbedeutung; keine zeitlichen Veränderungen (Keyes 2018; Gao et al. 2022). Wir vertreten den Standpunkt, dass in diesem Fall die begrenzten Erkenntnisse aus der geschlechtsspezifischen Analyse die Probleme durch die berechnete Ableitung des Geschlechts überwiegen (Posner 2017).
12. Für Implementierungsdetails und Analysedaten siehe https://github.com/hennyu/dhd-chronicles/tree/main/tm .
13. Die Topic-Scores sind Durchschnittswerte der Wahrscheinlichkeiten eines Topics in allen Beiträgen eines Jahres.
14. Hierbei ist zu beachten, dass Begriffshäufigkeiten gezählt wurden, und nicht Beiträge, in denen der Begriff vorkommt. Beiträge, die sich mit dem Begriff selbst befassen, verzerren die Zählung gegenüber denen, bei denen die Operationalisierung nur ein Teilschritt ist. Insbesondere im Jahr 2020 sind es zwei Workshops, die sich eingehend der Thematik widmen und zum großen Anstieg beitragen.
15. Bei dieser Zählung wurden Überschriften, Biographien von Autor:innen und Workshop-spezifischer Meta-Text ignoriert. Verschiedene Formen wurden sinnerhaltend vereinheitlicht, insbesondere auf Singularformen. Unter den Sonstigen häufiger sind: Daten, Wortschatz, Ansatz, Merkmal, Disziplin, Präsenz und Handlung.
16. Das Potential einer interaktiven Darstellung und Visualisierung deutete sich schon 2019 in der experimentellen Webanwendung dhd-boas-app (Andorfer 2019) an.
17. https://clarin03.ims.uni-stuttgart.de/dhd-chronicles/ .

Bibliographie

Andorfer, Peter. 2019. *dhd-boas-app*. https://dhd-boas-app.acdh-dev.oeaw.ac.at/.

Andorfer, Peter, Anna Busch, Fabian Cremer, Nickoal Eichmann-Kalwara, Patrick Helling, Andreas Henrich, Matthew Lincoln, u. a. 2021. „Die DHd-Abstracts im Zukunftslabor“. In *vDHd21 - Experimente*. Zenodo. https://doi.org/10.5281/ZENODO.4723039.

Andorfer, Peter, Fabian Cremer, und Timo Steyer. 2020. „Abstract Enhancement. Potentiale der DHd-Konferenzabstracts als Daten/Publikation“. In *DHd 2020: Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. Zenodo. https://doi.org/10.5281/ZENODO.4621705.

Borges, Rebekka, Anke Debbeler, und Patrick Helling. 2022. „Der DHd Data Steward - Maßnahmen zur Entwicklung einer nachhaltigen Datenstrategie für die Digital Humanities im deutschsprachigen Raum“. In *DHd 2022: Kulturen des digitalen Gedächtnisses*. Zenodo. https://doi.org/10.5281/ZENODO.6322482.

Burr, Elisabeth, Hrsg. 2017. *DHd 2016: Modellierung - Vernetzung - Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma* (Version 2. überarbeitete und erweiterte Ausgabe). Zenodo. https://doi.org/10.5281/ZENODO.3679331.

Busch, Anna, Fabian Cremer, Harald Lordick, Dennis Mischke, und Timo Steyer. 2022. „Strukturen und Impulse zur Weiterentwicklung der DHd-Abstracts“. In *DHd 2022: Kulturen des digitalen Gedächtnisses*. Zenodo. https://doi.org/10.5281/ZENODO.6328088.

Busch, Anna, und Peer Trilcke, Hrsg. 2023. *DHd 2023: Open Humanities, Open Culture*. Zenodo. https://doi.org/10.5281/ZENODO.7688632.

Gao, Jin, Julianne Nyhan, Oliver Duke-Williams, und Simon Mahony. 2022. „Gender Influences in Digital Humanities Co-Authorship Networks“. *Journal of Documentation* 78 (7): 327–50. https://doi.org/10.1108/JD-11-2021-0221.

Geierhos, Michaela, Hrsg. 2022. *DHd 2022: Kulturen des digitalen Gedächtnisses*. Zenodo. https://doi.org/10.5281/ZENODO.6304590.

Helling, Patrick, Rebekka Borges, Ingo Börner, Anna Busch, Fabian Cremer, Anke Debbeler, Henning Gebhard, Harald Lordick, und Timo Steyer. 2022a. „Der DHd-Verband und seine Abstracts – Betrachtungen des Einreichungsprozesses zu den DHd-Jahrestagungen“.

DHd-Blog. 21. November 2022. https://dhd-blog.org/?p=18599.

Helling, Patrick, Anke Debbeler, und Rebekka Borges. 2022b. „Konferenzbeiträge strategisch publizieren“. o-bib. Das offene Bibliotheksjournal / Herausgeber VDB, August, 1-17 Seiten. https://doi.org/10.5282/O-BIB/5835.

Henny-Krahmer, Ulrike, und Patrick Sahle. 2019. „Einreichungen zur DHd 2018“. *DHd-Blog*. 29. März 2019. https://dhd-blog.org/?p=9001.

Hoenen, Armin. 2019. „Einreichungen zur DHd 2019 II“. *DHd-Blog*. 29. März 2019. https://dhd-blog.org/?p=11418.

Keyes, Os. 2018. „The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition“. *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–22. https://doi.org/10.1145/3274357.

Kiefer, Katharina. 2019. „Einreichungen zur DHd 2019“. *DHd-Blog*. 29. März 2019. https://dhd-blog.org/?p=11358.

Nyhan, Julianne, und Oliver Duke-Williams. 2014. „Joint and Multi-Authored Publication Patterns in the Digital Humanities“. *Literary and Linguistic Computing* 29 (3): 387–99. https://doi.org/10.1093/llc/fqu018.

Pichler, Axel, und Nils Reiter. 2022. „Form Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities”. Journal of Cultural Analytics, 7(4). https://doi.org/10.22148/001c.57195.

Posner, Miriam. 2017. „Derive gender from a column of first names “. *Introduction to Digital Humanities*. 2017. http://miriamposner.com/classes/dh101f17/tutorials-guides/data-manipulation/derive-gender-from-a-column-of-first-names/.

Sahle, Patrick, Hrsg. 2019. DHd 2019: *Digital Humanities: multimedial & multimodal*. Zenodo. https://doi.org/10.5281/ZENODO.2596095.

Schöch, Christof, Hrsg. 2020. *DHd 2020: Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. Zenodo. https://doi.org/10.5281/ZENODO.3666690.

Spiro, Lisa. 2012. „‘This Is Why We Fight’: Defining the Values of the Digital Humanities“. In *Debates in the Digital Humanities*, herausgegeben von Matthew K. Gold, 16–35. University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0003.

Stiegler, Johannes, Hrsg. 2015. *DHd 2015: Von Daten zu Erkenntnissen. Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation*. Zenodo. https://doi.org/10.5281/ZENODO.3684490.

Stolz, Michael. 2017. *DHd 2017: Digitale Nachhaltigkeit*. Zenodo. https://doi.org/10.5281/ZENODO.3684825.

Tello, José Calvo. 2016. „DHd 2016: countries, cities and institutions of the speakers“. *CLIGS*. 11. März 2016. https://cligs.hypotheses.org/431.

Vogeler, Georg, Hrsg. 2018. *DHd 2018: Kritik der digitalen Vernunft*. Zenodo. https://doi.org/10.5281/ZENODO.3684897.

Weingart, Scott. 2015. „Submissions to DH2016 (Pt. 1)“. *The Scottbot Irregular*. 7. Dezember 2015. http://www.scottbot.net/HIAL/?p=41533.

Weingart, Scott B., und Nickoal Eichmann-Kalwara. 2017. „What’s Under the Big Tent?: A Study of ADHO Conference Abstracts“. *Digital Studies/Le Champ Numérique* 7 (1): 6. https://doi.org/10.16995/dscn.284.

Yalçın, Mehmet Adil, Niklas Elmqvist, und Benjamin B. Bederson. 2016. „Keshif: Out-of-the-box visual and interactive data exploration environment“. In *Proc. of IEEE VIS 2016 Workshop on Visualization in Practice: Open Source Visualization and Visual Analytics Software*.

DigEdTnT - Digital Edition Creation Pipelines: Tools and Transitions Optimierung digitaler Editionsworkflows: Erfahrungen und Herausforderungen im Projekt DigEdTnT

Steiner, Christian

christian.steiner@dhcraft.org
Digital Humanities Craft OG, Österreich
ORCID: 0000-0002-6658-4622

Pollin, Christopher

christopher.pollin@dhcraft.org
Digital Humanities Craft OG; Universität Graz, Österreich
ORCID: 0000-0002-4879-129X

Strutz, Sabrina

sabrina.strutz@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-7745-0558

Reiter, Georg

georg.reiter@uni-graz.at
Universität Graz, Österreich

Klug, Helmut
helmut.klug@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-7461-5820

Digitale Editionen dienen der Erschließung historischer Dokumente im digitalen Paradigma *(Sahle, 2013)* und haben zum Ziel, diese einer breiteren Öffentlichkeit zugänglich zu machen. Sie umfassen Text-, Bild- und ggf. quantitative Daten und erfordern häufig spezifische Schnittstellen und Werkzeuge, um fachspezifische Forschungsfragen bearbeiten zu können. Trotz der spezifischen Anforderungen jedes Editionsprojekts gibt es allgemeine Arbeitsschritte wie Transkription, Annotation, Normalisierung und Publikation. Im Rahmen des Projekts "Digital Edition Creation Pipelines: Tools and Transitions" (DigEdTnT) werden zwei repräsentative Workflows untersucht. Der erste konzentriert sich auf mittelalterliche Kochrezepte (http://gams.uni-graz.at/corema), der zweite auf Korrespondenzen aus dem 19. und 20. Jahrhundert (https://gams.uni-graz.at/hsa).

Im Zuge dieser Beispielprojekte werden Tools für die allgemeinen Arbeitsschritte getestet und beschrieben. Die detaillierten Ausarbeitungen dieser Best-Practice-Workflows und insbesondere der Übergänge (Transitions) von einem Tool zum anderen werden bis zum Projektende laufend unter https://digedtnt.github.io veröffentlicht. Dort finden sich einerseits Beschreibungen und Tutorials zu den einzelnen Tools, die es ermöglichen, deren Stärken und mögliche Herausforderungen einzuschätzen. Andererseits wird auch der für die Transitionen notwendige Code zur Verfügung gestellt und die dabei auftretenden Probleme beschrieben.

Ziel des Vortrags ist es, die Ergebnisse der Best-Practice-Workflows vorzustellen, zu kontextualisieren und kritisch zu reflektieren. Der Fokus liegt dabei auf den zuvor genannten Arbeitsschritten, die folgend näher erläutert werden. Die Werkzeuge FromThePage, Transkribus Lite, ediarum, FairCopy und OpenRefine wurden aufgrund unserer langjährigen Erfahrung in der Abwicklung von Editionsprojekten und der Entwicklung von Workflows ausgewählt. Ihre große Sichtbarkeit und Akzeptanz in der Community spielten ebenfalls eine wesentliche Rolle. Diese Tools decken verschiedene Aspekte der digitalen Editionsarbeit ab – von Transkription über Annotation bis hin zur Publikation.

Transkription

Die Transkription ist ein essenzieller Schritt in der digitalen Editionserstellung, wobei verschiedene Methoden und Werkzeuge mit jeweiligen Stärken und Schwächen zur Verfügung stehen. Im Zuge des DigEdTnT-Projektes werden dabei **zwei Tools** eingehend betrachtet: **FromThePage** und **Transkribus**.

FromThePage *(Guzman, 2019; Brumfield, 2012)* richtet sich speziell an Projekte, die auf gemeinschaftlicher Transkription *(Terras, 2016)* beruhen. Aufgrund seiner Benutzerfreundlichkeit ist es ideal für Personen mit grundlegenden Computerkenntnissen geeignet, um an der Transkription von Manuskripten mitzuwirken. Es bietet Zugang zu einer aktiven “Transkriptionsgemeinschaft” und ermöglicht die Zusammenarbeit an umfangreichen Sammlungen.

Hinsichtlich des Einsatzes von **FromThePage** für die Transkription von **Rezeptmanuskripten** hat sich im Projektverlauf gezeigt, dass der größte Vorteil dieses Tools darin liegt, dass es bereits mit minimalen Computerkenntnissen bedient werden kann und somit eine community-basierte Zusammenarbeit an umfangreichen Sammlungen möglich ist. Der TEI/XML-Export hingegen war zum Zeitpunkt der Erkundung des Tools weniger ausgereift, da zum einen in der Transkriptionsansicht keine Validierungsmöglichkeiten (z.B. bei Tippfehlern in Tags oder Positionierung von Tags an unzulässigen Stellen) vorhanden sind und zum anderen bei der Metadatenbeschreibung über die Benutzeroberfläche nicht klar wird, welchen TEI-Header-Elementen die Eingaben später im Output entsprechen könnten. Die zusätzliche Betrachtung der von FromThePage beworbenen Annotations- und Indexierungsmöglichkeiten hat zudem gezeigt, dass dieses Tool für umfassende Annotationen weniger geeignet ist.

Transkribus Lite *(Alvermann und Gut, 2021; Böhm und Gerhardt, 2019)* setzt im Gegensatz dazu auf maschinelles Lernen für die automatisierte Transkription *(Ó Raghallaigh und Palandri und Mac Cárthaigh, 2022; Walker, 2021)*. Es können entweder bereits trainierte öffentliche Layout- oder Schrifterkennungsmodelle verwendet oder eigene trainiert werden, um gedruckte und handgeschriebene Texte zu transkribieren. Die Anpassungsfähigkeit der selbst trainierten Modelle an die jeweiligen Handschriften macht Transkribus zu einem leistungsfähigen Werkzeug.

Der Einsatz von **Transkribus Lite** zur automatischen Transkription der handschriftlichen **Briefe** hat gezeigt, dass das Tool in seinem Kernanwendungsbereich sehr gute Ergebnisse liefert. Obwohl die Trainingsdaten, mit denen das Handschriftenerkennungsmodell trainiert wurde, teilweise heterogen waren - z.B. waren einige Briefe schwerer lesbar oder die Handschrift hatte sich im Laufe der Jahre verändert - konnte eine sehr gute Zeichenfehlerrate erreicht werden. Durch die Verwendung bereits trainierter und öffentlich verfügbarer Modelle als Trainingsgrundlage konnten diese guten Ergebnisse sogar noch verbessert werden. Auch das Erreichen der für das Modelltraining bei handschriftlichen Texten empfohlenen Anzahl von mindestens 10.000 Wörtern pro Hand erweist sich durch die Möglichkeit des kollaborativen Arbeitens an den Dokumenten als leicht erreichbar. Ein großer Nachteil ist allerdings, dass die Transkription des Trainingsmaterials für die Texterkennungsmodelle innerhalb von Transkribus Lite erfolgen muss, da in die Zwischenablage kopierter Text nur in einzelne Zeilen des Editors eingefügt werden kann und dieser Inhalt nicht in die anderen Zeilen umgebrochen werden kann. Im Texteditor können auch Annotationen mit vordefinierten Tags vorgenommen werden, durch die Erstellung eigener Tags mit optionalen Attributen kann sogar eine Kon-

formität mit den TEI-Richtlinien erreicht werden, so dass Transkribus Lite zunächst auch für die Annotation geeignet erscheint. Wie bei FromThePage hat sich aber auch bei Transkribus Lite gezeigt, dass das Tool bei komplexeren und umfangreicheren Annotationen schnell an seine Grenzen stößt. Grundsätzlich fehlt eine Validierungsfunktion, zudem kann nicht seitenübergreifend annotiert werden und verschachtelte und zeilenübergreifende Annotationen führen im TEI/XML-Export zu unerwünschten Ergebnissen, sodass es zielführender ist, auf solche Annotationen zu verzichten und sie nur in explizit dafür ausgelegten Annotationswerkzeugen vorzunehmen. Generell besteht die Notwendigkeit, den - in der Regel nicht validen - TEI/XML-Export durch XSL-Transformationen aufzubereiten, bevor er mit dedizierten Annotationswerkzeugen weiterverarbeitet wird.

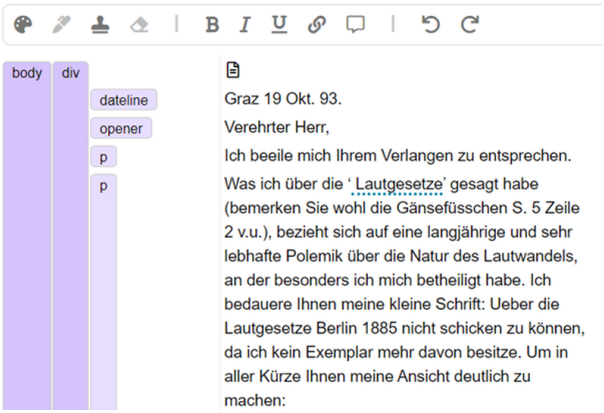
Annotation

Die Annotation von Texten umfasst den Prozess, dem transkribierten Text zusätzliche Informationen – semantische, strukturelle, interpretative – hinzuzufügen. Dieser Prozess erfordert oft spezielle Tools, die auf die jeweiligen Bedürfnisse und Anforderungen der Editionsprojekte zugeschnitten sind. Die Tools **ediarum** und **FairCopy** wurden dafür erprobt.

ediarum *(Mertgens, 2020; Vetter, 2022)* ist eine seit 2012 entwickelte digitale Arbeits- und Publikationsumgebung, die aus mehreren Softwarekomponenten besteht und im Wesentlichen einen Werkzeugkasten aus verschiedenen Modulen darstellt. Damit bietet ediarum eine **Schnittstelle** zwischen Editionsumgebung, XML-Datenbank und Rechercheportal, wobei die Kernfähigkeit in jedem Fall die Aufbereitung von XML-Dateien ist. Die digitale Arbeitsumgebung basiert auf einer **eXist-db** und ermöglicht neben der Transkription von Handschriften und Drucken vor allem die TEI-konforme Annotation und Erstellung von Text- und Sachapparaten sowie Registern. ediarum ist als Add-on zu Oxygen konzipiert und verfügt seit 2015 über eine generalisierte Eingabeoberfläche. Zu beachten ist, dass ediarum keine Plug-and-Play-Software ist, da für die Implementierung und den Betrieb von ediarum immer ein:e (DH-)Entwickler:in nötig ist. Der größte Vorteil von ediarum liegt aber jedenfalls darin, dass Transkriptionen sehr benutzerfreundlich mit TEI-konformem XML in einer gut individualisierbaren Editionsumgebung ausgezeichnet werden können.

Im Zuge des DigEdTnT-Projekts wurde von den diversen ediarum-Modulen für die Annotation der **Rezepte** das Hauptaugenmerk auf **ediarum.BASE.edit** gelegt. Um jedoch die in FromThePage transkribierten Dokumente, die bereits basal annotiert wurden, für die weitere Bearbeitung in ediarum vorzubereiten, wurde im Projektworkflow der **Übergang** mit einer XSL-Transformation gestaltet und beschrieben. Bei der Arbeit an den Rezept-Transkripten mit ediarum fiel besonders positiv auf, dass für Projektmitarbeitende die benutzerfreundliche Autoransicht eine Umgebung bietet, die es auch weniger computerversierten Personen ermöglicht, in einem Editionsprojekt mitzuwirken. Für eine auf das Projekt zugeschnittene Benutzeroberfläche oder das Anlegen neuer Buttons für projektspezifische Annotationen sind aber jedenfalls die Fähigkeiten einer (DH-)Entwickler:in erforderlich.

Bei **FairCopy** handelt es sich um einen erst wenige Jahre alten **TEI/XML-Editor**, der von der US-amerikanischen Softwarefirma Performant Software Solutions LLC entwickelt wird und sich als eine kostengünstigere Alternative zu etablierten Werkzeugen wie Oxygen präsentiert. Ein wesentlicher Unterschied zu anderen Editoren besteht darin, dass die TEI-Elemente nicht wie üblich angelegt werden, indem Text markiert und mittels Tastatureingabe von einem Tag umschlossen wird, sondern die Textpassagen werden zunächst mittels Maus markiert und dann durch die Auswahl des gewünschten Elements über verschiedene Dropdown-Menüs annotiert. Ein großer Teil der Elemente, die in FairCopy sogenannten Textstrukturelemente (wie z. B. ‘div’ oder ‘p’), werden gar per Drag-and-drop an die gewünschte Position links des Textes gezogen, wo sie dann als eingefärbte Kästchen erscheinen (siehe Screenshot).



Bei der Transition von Transkribus Lite zu FairCopy war eine XSL-Transformation des TEI-Exports aus Transkribus Lite nötig, da er nicht valide war und unnötige TEI-Elemente enthielt. Die Anwendung von FairCopy hat sich als etwas umständlich und zeitaufwendig erwiesen, da Tags über Toolbarmenüs ausgewählt werden müssen und Annotierende lernen müssen, wo die benötigten Elemente zu finden sind. Zudem sind nicht alle TEI-Elemente standardmäßig in den Submenüs enthalten, was eine individuelle, allerdings problemlos mögliche, Konfiguration erfordert. Die Submenüs sind an die Projekterfordernisse anpassbar. Eine Kenntnis der TEI-Guidelines ist vorteilhaft, um die benötigten Elemente schnell zu finden. FairCopy garantiert die Dokumentvalidität, indem es die Platzierung von Struktur-Elementen nur dort zulässt, wo sie erlaubt sind. Nachteile sind, dass Nicht-Strukturelemente nur durch eine gepunktete Unterstreichung gekennzeichnet sind und die Attribute der Elemente nicht ohne Klick darauf erkennbar sind. Bei stark verschachtelten Textstrukturen leidet

die Übersichtlichkeit und die Suchfunktion ist rudimentär, ohne komplexere Funktionen wie Suchen und Ersetzen.

Normalisierung

Digitale Editionen produzieren Forschungsdaten und machen diese idealerweise nach FAIR -Kriterien (Findability, Accessibility, Interoperability und Reusability) zugänglich. Dies erfordert die Integration von Normdaten oder kontrollierten Vokabularen, wobei die Normalisierung ein Schritt ist, der normalerweise parallel und iterativ zur Annotation durchgeführt wird. Bei der Annotation von Personennamen z. B. ist es auch fruchtbar, den entsprechenden Wikidata- oder GND-Identifier hinzuzufügen. Hierfür können Werkzeuge wie OpenRefine oder ba[sic?] zur Anwendung kommen.

OpenRefine ist ein Werkzeug zur Datenbereinigung und -transformation, das auch zur (halbautomatischen) Verknüpfung von Daten mit Normdatenbanken verwendet werden kann. Es ist nützlich für geisteswissenschaftliche Forschungsdaten. OpenRefine eignet sich besonders für die Arbeit mit Normdaten, da es ein flexibles Datenmodell für komplexe Strukturen bietet, Datenabgleiche mit Referenzdaten wie Wikidata oder GND unterstützt (Reconciliation) und vielfältige Transformationsoptionen bietet.

Für die Normalisierung der Daten der Beispielprojekte wurde OpenRefine gewählt, um die **Rezepttranskripte** mit Wikidata-Einträgen für die Zutaten anzureichern. Die Zutaten in frühneuhochdeutscher, hochdeutscher und englischer Schreibweise wurden in OpenRefine mit **Wikidata**-Q-Nummern angereichert und in eine leicht weiterverarbeitbare XML-Struktur exportiert. Obwohl die Bedienung von OpenRefine kaum technische Kenntnisse erfordert, sollte der Zeitpunkt des Einsatzes sorgfältig überlegt werden, insbesondere wenn die Daten in ediarum annotiert werden sollen. Die **Template-Option** in OpenRefine ermöglicht eine Anpassung des Exports an die Registerstruktur von ediarum. In diesem Beispielprojekt waren jedoch geringfügige projektspezifische Nachbearbeitungen des OpenRefine-Outputs mittels XSLT notwendig, da ein "Zutatenregister" in ediarum keine Standardlösung ist.

Für die halbautomatische Verknüpfung mit Normdaten können auch andere Werkzeuge wie ba[sic?] herangezogen werden. Dieses Tool ermöglicht die Verknüpfung von Daten mit spezifischen Vokabularen oder Normdatensätzen und kann auch zur Qualitätskontrolle von Daten eingesetzt werden.

Publikation

Forschungsdateninfrastrukturen und Publikationssoftware sind unverzichtbare Komponenten bei der Realisierung digitaler Editionen. Die Publikationssysteme, die im Rahmen von DigEdTnT näher untersucht werden, sind **ediarum.WEB** und der **teiPublisher**.

Der teiPublisher ist ein Open-Source-Framework, das speziell für die Publikation von TEI-Daten entwickelt wurde. Dieses Framework ermöglicht die Darstellung von TEI XML Dokumenten.

ediarum.WEB ist ein Werkzeug zur Publikation digitaler Editionen, das speziell für Projekte entwickelt wurde, die das ediarum-Backend nutzen. Es bietet eine Reihe von Funktionen zur Visualisierung von TEI XML Dokumenten, einschließlich der Unterstützung verschiedener Textansichten und Register.

Zum Zeitpunkt der Antragstellung wurden weder teiPublisher noch ediarum.WEB in den exemplarischen Workflows untersucht. Sowohl die projektspezifischen Beschreibungen dieser Tools als auch die Transitionen werden jedoch Inhalt dieser Präsentation sein und bis zum Zeitpunkt der DHd 2024 auf der DigEdTnT-Website veröffentlicht werden.

Zusammenfassung

Der Bedarf an Best-Practice-Leitfäden ergibt sich aus der zunehmenden Komplexität digitaler Editionsprojekte, die sich über alle Arbeitsschritte erstreckt. Jedes dieser Werkzeuge hat seine spezifischen Stärken und Herausforderungen, eignet sich für unterschiedliche Anwendungsfälle und wird aufgrund seiner besonderen Eigenschaften und Funktionen für bestimmte Editionsprojekte bevorzugt. Die Wahl des richtigen Werkzeugs hängt von den spezifischen Anforderungen des jeweiligen Projekts ab und kann wesentlich zur Qualität und Nachhaltigkeit der digitalen Edition beitragen.

Unser Ansatz zur Erprobung dieser Werkzeuge ist community-basiert: Entwickler:innen und Nutzer:innen treffen sich in zwei Workshops und mehreren Webinaren, um gemeinsam Anwendungsfälle und Feedback zu diskutieren. Der einleitende Workshop, der vom 23. bis 24. Februar 2023 abgehalten wurde, bot eine Fülle von Tools, die direkt von deren Entwickler:innen präsentiert wurden. Auf die Vorstellungen folgte eine "Tool-Dating"-Session, bei der die Fragen der Nutzer:innen von den Entwickler:innen beantwortet wurden.

Zu guter Letzt werden die Herausforderungen in Bezug auf die Nachhaltigkeit und Langzeitverfügbarkeit von Forschungsdateninfrastrukturen, insbesondere im Kontext von eXist-basierten Systemen wie teiPublisher und ediarum.Web im Vortrag diskutiert. Es wird dargelegt, dass keine Publikationslösung ohne regelmäßige Wartung auskommt.

Bibliographie

Alvermann, Dirk und Paweł Gut. 2021. Transkribus im Archiv – Ein polnisch-deutsches Projekt zur Handschriftentexterkennung an historischen Dokumenten. Archeion 122: 129–153. doi:10.4467/26581264ARC.21.006.14486, https://

www.ejournals.eu/Archeion/2021/122/art/20695/ (zugegriffen: 22. Juni 2023).

Brumfield, Ben. 2012. FromThePage: A Web-Based Tool for Transcribing, Indexing, and Annotating Handwritten Material. Präsentiert in Chicago, 7. Januar.

Ó Raghallaigh, Brian, Andrea Palandri und Críostóir Mac Cárthaigh. 2022. Handwritten Text Recognition (HTR) for Irish-Language Folklore. In Proceedings of the 4th Celtic Language Technology Workshop within LREC2022, 121–126. Marseille, France: European Language Resources Association, Juni. https://aclanthology.org/2022.cltw-1.17 (zugegriffen: 22. Juni 2023).

Sahle, Patrick. 2013. Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. [Finale Print-Fassung]. Bd. 8. Norderstedt: BoD. (zugegriffen: 14. Februar 2023).

Terras, Melissa. 2016. Crowdsourcing in the Digital Humanities. In A New Companion to Digital Humanities, 2nd Edition, hg. von Susan Schreibman, Ray Siemens, und John Unsworth, 420–439. Wiley-Blackwell, 6. Januar. doi:10.1002/9781118680605.ch29.

Vetter, Angila. 2022. ediarum.MEDIAEVUM. Eine Arbeitsumgebung zur Edition mittelalterlicher (Prosa)Texte. Beiträge zur mediävistischen Erzählforschung (28. November): 47–64. doi:10.25619/BmE20223194, https://ojs.uni-oldenburg.de/ojs/index.php/bme/article/view/194 (zugegriffen: 14. Februar 2023).

Digitale Editionen und ihre (potenziellen) Nutzer*innen. Konzeptionelle Überlegungen für ein Editions-Registers

Esch, Claudia
claudia.esch@uni-bamberg.de
Universität Bamberg, Deutschland

Problemstellung

Digitale Editionen stellen ein wichtiges und nach wie vor dynamisches Forschungsfeld innerhalb der DH dar (Burckhardt et al., 2023). Während das Potenzial digitaler Editionen nicht nur theoretisch fundiert (Sahle, 2013), sondern auch innerhalb der DH-Community unbestritten ist, ist ihre Rezeption in den Fachwissenschaften zum Teil verhalten (O'Sullivan et al., 2022; Del Rosselli Turco, 2016). Offenbar gibt es eine Diskrepanz zwischen den Bedürfnissen der potenziellen Nutzer*innen und den Entwicklungen im Bereich der digitalen Editionen (O'Sullivan et al., 2022). Ein Weg, diese Lücke zu verkleinern, kann die Analyse der konkreten Erwartungen an digitale Editionen aus Nutzer*innensicht sein (Franzini et al., 2019; Leblanc, 2018; Caria und Mathiak, 2018; Resch und Rastinger). Diese Forschungen sind immens wichtig, stellen allerdings die digitale Edition als Zielobjekt in den Vordergrund. Im Folgenden sollen digitale Editionen dagegen als einer von mehreren Zugängen zu Texten im Arbeitsalltag von Forscher*innen betrachtet werden. Denn neben der wachsenden Zahl an digitalen Editionen werden auch gedruckte und digitalisierte Editionen noch lange eine wichtige Rolle für die fachliche Auseinandersetzung mit Texten spielen. Das legen nicht nur bisherige Studien nahe (Porter, 2013, 2016), sondern auch die in der Praxis zu beobachtende Langlebigkeit von Editionen. So liegen in vielen Disziplinen zentrale Texte noch immer in Editionen aus dem 19. Jahrhundert vor, während zugleich heute noch wissenschaftliche Editionen ausschließlich im Druck erscheinen (Burkhardt, 2020). Dennoch findet die Diskussion über digitale Editionen meist eher innerhalb der DH-Community statt. So sind die aktuell wichtigsten Rechercheportale (Sahle, 2022; Franzini, 2012) vorwiegend aus einem Interesse an der Entwicklung digitaler Editionstechniken hervorgegangen, was sich auch in deren Struktur wiederspiegelt. Dies verstärkt die Tendenz, digitale Editionen losgelöst von anderen Editionsformaten zu betrachten. Es wird hier jedoch argumentiert, dass es eine inklusive Perspektive braucht, um potenzielle Nutzer*innen aus den Fachdisziplinen dort abzuholen, wo ihre Bedürfnisse und Interessen liegen. Ein Schritt in diese Richtung könnte ein Register für gedruckte, digitalisierte und digitale Editionen sein, wie es im Rahmen des NFDI-Konsortiums Text+ entsteht. Im Folgenden wird ein Konzept vorgestellt, das diesen Ansatz in den Fokus rückt und auf einer Analyse der bisherigen Suchstrategien von Forscher*innen bezüglich Editionen und ihren Erwartungen an ein Register basiert.

Projekthintergrund

Datengrundlage für die Ausführungen sind dreizehn semistrukturierte Interviews, die Ende 2022 mit Forscher*innen aus verschiedenen Disziplinen geführt wurden. Die Interviews fanden im Rahmen eines Projekts statt, das als Masterarbeit an der Universität Bamberg eingereicht wurde, und einen Vorschlag für ein nutzerorientiertes Konzept für die Registry der Taskforce Editionen des NFDI-Konsortiums Text+ erarbeitete (Esch, 2023). Die Registry ist als kuratiertes Verzeichnis für die im deutschen Forschungs- und Förderraum angesiedelten Editionen, unabhängig von ihrem Medienformat, geplant. Neben der Verbesserung der Zugänglichkeit von Editionen ist auch eine Einwirkung auf die Praxis digitalen Edierens durch

die Sichtbarmachung von Best-Practice-Beispielen angestrebt (Blumtritt et al., 2023). Damit besitzt die Registry das Potenzial, Verbindungen zwischen der fachwissenschaftlichen Nutzung von Editionen und der Konzeption und Weiterentwicklung digitaler Editionen herzustellen, die in der Praxis oftmals noch als unterschiedliche Bereiche wahrgenommen werden (Problems with Digital Scholarly Editions). Daher wurde bei der Auswahl der Interviewpartner*innen darauf geachtet, Personen sowohl aus verschiedenen geisteswissenschaftlichen Disziplinen als auch mit unterschiedlichen Erfahrungen im Bereich digitalen Arbeitens einzubeziehen, um einen Querschnitt potenzieller Nutzer*innen von digitalen Editionen zu erfassen. Die unterscheidet die Befragungen von den Interviews, die im Kontext des „C21 Edition“-Projekts überwiegend mit Expert*innen aus dem Bereich digitalen Edierens geführt wurden (O'Sullivan und Kurzmeier, 2023).

Ausgewählte Ergebnisse

In den Interviews wurde unter anderem die aktuelle Nutzung verschiedener Editionsarten erfasst. Hier konnten die Ergebnisse älterer Studien dahingehend bestätigt werden, dass genuin digitale Editionen, wie sie von Patrick Sahle definiert werden (Sahle, 2013), aktuell nur einen kleinen Teil der verwendeten Editionen ausmachen (Porter, 2016). Digitalisierte Editionen werden dagegen häufiger genutzt. Auch das fügt sich in die Resultate früherer Studien ein (Caria und Mathiak, 2018).

Interessant waren insbesondere die Angaben zu den Gründen für das Nutzungsverhalten. Trotz der faktischen Dominanz gedruckter bzw. digitalisierter Editionen wurden kaum grundsätzliche Bedenken gegen digitale Editionen geäußert. Die durchaus vorhandenen, zum Teil noch nicht vollkommen gelösten Herausforderungen digitaler Editionen, wie etwa langfristige Verfügbarkeit und Referenzierbarkeit (Del Rosselli Turco, 2016; Stäcker, 2020), spielten eine untergeordnete Rolle. Vielmehr wurden die mangelnde Verfügbarkeit digitaler Editionen im eigenen Fachbereich sowie eine situative Nutzung verschiedener Medienformen hervorgehoben. Während die Durchsuchbarkeit und der schnelle Zugriff als Vorteile digitaler Editionen genannt wurden, galt einigen Teilnehmer*innen das Druckwerk oder eine PDF-Druckversion als bester Zugang für eine intensive Beschäftigung mit einzelnen Textpassagen.

Angelehnt an die Unterscheidung zwischen medialer und konzeptioneller Digitalität von Torsten Hiltmann (2022) ließe sich folgern, dass viele Anwender*innen in ihrer Nutzung von Editionen noch im Bereich der ersteren bewegen, in der die digital verfügbaren Daten vorwiegend zur Simulation analoger Quellen dienen. Eine datenzentrierte Nutzung von Editionen, also eine Anwendung im Bereich der konzeptionellen Digitalität, spielte dagegen nur bei Befragten aus der Sprachwissenschaft oder unmittelbar aus dem Bereich der DH eine größere Rolle. Im Bereich der medialen Digitalität erfüllen digitalisierte Editionen jedoch bereits viele der Ansprüche, so dass der höhere Aufwand bei digitalen Editionen oft nicht in Einklang mit dem (momentan wahrgenommenen) Nutzen steht. Wie aus den Interviews hervorgeht, wird sowohl bei der Erstellung als auch bei der Nutzung digitaler Editionen ein höherer Aufwand wahrgenommen. Dieser ist auf der Produktionsseite einerseits durch technische Herausforderungen, aber auch durch gestiegene methodische Ansprüche an ein flexibles Textverständnis (Sahle, 2013) und die daraus resultierende Forderung nach Transkription möglichst vieler Textzeugen bedingt, während bei der Anwendung die noch fehlende Standardisierung von Interfaces eine Hürde darstellt.

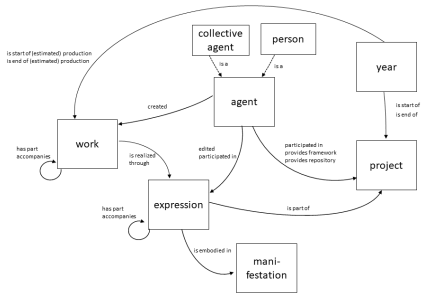
Die Befunde verweisen auf ein grundlegendes Problem. Die Verhaftung vieler digitaler Editionen im methodischen Print-Paradigma wurde bereits vielfach beobachtet (van Zundert 2016), wobei oft Versuche zur Behebung dieses als Defizit wahrgenommenen Zustands im Fokus stehen (Cugliana, 2023). Eine solche fortschrittsorientierte Sichtweise, die von einer logischen Entwicklung vom Print-Paradigma zur datenbasierten Edition ausgeht, kann jedoch unter Umständen den Blick auf andere Zusammenhänge verdecken. Wie Patrick Sahle überzeugend darlegt (Sahle, 2013) gibt es keine objektive Gestaltung von Editionen, da das zu Grunde liegende Textverständnis implizit oder explizit die editorischen Entscheidungen beeinflusst. Die Möglichkeit zur Darstellung multipler Textbegriffe, die durch die Transmedialität digitaler Editionen ermöglicht wird, ist sicherlich einer der großen Vorteile dieser Editionsform (Vogeler, 2019). Daraus sollte jedoch keine neue Objektivität des Digitalen abgeleitet werden. Vielmehr erfordert die Umsetzung eines bestimmten digitalen Paradigmas ebenfalls editorische Entscheidungen, die zu einer veränderten Schwerpunktsetzung auch im Hinblick auf das Textverständnis führen, was sich etwa in der Tendenz zu hyperdiplomatischen Editionen oder der gestiegenen Bedeutung der Materialität (Cappellotto, 2020) ausdrückt. Es ließe sich vielleicht sogar überlegen, ob der Blick auf Texte als Daten nicht ein Textverständnis darstellt, das Patrick Sahles Textrad ergänzen kann.

Die Rolle des Textverständnisses wird im Kontext der Frage nach der Akzeptanz digitaler Editionen jedoch eher selten diskutiert, obwohl die Bewertung der Vorteile digitaler Editionen – so die hier vorgetragene These – auch vom zu Grunde gelegten Textbegriff abhängt. So kann die Entscheidung für dem Buchwesen stark verwandte digitale Editionen bei einer Priorisierung des Werkcharakters eine innerhalb des Fachkontextes durchaus rationale Kosten-Nutzen-Abwägung sein (Huskey, 2022), vor allem angesichts des erhöhten Ressourcenaufwand digitaler Editionen (Šimek, 2022). Zwar ermöglicht die Transmedialität digitaler Editionen grundsätzlich auch Printformate als zusätzliche Ausgabeform, doch ist das Ausgabeformat nur ein Teil des Problems. Vielmehr hat die Verwirklichung eines multiplen Textbegriffs eben ihren Preis, dessen Bewertung vom erwarteten Nutzen und damit auch vom zu Grunde liegenden Textverständnis abhängt. Letztlich konkurrieren also weniger Printwerke und digitale Formate als vielmehr verschiedene Textverständnisse.

Unter diesem Blickwinkel könnten sich erweiterte Diskussionsräume für DH und Fachwissenschaften eröffnen, die den Abstand zwischen den Communities überbrücken helfen. Denn so ließe sich die verhaltene Rezeption digitaler Editionen nicht (nur) mit einem Informationsdefizit auf Seite der Fachwissenschaften erklären, sondern als Resultat einer auf dem dominanten Textverständnis beruhenden Grundhaltung verstehen, die sich aus den jeweiligen Forschungsfragen und den Fachmethoden speist. Diese sind grundsätzlich wandelbar, doch sind die zur Verfügung stehenden Editionen nur ein Faktor von vielen. Digitale Editionen, deren Textverständnis sich aus einem überwiegend innerhalb der DH entwickelten digitalen Paradigma speist, müssen erst ihren Weg in das komplexe Zusammenspiel von Textgrundlage, Forschungsfragen und Fachmethoden finden. Es ist anzunehmen, dass eine zunehmende Rezeption digitaler Methoden in den Fachwissenschaften auch einen größeren Bedarf an digitalen Editionen nach sich ziehen wird. Voraussetzung dafür ist aber zum einen zunächst eine gewisse kritische Masse an Editionen, die das Potenzial des digitalen Paradigmas ausschöpfen und die Anwendung digitaler Methoden ermöglichen. Zugleich wird es aber auch weiterhin Fragestellungen und Methoden geben, für die der Mehrwert bestimmter Formen digitalen Edierens überschaubar bleibt. Statt einer Ablehnung digitaler Editionen kann hier aber auch ein Bedarf an spezifischen Formen digitalen Edierens gesehen werden. Ein interessanter Lösungsansatz in diese Richtung stellt etwa das Konzept der *assertive edition* dar, das die Faktenfokussierung der Geschichtswissenschaften konstruktiv aufgreift (Vogeler et al, 2022). Digitale Editionen können also durchaus einen wichtigen Beitrag zum Wandel in den Geisteswissenschaften leisten, doch sind die Erfolgsaussichten umso besser, je intensiver sie in den jeweiligen Fachdiskurs eingebunden sind. Aus dieser Perspektive erscheinen gemeinsame Kommunikationsräume besonders wichtig, um die digitale Edition aus dem DH-internen Resonanzraum herauszuholen und für verschiedene Textverständnisse produktiv zu öffnen bzw. einen Wandel innerhalb der einzelnen Fachbereiche und ihrer Methoden anzustoßen. Ein möglicher Ansatzpunkt könnte ein Editionsverzeichnis sein, das eine Plattform für alle Editionsformate bietet und jenseits der Dichotomie von Print und Digital als praxisorientiertes Hilfsmittel für DH-Forschende ebenso wie für Fachwissenschaftler*innen ohne DH-Affinität funktioniert.

Konsequenzen für das Konzept des Editionsregisters

Die Perspektive auf das edierte Material bietet sich dabei als gemeinsamer Nenner an. Die Fachwissenschaftler*innen gaben in den Interviews an, primär nicht nach Editionen, sondern nach Quellen bzw. Texten zu suchen, idealerweise in der für ihre Zwecke am besten aufbereiteten, notfalls aber auch in einer anderen zugänglichen Form. Bemerkenswerterweise wurde diese Sicht sowohl von DH-affinen als auch eher traditionell aufgestellten Personen geäußert. Ein zentrales Findmittel könnte dabei durchaus auch Defizite bei der Suche nach Editionen in traditionellen Systemen wie Bibliothekskatalogen beheben, da dort nicht nur digitale Editionen unterrepräsentiert sind, sondern auch die getrennte Modellierung von Quellen und Edition zwar im RDA-Standard mitgedacht, aber noch nicht konsequent umgesetzt ist (Gantert und Lauber-Reymann, 2023). Innerhalb ihres engeren Fachgebiets verfügen die interviewten Expert*innen zwar über ausgefeilte Suchstrategien unter Einbezug fach- und themenspezifischer Print- und Onlineressourcen, die diese Defizite kompensieren, sie sehen aber durchaus Bedarf nach einer besseren Verzeichnung von Editionen im Kontext der universitären Lehre oder bei interdisziplinären Forschungsfragen. Die im Vortrag vorgeschlagene Modellierung eines zentralen Editionen-Registers basierend auf einem Entity-Relationship-Modell, beispielsweise angelehnt an IFLA-LRM, würde für die Auffindbarkeit aller Arten von Editionen aus Nutzersicht daher einen Mehrwert bringen (Esch, 2023).



Konzeptionelles Modell (vereinfacht)

Zugleich wurde in den Interviews aber auch der Wunsch nach einer systematischen Erfassung digitaler Editionen geäußert, um die Entwicklung in diesem Bereich zu erfassen und in Form von Best-Practice-Modellen weiterzuentwickeln. Dieser Wunsch ist nicht als Gegensatz zur inhaltsorientierten Perspektive aufzufassen, sondern beide Sichtweisen wurden zum Teil von denselben Personen geäußert. Hier zeigt sich die Verzahnung von Fach- und Editionsperspektive, die für die Brückenfunktion des Registers zentral ist. Das Datenmodell ermöglicht es daher, beide Perspektiven gleichberechtigt zu integrieren.

Neben den oben beschriebenen Chancen, die ein solches Register für die Verknüpfung von DH und Fachwissenschaften bietet, bestehen in Hinblick auf Realisierbarkeit, Vollständigkeit und Nachhaltigkeit allerdings auch erhebliche Risiken. Große Einigkeit bestand unter den Interviewpartner*innen darin, dass eine zumindest angestrebte inhaltliche Vollständigkeit und hohe Aktualität der Einträge wichtige Kriterien für den Erfolg eines solchen Projekts sind, was zu einem hohen Ressourcen-Bedarf führt. Die zeitlich begrenzte Förderung der NFDI-Konsortien stellt zudem für die langfristige Pflege der Daten eine Herausfor-

derung dar. Eine Kooperation mit bereits etablierten Fachressourcen wäre eine mögliche Kompensationsstrategie, insbesondere da eine Verknüpfung mit weiteren Hilfsmitteln und Handschriften von den Interviewteilnehmer*innen als wünschenswert betrachtet wurde. Das Datenmodell sollte daher perspektivisch möglichst anschlussfähig für verschiedene Fachressourcen sein, so dass ein Datenaustausch erfolgen kann. Der hohe Grad an Spezialisierung der Ressourcen und die sich daraus ergebende Zersplitterung der Landschaft bei Hilfsmittel stellt jedoch eine große Herausforderung dar. Es existieren durchaus Projekte wie etwa die Digital Latin Library (Digital Latin Library), deren Katalog mit FRBRoo eine ebenfalls auf FRBR basierende Ontologie nutzt und alle Daten als Linked Open Data zur Verfügung stellt. Weitere Kooperationspartner könnten perspektivisch Angebote wie die Mittelhochdeutsche Begriffsdatenbank (Zeppezauer-Wachauer, 2022) oder das Handschriftenportal (Handschriftenportal) sein, wobei die Schnittstellen meist noch in Entwicklung sind. Im Moment stellen zahlreiche relevante Fachressourcen allerdings noch Datensilos dar. In wie weit das hier vorgeschlagene Modell eine Integration der Daten ermöglicht, muss daher im Moment noch offen bleiben. Eine Weiterentwicklung in Kooperation mit zumindest einigen wichtigen Projekten, die sich der systematischen Erschließung von Quellen widmen, wäre jedoch ein lohnendes Forschungsprojekt und könnte das hier entworfene Ziel einer Vernetzung von Editionen auf der Metadatenebene unabhängig von ihrem Medienformat in größere Nähe rücken. Ob sich dies im Rahmen der Registry von Text+ verwirklichen lässt oder ob hierzu darüberhinausgehende Ressourcen und Strukturen notwendig sind, wäre noch zu eruieren.

Fazit

Das im Vortrag vorgestellte Datenmodell setzt einige wichtige Erkenntnisse um, die sich aus der Befragung von Fachwissenschaftler*innen zur aktuellen Nutzung von Editionen und den Erwartungen an ein zentrales Editions-Register ergeben haben. Die Abstraktion vom Medienformat der Editionen ermöglichte es, die Suchstrategien von Forschenden sowohl auf digitale als auch digitalisierte und gedruckte Editionen auszuweiten und damit eine Brücke zwischen den oftmals getrennt wahrgenommenen Bereichen zu schlagen. Ein zentrales Bindeglied ist dabei die inhaltsorientierte Perspektive, die auf die Recherche des edierten Materials zielt. Die getrennte Modellierung von ediertem Material und Edition im Datenmodell ermöglicht nicht nur eine gezielte Inhaltsrecherche auf Metadatenebene, sondern auch die Integration sowohl des Blicks auf die Eigenschaften digitaler Editionen als auch auf den edierten Inhalt. Beide Perspektiven sind für unterschiedliche Forschungszwecke relevant, wobei sich die Zielgruppen durchaus überschneiden. Damit kann auf pragmatischer, sachorientierter Ebene die Wahrnehmung digitaler Editionen verbessert und ein fruchtbarer Austausch angestoßen werden kann.

Bibliographie

Blumtritt, Jonathan, Elisa Cugliana, Nils Geißler, Philipp Hegel, Kilian Hensen, Jörg Hörnschemeyer, Christoph Kudella, Karoline Lemke, Harald Lordick, Frederike Neuber, Claes Neuefeind, Daniela Schulz, Melanie Elisabeth-H. Seltmann, Martin Sievers, and Tessa Gengnagel. 2023. „Offene Editionen – Die Task Area Editionen im NFDI-Konsortium Text+“ In DHd2023: Open Humanities, Open Culture. 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V., 404–406. Luxemburg; Trier. https://doi.org/10.5281/zenodo.7688632.

Burckhardt, Daniel, Jörg Hörnschemeyer, Mareike König, Julian Schulz, Till Grallert, und Jana Keck. 2023. „Opening Sources – modulare Wege zur Quellenbereitstellung und –edition“ In DHd2023: Open Humanities, Open Culture. 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V., 83–86. Luxemburg; Trier. https://doi.org/10.5281/zenodo.7688632.

Burkhardt, Julia. 2020. Von Bienen lernen: Das Bonum universale de apibus des Thomas von Cantimpré als Gemeinschaftsentwurf : Analyse, Edition, Übersetzung, Kommentar. Regensburg: Schnell + Steiner.

Cappellotto, Anna. 2020. From codex to apps: the medieval manuscript in the age of its digital reproduction. Umanistica Digitale 9, 1-18. https://doi.org/10.6092/issn.2532-8816/11459.

Caria, Federico, und Brigitte Mathiak. 2018. “A Hybrid Focus Group for the Evaluation of Digital Scholarly Editions of Literary Authors” In Digital scholary editions as interfaces, hg. von Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber und Gerlinde Schneider, 267–286. Norderstedt: Books on Demand. http://kups.ub.uni-koeln.de/9085/.

Cugliana, Elisa. 2023. “Coding editions. Computational approaches to the editing of pre-modern texts” In DHd2023: Open Humanities, Open Culture. 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V., 102–105, Luxemburg; Trier. https://doi.org/10.5281/zenodo.7688632.

Del Rosselli Turco, Roberto. 2016. “The Battle We Forgot to Fight: Should We Make a Case for Digital Editions?” In Digital Scholarly Editing: Theories and Practices, hg. von Matthew James Driscoll und Elena Pierazzo, 219–238. Cambridge: Open Book Publishers. https://books.openedition.org/obp/3381.

“Digital Latin Library”, https://catalog.digitallatin.org/ (zugegriffen: 04.12.2023).

Esch, Claudia. 2023. “Forschungsdatenmanagement für Editionen - Standortanalyse und Empfehlungen für den Aufbau eines nationalen Registers im Rahmen von Text+". https://doi.org/10.5281/zenodo.7943845.

Franzini, Greta. 2012. “A Catalogue of Digital Editions”. https://doi.org/10.5281/zenodo.1161425.

Franzini, Greta, Melissa Terras, und Simon Mahony. 2019.„Digital Editions of Text: Surveying User Requirements in the Digital Humanities.” Journal on Computing and Cultural Heritage 12, no 1: 1–23. https://doi.org/10.1145/3230671.

Gantert, Klaus, und Margrit Lauber-Reymann. 2023. „Bibliothekskataloge und Discovery-Systeme“ In Informationsressourcen: Ein Handbuch für Bibliothekare und Informationsspezialisten, hg. von Klaus Gantert und Margrit Lauber-Reymann, 67–98. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110673272-007.

„Handschriftenportal“, https://handschriftenportal.de/info/about (zugegriffen: 04.12.2023).

Hiltmann, Torsten. 2022. „Vom Medienwandel zum Methodenwandel“ In Digital History: Konzepte, Methoden und Kritiken digitaler Geschichtswissenschaft, hg. von Karoline Dominika Döring, Stefan Haas, Mareike König und Jörg Wettlaufer, 13–44. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110757101-002.

Huskey, Samuel. 2022. “The Visual [Re]Presentation of Textual Data in Traditional and Digital Critical Editions” magazén 3, no. 1: 116–138. https://doi.org/10.30687/mag/2724-3923/2022/05/005.

Leblanc, Elena. 2018. “Design of a Digital Library Interface from User Perspective, and its Consequences for the Design of Digital Scholarly Editions: Findings of the Fonte Gaia Questionnaire” In Digital scholary editions as interfaces, hg. von Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber und Gerlinde Schneider, 287–315. Norderstedt: Books on Demand. http://kups.ub.uni-koeln.de/9085/.

O‘Sullivan, James, und Michael Kurzmeier. „C21 Editions Interviews“. Uploaded May 19, 2023, https://www.dhi.ac.uk/data/c21editions (zugegriffen: 04.12.2023).

O‘Sullivan, James, Michael Pidd, Órla Murphy, und Bridgette Wessels. 2022. “Perspectives on the Future of Digital Editions & Publishing” In Digital Humanities 2022, Conference Abstracts: The University of Tokyo, Japan, 25-29 July 2022, 535–536. https://dh2022.adho.org/.

Porter, Dot. 2013. “Medievalists and the Scholarly Digital Edition”. Scholarly Editing 34, http://scholarlyediting.org/2013/essays/essay.porter.html.

Porter, Dot. 2016. “What is an edition anyway? My Keynote for the Digital Scholarly Editions as Interfaces conference”, University of Graz. https://www.dotporterdigital.org/what-is-an-edition-anyway-my-keynote-for-the-digital-scholarly-editions-as-interfaces-conference-university-of-graz/ (zugegriffen: 04.12.2023).

“Problems with Digital Scholarly Editions”, https://www.c21editions.org/problems/ (zugegriffen: 04.12.2023).

Resch, Claudia, und Nina Rastinger. 2022. “Digitale Editionen im Spannungsfeld zwischen Formalisierung und Interpretation: Rezensionen der Online-Zeitschrift RIDE als Gradmesser für die Zukunft“. In DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation: 7. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, hg. von Christof Schöch, Paderborn. https://doi.org/10.5281/zenodo.3666690.

Sahle, Patrick. 2013. Digitale Editionsformen: Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. 3 Bände. Norderstedt: Books on Demand. http://kups.ub.uni-koeln.de/id/eprint/5351.

Sahle, Patrick. 2023. A catalog of Digital Scholarly Editions, last modified November 29, 2023. https://www.digitale-edition.de/exist/apps/editions-browser/index.html (zugegriffen: 04.12.2023).

Stäcker, Thomas. 2020. “A digital edition is not visible– some thoughts on the nature and persistence of digital editions” Zeitschrift für digitale Geisteswissenschaften. https://doi.org/ 10.17175/2020_005.

van Zundert, Joris. 2016. „Barely Beyond the Book?“ In Digital Scholarly Editing: Theories and Practices, hg. von Matthew James Driscoll and Elena Pierazzo, 83–106. Cambridge: Open Book Publishers. https://books.openedition.org/obp/3381.

Vogeler, Georg. 2019. “The assertive edition” International Journal of Digital Humanities 1, no. 2: 309–322. https://doi.org/10.1007/s42803-019-00025-5.

Vogeler, Georg, Roman Bleier, und Christopher Pollin. 2022. „Ich glaube, Fakt ist…: Der geschichtswissenschaftliche Zugang zum Edieren“ In Digital History: Konzepte, Methoden und Kritiken digitaler Geschichtswissenschaft, hg. von Karoline Dominika Döring, Stefan Haas, Mareike König und Jörg Wettlaufer, 171-190. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110757101-010.

Zeppezauer-Wachauer, Katharina. 2022. „50 Jahre Mittelhochdeutsche Begriffsdatenbank (MHDBDB): Eine Jubiläums-Zeitreise zwischen Lochkarten, Pixel- Drachen, relationaler Datenbank und Graphdaten“ In Digitale Mediävistik: Perspektiven der Digital Humanities für die Altgermanistik, hg. von Elisabeth Lienert, Joachim Hamm, Albrecht Hausmann und Gabriel Viehauser, 161–186. Oldenburg: BIS-Verlag, https://doi.org/10.25619/BmE20223203.

„Digital Humanities interessieren uns nicht, das haben wir schon ausgeforscht“ - Resonanz der DH am Beispiel der Theologie

Nunn, Christopher
christopher.nunn@theologie.uni-heidelberg.de
Universität Heidelberg, Deutschland
ORCID: 0000-0001-7208-8636

Das Verhältnis der Theologie zu den Digital Humanities

„DH Quo Vadis ?“ Zur Beantwortung dieser Frage scheint es sinnvoll zu betrachten, welchen Anklang die DH in den wissenschaftlichen Disziplinen gefunden haben und welche Maßnahmen aufgrund dieser Betrachtungen getroffen werden sollten. In diesem Vortrag soll dies am Beispiel der Theologie vorgenommen werden.

Theologische Forschungen spielten vor allem in den frühen Jahren der DH eine herausragende Rolle. So wird der *Index Thomisticus* von Roberto Busa häufig als Gründungsmythos zitiert (Thaller 2017, 3). Dessen Zusammenarbeit mit IBM brachte den DH (bzw. damals Humanities Computing) so viel Popularität, dass sich dessen Verdienste noch heute im *Roberto Busa Prize* spiegeln, mit dem die ADHO alle drei Jahre Digital Humanists für ihr Lebenswerk auszeichnet (s. Jones 2016, 1-2). Es gibt allerdings gute Gründe, andere als Pioniere der DH anzusehen, z. B. Josephine Miles (s. Blaney et al. 2021, 7); Nyhan (2022) problematisiert zudem das Projekt Busas selbst, da die Arbeit vieler — insbesondere weiblicher — Kollaborateure am *Index Thomisticus* nicht gewürdigt wurde.

Ungeachtet dessen wäre aufgrund der Wirkungsgeschichte des *Index Thomisticus* zu erwarten, dass die DH in theologischen Forschungen breit rezipiert werden und die Theologie somit auch aktuell noch eine bedeutende Rolle im Fächerkanon der DH spielt. Faktisch ist sie jedoch allenfalls in dessen Peripherie zu verorten. Literaturwissenschaften wie die Germanistik (Digital Philology) oder Geschichtswissenschaften (Digital History) sind deutlich dominanter in den DH vertreten. Im häufig zitierten Sphären-Modell Sahles (2013, 6) wird die Theologie nicht einmal erwähnt. Man könnte argumentieren, dass Sahle selbst hier nur eine Auswahl trifft und nicht auf Vollständigkeit bedacht ist. Es könnte auch angeführt werden, dass seine Wahl subjektiv ist und z. B. die Philosophie deshalb genannt wird, weil er u. a. diese studiert hat. Dennoch wird deutlich, dass die Theologie keinen großen Eindruck in den DH hinterlassen hat und Theolog*innen weitestgehend der DH-Community fernbleiben. Gramelsberger (2023, 111-121) bietet zudem unter dem Stichwort einer digitalen Philosophie eine Kartierung von Philosophie-Projekten im Kreis der Digital Humanities . Eine vergleichbare Studie ist für die Theologie noch nicht geschrieben – so gesehen ist die Philosophie durchaus einen Schritt weiter. Wie kommt es also, dass die Theologie in den DH nicht mehr zentraler Akteur, sondern nur Randfigur ist und DH im theologischen Forschungsalltag nur eine äußerst marginale Rolle spielt?

Ablehnung von DH

Innerhalb der Theologie finden sich die gleichen Ressentiments gegen die DH wie auch in anderen Geisteswissenschaften (zu deren Skepsis s. z. B. Krämer 2018, 7). Hinzu kommt die Sorge einiger Vertreter*innen, die heiligen Schriften könnten durch digitale Analysen entwertet werden, s. z. B. Stoellger (2021) 110: „Der neue garstige Graben zwischen unsinnlich sinnfrei operierenden digitalen Medien und der habitualisierten sozialen wie religiösen Orientierung am Symbolischen mit Sinn, Sinnlichkeit und vielleicht sogar Wahrheitsfragen ist eine nachhaltige Irritation und Beunruhigung, nicht nur der Geisteswissenschaften oder von Theologie und Kirche.“

Damit zusammenhängend lässt sich der Eindruck gewinnen, dass der Umfang der DH-Möglichkeiten nur in einem sehr reduzierten Maß wahrgenommen wird, etwa beschränkt auf die Erstellung digitaler Editionen und Analyse von Netzwerken. Heil (2022, 104) resümiert z. B. in einem Beitrag zum Thema „Digital Humanities – zwischen Fortschritt und Rückschritt: Ein Standpunkt“: „ Die Probleme liegen allerdings vor allem in der Gefahr eines Rückschritts im Fortschritt, wenn a) veraltete Editionen zum neuen Standard werden, da nur sie digital ohne Copyright verfügbar sind, wenn b) computergestützte Textanalysen sich nur auf einen kleinen Teilbereich der Texte stützen können, da weitere Textcorpora entweder digital nicht aufbereitet oder rechtlich nicht freigegeben sind [...], und wenn c) die Lust an neuen Visualisierungen simplifizierte Textanalysen erzwingt. Trotzdem eröffnen sich natürlich neue Möglichkeiten, die nicht nur in Zeiten von Lockdowns aufgrund von Corona gerne genutzt werden.“

Hier werden einzig Textanalysen berücksichtigt, die häufig durch den fragmentarischen oder unzureichend edierten Zustand ihrer Quellen behindert oder gar grundsätzlich in Frage gestellt werden. Andere Medien wie z. B. Tonaufnahmen oder Videospieltechnologie sind nicht im Blick. Der letzte von Heil angeführte Punkt führt zu einem Problem, das mir vermehrt im Gespräch mit Vertreter*innen bibelexegetischer Wissenschaften begegnet ist. Angesichts von Konkordanzprogrammen wie BibleWorks wird davon

ausgegangen, dass eine weitere Investition in die DH nicht mehr benötigt wird. „DH interessieren uns nicht, das haben wir schon ausgeforscht.“ Hierbei wird übersehen, dass Programmoberflächen die Analysemöglichkeiten beschränken und steuern (vgl. Berry und Fagerjord 2017, 127 zur Interface-Theorie von Johanna Drucker).

Einen letzten Aspekt, der dazu führt, dass sich Theolog*innen nicht mit DH auseinandersetzen, beobachtet Anderson (2019, 76): „My suspicion is that theological scholars may appreciate what their colleagues in other disciplines are doing but see them as irrelevant to theological enquiry.“

Die ausgeprägte Interdisziplinarität der Theologie führt hier dazu, dass das Engagement in den DH den Nachbardisziplinen überlassen wird, um sich vermeintlich auf das theologische Kerngeschäft konzentrieren zu können. In Evaluationen zu Veranstaltungen mit computationellen Methoden lasen wir wiederholt, dass die Studierenden darin keine Relevanz für ihren späteren Berufsalltag erkannten. Es ist daher wichtig, dass DH bereits in den Lehrcurricula implementiert wird, um deren Wert für die Theologie zu verdeutlichen (vgl. Hunze (2021, 117-118) zur Digitalisierung im Theologiestudium.

Andere Inhalte unter dem Label der DH – zum Spektrum einer Digital Theology

Digital History wird von Lässig (2021, 10) wie folgt beschrieben: „ In using the term digital history , then , historians aim to carve out a field that reflects their interest in accessing the digital space from the questions and issues raised in their field , that takes their specific types of primary sources and epistemology into account , and that uses digital technology to answer their research questions without becoming an end in itself . “

Ähnlich verhält es sich auch mit den Digital Classics und anderen auf spezifische Disziplinen zugespitzte Bereiche der DH. Digital Theology hat jedoch eine breitere Verwendung erfahren. Der erste Inhaber einer deutschen Professur für digitale Theologie, Florian Höhne von der Universität Erlangen, ist z. B. kein Digital Humanist, sondern Medienethiker. Seine vollständige Denomination lautet entsprechend „Medienkommunikation, Medienethik und Digitale Theologie.“ Van Oorschot (2020, 165) teilt die Verwendungsweisen von digitaler Theologie in vier verschiedene Bereiche auf: „1. Theologie in digitalen Räumen […]. 2. Theologie mit digitalen Mitteln, Tools oder Methoden […]. 3. Theologische Reflexion auf Digitalisierung […]. 4. Digitaler Wandel der Theologie […]." Nur der zweite Aspekt von digitaler Theologie lässt sich mit der oben genannten Definition von Digital History et al. in Verbindung bringen.

Phillips et al. (2019) versuchen, digitale Theologie als Bestandteil der DH zu definieren, umschreiben hierbei jedoch eher die Analyse religiöser Phänomene in digitalen Kulturen. Derartige Untersuchungsgegenstände rund um das Wechselspiel von Religiosität und Digitalisierung sind auch Schwerpunkt des Forschungsclusters „Digital Religions “ an der Universität Zürich. Hier grenzt man sich allerdings korrekterweise unmittelbar vom Begriff der DH ab, wie etwa Ulshöfer und Kirchschläger (2021, 12) zum Thema digitaler theologischer Ethik ausführen: „ Insofern bietet der Band zwar in seinen Artikeln noch kein Beispiel, wie Digital Humanities -Methoden beispielsweise der Datenauswertung [...] auch für die Ethik fruchtbar gemacht werden können, aber im Sinne einer „digital theological ethics “ soll die doppelte Stoßrichtung von digital theology , die Phillips et al. beschreiben, aufgenommen werden, so dass es darum geht, dass Digitalisierung zu ethisch-theologischen Fragen führt und dass theologische und auch philosophische Ethik durch Digitalisierung hinsichtlich ihrer Themenbereiche und Methoden herausgefordert wird. “

Diese Spielart einer in der Theologie vielfach erprobten Reflexion auf Digitalität kann auch für Diskurse der DH unmittelbar fruchtbar gemacht werden, wenn sie epistemologisch auf diese angewandt wird, wie van Oorschot (2021) demonstriert. Hier kann die Theologie dem „ theoretical turn“ (Burghardt 2020) der DH gerecht werden und infolgedessen einen wertvollen Mehrwert leisten.

DH unter anderem Label

Die Offenheit und Unschärfe des Begriffs der DH führte zu dessen Bezeichnung als „ big tent .“ So kritisiert Terras (2011): „The latest definition , ‚Big Tent Digital Humanities ‘, deliberately obfuscates the focus of the field . Roll up roll up ! Everything is Digital Humanities ! Everyone is a digital humanist ! The concept of a , big tent ‘ to demarcate a group of individuals is a pragmatic and flexible description usually used to give strength in numbers , permitting a broad spectrum of views or approaches across the constituency .“

Hierin mag begründet sein, weshalb sich in der Theologie durchaus computationell gestützte, teils auch sehr elaborierte Projekte finden lassen, die sich den DH jedoch nicht explizit zuschreiben. Zu nennen sind hier nicht nur zahlreiche digitale Editionen (z. B. das patristische Textarchiv an der Berlin-Brandenburgischen Akademie der Wissenschaften), sondern auch kollaborative Projekte wie *Cursor, die Zeitschrift für explorative Theologie* , VREs wie der *New Testament Virtual Manuscript Room* oder Projekte um DH-typische Fragestellungen wie z. B. *GenderVarianten – Revisionen von Genderkonstruktionen in Textüberlieferungen* . Am deutlichsten wird dies bei Sutinen und Cooper (2021), die in einer Monographie das Phänomen der Digital Theology umfassend beleuchten und in einem Kapitel „ How to Research Digital Theology " exemplarisch Methoden anführen, die zum etablierten Analysebesteck der DH gehören, hier jedoch nur als Input der „Computer Sciences" bezeichnet werden. Technische Aspekte rücken bei

diesen Projekten zugunsten der vermittelten Inhalte meist in den Hintergrund. Dies hat zur Folge, dass sie jenseits des methodischen Diskurses stehen und somit an den Debatten der DH-Community nicht teilnehmen. Die Theologie als Disziplin wird somit weiterhin als rückgeschritten wahrgenommen, obgleich sie *realiter* technisch fundierte Projekte aufzuweisen hat.

Theologie als Teil der DH

Bisher traten im Bereich der DH vor allem Akteure der Religionswissenschaft wie Frederik Elwert von CERES hervor, die sich selbst nicht der theologischen Forschungslandschaft zuordnen, jedoch häufig seitens der DH-Community mangels Alternativen als Repräsentant*innen der Theologie aufgefasst werden (zur Abgrenzung von Religionswissenschaft und Theologie s. z. B. Moenikes, 1997; es gibt dagegen jedoch auch die Position, Religionswissenschaft als integralen Bestandteil der Theologie zu betrachten, s. z. B. Feldtkeller 2006, 121-139). Doch auch Theolog *innen treten in den letzten Jahren in den DH-Diskurs ein. So präsentiert Phillips die Arbeiten von CODEC zu digitaler Theologie auf der ADHO-Konferenz von 2017 (s. Phillips et al. 2019, 40). Anderson (2019) stellt wichtige Überlegungen zur Zukunft der Theologie in den DH an und im deutschsprachigen Raum können manche Beiträge im Band *Verkündigung und Forschung* 65/2 (z. B. von Zahnd (2020, 114-123) zur mittleren und neuen Kirchengeschichte oder von Karcher (2020, 132-142) zur Praktischen Theologie) wichtige Impulse setzen. Ein ähnlicher Beitrag zur Patristik liegt dank Volp (2020) vor. Hutchings und Clivaz (2021) sowie Clark und McBride Lindsey (2022) bieten indes einen internationalen Querschnitt zur Thematik an. Im Rahmen des TheoLabs hoffen meine Kollegin Frederike van Oorschot und ich, diesen Aufbruch in der theologischen Forschungslandschaft zu intensivieren, indem wir die Herausgabe eines Kompendiums Computational Theology planen, das die Potentiale der DH-Praktiken für theologische Fragestellungen verstärkt in den Blick nehmen wird (s. Nunn et al. 2023).

Fazit

Der Überblick zum (fehlenden) Zusammenspiel von Theologie und DH hat verdeutlicht, dass hierfür unterschiedliche Gründe existieren. Neben den in allen Geisteswissenschaften tradierten Ressentiments gegen die DH hat sich gezeigt, dass digitale Theologie über den Einsatz von DH hinaus ein wesentlich größeres Begriffsspektrum umfasst und Theolog*innen in der Folge andere Dimensionen von Digitalität in den Fokus stellen. Zugleich wurde sichtbar, dass auch DH-typische Projekte in der theologischen Forschungslandschaft vorhanden sind, diese aber anders gelabelt werden. Die Problematik des Aufbaus von Forschungsstrukturen unter Ausschluss der DH-Community besteht in der Gefahr, aufgrund von Unkenntnis redundante Arbeitsschritte durchzuführen, indem Forschungsprozesse mühsam entwickelt werden, die andernorts bereits praktiziert werden. Im Sinne guter wissenschaftlicher Praxis wäre demnach eine Öffnung gegenüber der DH-Community geboten, was nicht heißt, dass die gleichen Begrifflichkeiten genutzt werden müssen, jedoch zumindest die Projekte im Rahmen von ADHO Konferenzen vorgestellt werden sollten. Seitens der Digital Humanists kann dieser Prozess unterstützt werden, indem verstärkt Ressourcen in die Wissenschaftskommunikation investiert werden. Die jüngsten Entwicklungen in der theologischen Forschungslandschaft zeigen erste Erfolge. Die Hoffnung scheint aktuell berechtigt, dass die Theologie wieder als Player der DH etabliert werden kann und umgekehrt auch der Umgang mit den DH langsam aber stetig in das Selbstverständnis theologischer Forschung integriert wird. Durch die vielfältigen Quellen, die im theologischen Bereich noch digital zu erschließen sind, stellt dies ein lohnenswertes Unterfangen für beide Seiten dar.

Bibliographie

Anderson, Clifford. 2019. „Digital Humanities and the Future of Theology .“ *Cursor_ Zeitschrift für Explorative Theologie* 1, Nr. 1: 75-103. https://doi.org/10.17885/heiup.czeth.2019.1.24000 (zugegriffen: 19. Juli 2023).

Berry, David M. und Anders Faggerjord. 2017. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge und Malden: Polity.

Blaney , Jonathan, Sarah Milligan , Marty Steer und Jane Winters. 2021. *Doing Digital History : A Beginner's Guide to Working with Text as Data.* Manchester: Manchester University Press.

Burghardt, Manuel. 2020. „Theorie und Digital Humanities: Eine Bestandsaufnahme.“ Digital Humanities Theorie. https://dhtheorien.hypotheses.org/680 (zugegriffen: 19. Juli 2023).

Clark, Emily Suzanne und Rachel McBride Lindsey , ed . 2022. *Digital Humanities and Material Religion : An Introduction* . Berlin: De Gruyter.

Da, Nan Z. 2019. „The Digital Humanities Debacle.“ *The Chronicle for Higher Education*, 27. März 2019. https://www.chronicle.com/article/the-digital-humanities-debacle (zugegriffen: 19. Juli 2023).

Feldtkeller, Andreas. 2006. „Religions- und Missionswissenschaft: Was den Unterschied ausmacht für das Gesamtprojekt Theologie,“ in *Eine Wissenschaft oder viele? Die Einheit evangelischer Theologie in der Sicht ihrer Disziplinen*, hg. von Ingolf U. Dalferth, 121-139. Leipzig: Evangelische Verlagsanstalt.

Gramelsberger , Gabriele. 2023. *Philosophie des Digitalen zur Einführung.* Hamburg: Junius Verlag GmbH.

Heil, Uta. 2022. „Digital Humanities - Zwischen Fortschritt und Rückschritt. Ein Standpunkt.“ *Journal of Ethics in Antiquity and Christianity* 4: 101-104.

https://doi.org/10.25784/jeac.v4i0.1027 (zugegriffen: 19. Juli 2023).

Hunze, Guido. 2021. „Technisches Upgrade oder soziokulturelle Transformation? Warum Digitalisierung mehr als der Einsatz digitaler Medien in der Lehre ist," in *Theologiestudium im digitalen Zeitalter*, hg. von Andree Burke, Ludger Hiepel, Volker Niggemeier und Barbara Zimmermann, 97-119. Stuttgart: Verlag W. Kohlhammer.

Hutchings, Tim und Claire Clivaz , ed . 2021. *Digital Humanities and Christianity : An Introduction* . Berlin: De Gruyter.

Jones, Steven. 2016. *Roberto Busa , S. J., and the Emergence of Humanities Computing: The Priest and the Punched Cards.* New York, London: Routledge.

Karcher, Stefan. 2020. „Praktische Theologie und Digital Humanities ." *Verkündigung und Forschung* 65, Nr. 2: 132-142.

Krämer, Sybille. 2018. „Der ‚Stachel des Digitalen' – ein Anreiz zur Selbstreflexion in den Geisteswissenschaften? Ein philosophischer Kommentar zu den Digital Humanities in neun Thesen." *Digital Classics Online* 4, Nr. 1: 5-11. https://doi.org/10.11588/dco.2018.0 (zugegriffen: 19. Juli 2023).

Lässig, Simone. 2021. „Digital History : Challenges and Opportunities for the Profession." *Geschichte und Gesellschaft* 47, Nr. 1: 5-34.

Moenikes, Ansgar. 1997. „Zum Verhältnis zwischen Religionswissenschaft und Theologie." *Zeitschrift für Religions- und Geistesgeschichte* 49, Nr. 3: 193-207. https://www.jstor.org/stable/23899600 (zugegriffen: 19. Juli 2023).

Nunn, Christopher, Frederike van Oorschot und Selina Fucker. 2023. „Revolution through collaboration ? An attempt to familiarize " old guards " with DH." *Digital Humanities 2023. Collaboration as Opportunity (DH2023).* Graz: 1-2. https://doi.org/10.5281/zenodo.8118900 (zugegriffen: 19. Juli 2023).

Nyhan , Julianne. 2022. *Hidden and Devalued Feminized Labour in the Digital Humanities : On the Index Thomisticus Project 1954-67*. London: Routledge.

Phillips, Peter, Kyle Schiefelbein-Guerrero und Jonas Kurlberg . 2019. „ Defining Digital Theology : Digital Humanities , Digital Religion and the Particular Work of the CODEC Research Centre and Network." *Open Theology* 5: 29-43. https://doi.org/10.1515/opth-2019-0003 (zugegriffen: 19. Juli 2023).

Sahle , Patrick. 2013. „DH studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities ." *DARIAH-DE Working Papers* 1: 1-37. http://resolver.sub.uni-goettingen.de/purl/?dariah-2013-1 (zugegriffen: 19. Juli 2023).

Stoellger, Philipp. 2021. „Was bedeutet Digitalisierung – für die Schrift als Schrift? in *Digitalisierung: Neue Technik – neue Ethik: Interdisziplinäre Auseinandersetzung mit den Folgen der digitalen Transformation*, hg. von Benjamin Held und Frederike van Oorschot, 105-141. Heidelberg: heiBOOKS.

Sutinen , Erkki und Anthony-Paul Cooper. 2021. *Digital Theology : A Computer Science Perspective* . Bingley : Emerald Publishing.

Terras , Melissa. 2011. „ Peering Inside the Big Tent : Digital Humanities and the Crisis of Inclusion. " *Adventures in Digital Cultural Heritage*. https://melissaterras.org/2011/07/26/peering-inside-the-big-tent-digital-humanities-and-the-crisis-of-inclusion (zugegriffen: 19. Juli 2023).

Thaller, Manfred. 2017. „Geschichte der Digital Humanities," in *Digital Humanities: Eine Einführung*, hg. von Fotis Jannids, Hubertus Kohle und Malte Rehbein, 3-12. Stuttgart: J.B. Metzler Verlag.

Ulshöfer, Gotlind und Peter G. Kirchschläger . 2021. „Digitalisierung aus theologischer und ethischer Perspektive: Eine Einführung," in: *Digitalisierung aus theologischer und ethischer Perspektive: Konzeptionen – Anfragen – Impulse*, hg. von Gotlind Ulshöfer, Peter G. Kirchschläger und Markus Huppenbauer, 9-22. Baden-Baden: Nomos Verlag.

van Oorschot, Frederike. 2020. „Digital theology." *Verkündigung und Forschung* 65, Nr. 2: 162-171.

van Oorschot, Frederike . 2021. „ Neue Technik – neue Wissenschaft? Wissenschaftstheoretische und -ethische Herausforderungen der Digital Humanities," in *Digitalisierung: Neue Technik – neue Ethik: Interdisziplinäre Auseinandersetzung mit den Folgen der digitalen Transformation*, hg. von Benjamin Held und Frederike van Oorschot, 143-164. Heidelberg: heiBOOKS.

Volp, Ulrich. 2020. „computatoribus utamur! Herausforderungen der Digital Humanities für die Alte Kirchengeschichte", in *Kirchengeschichte: Historisches Spezialgebiet und/oder theologische Disziplin. Wolfram Kinzig zum 60. Geburtstag*, hg. von Claudia Kampmann, Ulrich Volp, Martin Wallraff und Julia Winnebeck, 439-462. Leipzig: Evangelische Verlagsanstalt.

Zahnd, Ueli. 2020. „Netzwerke, historisch und digital: Digital Humanities und die Mittlere und Neue Kirchengeschichte." *Verkündigung und Forschung* 65, Nr. 2: 114-123.

Disambiguierung von Wortbedeutungen aus dem Thesaurus Linguae Latinae mittels Fine-tuning von Latin BERT

Lendvai, Piroska
piroska.lendvai@badw.de
BAdW, Deutschland

Wick, Claudia
claudia.wick@thesaurus.badw.de
BAdW, Deutschland

1 Einleitung

Im Bereich des historischen Natural Language Processing (NLP) existieren für eine wachsende Anzahl von Sprachen kontextuelle Sprachmodelle. Für das Latein wurde von Bamman und Burns (2020) ein vortrainiertes BERT-Modell (vgl. Devlin *et al.*, 2018) veröffentlicht, das sie in vier klassische NLP-Aufgaben jeweils einer fine-tuning unterzogen haben, u.a. für Word Sense Disambiguation (WSD). Im Bereich der computergestützten Semantikanalyse der WSD kamen bereits verschiedene Ansätze des maschinellen Lernens zur Anwendung (einen Überblick bietet z.B. Navigli, 2009), neuere Arbeiten setzen auch auf neuronale Modelle und -Architekturen, auch in Kombination mit lexikalischen Wissensdatenbanken und enzyklopädischen Ressourcen (vgl. Bevilacqua *et al.*, 2021). WSD wird in der Regel als überwachte Klassifikation durchgeführt, wobei der Algorithmus lernen soll, die passende Zuordnung zu einer Bedeutungsgruppe für ein oder mehrere Fokuswörter im jeweiligen Kontext, z.B. innerhalb eines Satzes, vorherzusagen. Je nach Anwendungsziel können solche Bedeutungskennzeichnungen ('sense labels') auf unterschiedliche Weise definiert werden: Das angestrebte grobe oder feine Unterscheidungsspektrum kann zwei oder auch mehrere Kategorien umfassen.

Für WSD existieren — mit Ausnahme des Englischen — kaum allgemein verfügbare, große Benchmark-Datensätze: Die manuelle Erstellung von annotierten Datengrundlagen für eine überwachte WSD-Anwendung ist oft mühselig und langsam. Die Suche nach bereits vorhandenen, anpassbaren Ressourcen und Methoden ist daher ein wichtiges Forschungsthema. Eine vielversprechende Ressource, die für lateinische WSD genutzt werden könnte, ist Latin WordNet; für Details sowie Referenzen siehe Franzini *et al.* (2019). Eine andere Möglichkeit bestünde in der Nachnutzung von Wörterbuchdaten, weil deren Artikel Textstellen klassieren, die manuell annotierte Daten darstellen, die für Training und Test genutzt werden könnten. Unsere Pilotstudie Lendvai und Wick (2022) hatte das Ziel, Einblicke in Methoden zu gewinnen, die solche Wörterbuchdaten als Basis für die automatische Zuweisung von Bedeutungen nutzen könnten. Wir haben gezeigt, wie wir 40 Lemmata aus einem hauseigenen Wörterbuch zum Zwecke der historischen WSD verarbeitet hatten. Nach dieser Methodik haben wir die Studie nun weitergeführt, und insgesamt 115 Lemmata bearbeitet. Die Details dazu sowie die Resultate werden wir in diesem Beitrag darstellen.

Bereits Bamman und Burns (2020; im Folgenden: B&B) stellten Trainingsdaten für BERT — eine Transformer-basierte Deep-Learning-Architektur für Sprachmodellierung — zusammen, welche sie dem lateinisch-englischen Wörterbuch von Lewis und Short (1879; fortan: L&S) entnahmen. Für jedes Stichwort im Fokus der WSD (ein sog. 'Lemma') wählten sie eine Anzahl antiker Belegstellen aus den ersten beiden semantischen Hauptgruppen des jeweiligen Artikels des Fokus-Lemmas. Der binären Klassifikation von B&B folgend haben wir für dieselbe Auswahl von Lemmata Daten aus einer derzeit noch proprietären Quelle erhalten: aus dem Thesaurus Linguae Latinae (TLL)[1].

Der TLL ist ein einsprachiges Wörterbuch, das sämtliche Bedeutungen aller lateinischen Wörter der Antike verzeichnet und eine repräsentative Auswahl von Originalbelegen für jede Bedeutung zitiert. Der gesamte lateinische Wortschatz umfasst schätzungsweise 53‘000 bis 56‘000 Wörter. Die Aussicht, die WSD-Leistung bei Datensätzen, die aus zwei verschiedenen Wörterbüchern stammen, zu vergleichen, wäre in mehrfacher Hinsicht interessant: etwa um quantitative Einblicke in die Strukturierungspraxis von Wörterbüchern zu gewinnen, oder sogar zu versuchen, deren Einordnungspraxis empirisch zu überprüfen. Nach Sichtung der Daten wurde uns allerdings klar, dass ein direkter Vergleich von Daten aus dem TLL bzw. aus L&S methodisch kaum vertretbar wäre.

Vorgängig durchgeführte Vergleiche zwischen Artikeln zum jeweils gleichen Lemma ergaben, dass die beiden Lexika bei der Kategorisierung unterschiedlich verfahren. Für dasselbe Lemma waren beispielsweise manche Zitate, welche in L&S der Hauptbedeutung I zugewiesen werden, im TLL auf die ganz anders definierten Hauptgruppen I und II verteilt (oder umgekehrt). Dieser Unterschied rührt daher, dass Lexikographen die Bedeutungen meist nicht als flache, durchnummerierte Liste präsentieren, sondern sie nach übergeordneten Kategorien in eine systematische Ordnung bringen. Eine Norm hierfür existiert allerdings nicht. Eine solche Gliederung, oft mit zwei Hauptgruppen, ist ein künstliches Darstellungskonstrukt, welches für ein und dasselbe Wort ganz unterschiedliche Formen annehmen kann (vgl. Punkt 2.1). Leider führt dies dazu, dass kein direkter Vergleich der WSD auf so unterschiedlichen Datensätzen möglich ist.

Wir haben uns gegen B&Bs Beschränkung der Datenmenge entschieden, welche für beide Hauptbedeutungen nur jeweils gleichgroße Zitatmengen zulassen, da wir sonst das volle Potenzial des TLL — seinen Umfang — nicht ausschöpfen könnten. Wir kürzten deshalb die oft um ein Mehrfaches größere Bedeutungsgruppe nicht, sondern nahmen deren quantitative Unausgeglichenheit in Kauf. Unser TLL-Datensatz ist daher deutlich umfangreicher und in Bezug auf das Labeling möglicherweise grobkörniger geworden als die B&B-Daten. Diese enthalten allerdings Lemmata, welche im TLL noch fehlen, d.h. Wörter nach „Re-“.

Unsere Arbeit verfolgt folgende Ziele:

- Suche nach Methoden und Testmöglichkeiten für die experimentelle Einordnung oder sogar Validierung von Bedeutungsrepräsentationen und deren Unterscheidung.

- Aufzeigen von Zusammenarbeit zwischen den Geisteswissenschaften und der NLP-Gemeinschaft, die komplementäres Fachwissen liefern.
- Weiterverwendung des von B&B vortrainierten neuronalen Sprachmodells für Latein.
- Wiederholung des WSD-Experiments von B&B anhand der von ihnen veröffentlichten Benchmark-Daten, des Codes und des Vergleichsmodells.
- Wiederholung des WSD-Experiments durch fine-tuning von Latin BERT mit Daten, die wir aus dem TLL konstruieren.
- Beobachtung der Einordnungspraxis und des Umfangs in den B&B und TLL Datensätzen.
- Verbesserung der experimentellen Methodik von B&B durch eine detaillierte Auswertung in Form von F-Makro in einem Pro-Lemma-Aufbau der Experimente.

2 Explorative Datenanalyse

Das Wörterbuch von L&S ist eine Übersetzung von Wilhelm Freunds deutsch-lateinischem Wörterbuch aus dem 19. Jahrhundert, welches auf älteren Lexika aufbaut.[2] B&B wählten daraus 201 Lemmata und entnahmen ihnen 8‘354 Instanzen für ihren Datensatz. Der TLL wird seit 1900 erarbeitet, wobei sich die redaktionellen Prinzipien über die Zeit verändert haben.[3] Typisch ist die verschachtelte Struktur der Artikel, welche über zehn Ebenen tief reichen kann. Bedeutungsgruppen auf derselben Ebene sollten einander ausschließende Parameter syntaktischer oder semantischer Natur aufweisen, was oft zu einer dichotomischen Anordnung führt. Diese Systematisierung von Bedeutungsunterschieden orientiert sich also nicht an den verschiedenen Übersetzungsmöglichkeiten des Wortes, sondern klassifiziert semantische Merkmale innerhalb der Ausgangssprache Latein. Zur Illustration der Bedeutungen werden passende Textzitate aufgelistet (auch „Belege“ oder „Stellen“ genannt).

Die TLL-Daten standen uns im TEI-XML-Format zur Verfügung. Genau wie B&B haben wir die Trainingsdaten aus jeweils einem einzigen Wörterbucheintrag generiert (Homonyme bleiben ausgeschlossen) und nur Belege mit einer Mindestlänge von fünf Wörtern berücksichtigt. Dieses Datenmaterial wurde ausschließlich jenen zwei Hauptbedeutungsgruppen zugewiesen, welche im Artikel als Teilung auf der höchsten (ersten) Ebene definiert werden (üblicherweise als „I“ und „II“ bezeichnet). Durch rekursives Hinabsteigen in die verschachtelte Struktur des gedruckten Artikels (s. Abbildung 1) wurde diese abgeflacht (s. Abbildung 2).

Der so gewonnene TLL-Datenbestand umfasst 65‘522 Textstellen für 115 Lemmata (die einer Stichprobe aus dem 120 von B&B entsprechen), die sich wie folgt über die Wortarten verteilen: 49% Verben, 14,5% Adjektive, 15,5% Substantive, 21% andere (Adverbien, Pronomina, Präpositionen, Konjunktionen, Partikeln). Ab dem Buchstaben C wird das Lemma in den TLL-Artikeln zunehmend abgekürzter, d.h. es erscheint lediglich die Endung im Zitat. Die Ergänzung der oft sehr unterschiedlichen Wortstämme war für die Durchführung aussagekräftiger WSD-Experimente unabdingbar, da die Lemmaformen eine zentrale Information für die Lernalgorithmen darstellen. Dies erforderte ein Human-in-the-Loop-Verfahren, bei dem für jedes Lemma Ergänzungsanweisungen in Form von Python-Code geschrieben wurden.

βres definitae:
① *specimina pauca ad illustrandas notiones selecta*:
Ⓐ *tolerandi, sustinendi* (*sc. fortiter sim.*):
① *in universum*: PLAVT. Men. 721 viduam esse mavelim, quam istaec flagitia tua -i (779 perpeti). 978 magis multo -or facilius verba: verbera ego odi. TER. Eun. 244 neque ridiculus esse neque plagas -i possum. PACVV. trag. 279 -or facile iniuriam, si est vacua a contumelia. CIC. Verr. II 3, 95 quem *contumeliae aculeum* -i … viri boni difficillime possunt. 3, 201 si hoc vectigal aratio tolerare, hoc est Sicilia ferre ac -i potest. Phil. 6, 19 aliae nationes servitutem -i possunt, populi Romani est propria libertas. fin. 3, 42 si dolores eosdem tolerabilius -untur qui excipiunt eos pro patria quam qui leviore de causa (*item in philosophia:* 4, 23 *Panaetius* cum … de dolore -endo scriberet. Tusc. 4, 60 qui non turbulente humana -antur. *sim. al.*). BRVT. Cic. ad Brut. 24, 6 W. servire et -i contumelias … odero. VARRO rust. 2, 10, 3 *senes* callium difficultatem ac montium arduitatem … *non* facile ferunt, quod -undum *est pastoribus. et passim.*
② *-untur qui quid incolumes, sine noxa sim. sustinent* (*exempla potiora; cf. p. 725, 31*): OV. trist. 3, 3, 7 nec caelum -or nec aquis adsuevimus istis. CELS. 2, 18, 3 pisces …, qui salem non -untur. SEN. epist. 51, 10 quamlibet viam iumenta -untur, quorum durata … ungula est. COLVM. 8, 17, 8 *mullus* raro … *vivarii* claustra -itur. PLIN. nat. 31, 23 *fluvii cuiusdam* gurgitem periuri negantur -i velut flammam.
③ *-untur qui pondera corporea sustinent* (*proprie et in imagine; cf. e. g. p. 724, 7*): SEN. contr. 3 praef. 9 quidam *equi* melius equitem -antur, quidam iugum (*addas imagines vol. VII 2, 641, 70 sq.; aliter p. 722, 4*). suas. 2, 1 (*ironice*) insueta … arma non -surae manus (STAT. Theb. 11, 551 [*Pofynices ad fratrem*] exercita … membra vides *mea*; disce a. -i). SEN. Thy. 931 (*in imag.*) pondera regni non inflexa cervice -i (SIL. 14, 90).
Ⓑ *subeundi, experiendi* (*sc. mala, quibus quis afficitur neglecto respectu fortiter, laboriose sim. perpetiendi; bona v. sub B3*): CIC. rep. 3, 23 cum de tribus unum est optandum, aut facere iniuriam nec accipere, aut et facere et accipere, aut neutrum, optumum est facere impune …, secundum nec facere nec -i, miserrimum digladiari semper tum faciendis tum accipiendis iniuriis. NIGID. Gell. 9, 12, 6 imminentia fraudis,

Abbildung 1: Ausschnitt aus der verschachtelten Struktur des TLL-Artikels für das Lemma *patior*.

```
77  patior  I   viduam esse mavelim , quam istaec flagitia tua   pati
78  patior  I   magis multo   patior  facilius verba : verbera ego odi
79  patior  I   neque ridiculus esse neque plagas pati   possum
80  patior  I    patior  facile iniuriam , si est vacua a contumelia
81  patior  I   quem contumeliae aculeum  pati   … viri boni difficillime possunt
82  patior  I   si hoc vectigal aratio tolerare , hoc est sicilia ferre ac   pati   potest
83  patior  I   aliae nationes servitutem pati   possunt , populi romani est propria libertas
84  patior  I   si dolores eosdem tolerabilius   patiuntur  qui excipiunt eos pro patria quam qui leviore
    de causa
85  patior  I   panaetius cum … de dolore patiendo   scriberet
86  patior  I   qui non turbulente humana patiantur
87  patior  I   servire et   pati   contumelias … odero
88  patior  I   senes callium difficultatem ac montium arduitatem … non facile ferunt , quod  patiundum
    est pastoribus . et passim
89  patior  I   nec caelum   patior  nec aquis adsuevimus istis
90  patior  I   pisces … , qui salem non  patiuntur
91  patior  I   quamlibet viam iumenta   patiuntur  , quorum durata … ungula est
92  patior  I   mullus raro … vivarii claustra   patitur
93  patior  I   fluvii cuiusdam gurgitem periuri negantur pati   velut flammam
94  patior  I   quidam equi melius equitem   patiantur  , quidam iugum
95  patior  I   insueta … arma non   passurae   manus
96  patior  I   exercita … membra vides mea ; disce a .  pati
97  patior  I   pondera regni non inflexa cervice pati
98  patior  I   cum de tribus unum est optandum , aut facere iniuriam nec accipere , aut et facere et
    accipere , aut neutrum , optumum est facere impune … , secundum nec facere nec  pati   , miserrimum
    digladiari semper tum faciendis tum accipiendis iniuriis
99  patior  I   imminentia fraudis , quam quis vel facturus cuipiam vel   passurus   est
```

Abbildung 2: Geglättete, mit sense-label versehene Trainingsdaten aus dem TLL-Artikel für das Lemma *patior*.

2.1 Beobachtungen zu Bedeutungseinordnungen in den beiden Wörterbüchern

Für das Vortraining von Latin BERT wurden Texte aus einem noch größeren Zeitraum (ca. 2‘000 Jahre) benutzt, der neben der Antike auch das Mittelalter, den Humanismus und die Neuzeit umfasst. Der TLL dokumentiert das antike Latein umfassender als L&S: er berücksichtigt auch nicht-

literarische Textgattungen wie Inschriften, juristische oder medizinische Quellen, sowie die spätantiken christlichen Texte (bis ca. 700 n. Chr.). Das dargestellte Bedeutungsspektrum ist größer als bei L&S, welche vor allem die klassischen Autoren aus ca. 200 Jahren berücksichtigen. Das bedeutet z.B. für *religio* In der klassischen, „heidnischen" Zeit bezeichnete das Wort Gefühle der (Ehr)furcht gegenüber Göttern oder strikt fixierte Kulthandlungen. Im heute bekannten Sinn von 'Religion' als dogmatisches Glaubenssystem, das auf Offenbarung beruht, wurde es erst bei christlichen Autoren verwendet.

Sowohl in den B&B- als auch in den TLL-Daten wird nicht immer wörtlich zitiert, sondern die Belege werden häufig leicht verändert. L&S schreiben sie der Lesbarkeit oder Kürze wegen zuweilen um, was zur Paraphrasierung des Originalzitats führen kann; der TLL sucht dies zu vermeiden.

Die häufigsten Schemata bei der Bedeutungseinordnung sind die folgenden:

- Unterscheidung nach natürlichen, semantischen Kriterien. Typisch hierfür ist die Unterteilung „I: konkret (körperlich, räumlich etc.)" vs. „II: übertragen (bildlich, metaphorisch etc.)".
- Unterscheidung nach künstlichen Kriterien wie z.B. „I: Allgemein" vs. „II: im Besonderen", wobei die unter II aufgeführten Spezialitäten weder syntaktisch noch semantisch homogen sind.
- Hierarchische Unterscheidung wie Lemma vs. Sublemma. L&S tendieren dazu, beispielsweise ein adjektivisch verwendetes Partizip Perfekt als Klasse II zu definieren, wogegen der TLL z.B. *remissus* nicht zusammen mit *remitto* präsentiert, sondern es in einem gesonderten Artikel behandelt. Die entsprechenden Zitate solcher Sublemmata erscheinen deswegen nicht in den TLL-Trainingsdaten.

3 WSD-Experimente

Beim fine-tuning von Latin BERT wird das vortrainierte Sprachmodell explizit für die WSD-Aufgabe trainiert, d.h. für die Klassifizierung der Zitate, die den Klassen I oder II im Thesaurus entspricht. Das Klassifizierungsmodell lernt dabei, genau zwei Bedeutungen für ein Fokus-Lemma (jeweils ein Wort pro Zitat) zu unterscheiden. Diese Aufgabe steht im Gegensatz zu dem, was in der ersten Phase des Vortrainings von Latin BERT, stattfand. Dort bestand die Aufgabe darin, dass das lateinische Sprachmodell von BERT so viele Bedeutungen eines Wortes wie möglich erlernt.

3.1 Aufbau der Experimente

Wir haben 100 epochs (Trainingszyklen) pro Lemma durchgeführt, d.h. Training und Testing wurden für jedes Lemma einzeln durchgeführt. Die Leistung wurde anhand des ungewichteten Makro-F1-Scores jedes Lemmas unter Verwendung von Pedregosa *et al*. (2011) bewertet. Auf die Genauigkeit (accuracy) zu evaluieren wäre suboptimal, da sie nicht klar ausdrückt, wie gut die beiden Klassen im Einzelnen erlernt werden, und sie korrigiert auch nicht die quantitative Unausgewogenheit beider Klassen. Für jede epoch wurde das Makro F1 auf den Entwicklungsmengen-Datensatz (development set) des Lemmas berechnet. Die Parameter, die auf dem development set mit der besten Leistung erlernt waren, wurden verwendet, um das Makro F1 für jedes Lemma auf der im Training nicht verwendeten, d.h. von BERT noch ungesehenen Testmenge (heldout testset) zu messen.

B&B verwendeten für das Training den ungewichteten cross-entropy loss. Da in unseren Daten die beiden Klassen jedes Lemmas unausgewogen sind, haben wir die Gewichte für jede Klasse für die loss function berechnet. Als Vergleichsmodell (baseline classifier) verwendeten wir dieselben 200-dimensionalen statischen word2vec-embeddings in einem biLSTM-Klassifikator wie B&B (vgl. Mikolov *et al.*, 2013). Enklitika wurden in den Zitaten nicht abgetrennt, da davon ausgegangen wurde, dass der wordpiece tokenizer von Latin BERT diese berücksichtigt.

3.2 Auswertung der experimentellen Resultate

Wir reproduzierten die B&B-WSD-Studie und bekamen einen vergleichbaren Genauigkeitswert (.737; bei B&B: .754). Anschließend haben wir aus dem B&B-Datensatz mit 201 Lemmata Datensätze pro Lemma für diejenigen 115 Lemmata abgeleitet, wofür wir aus dem TLL Daten hatten. Auf diesen haben wir Latin BERT sowie biLSTM als Vergleichsmodell trainiert, wobei wir den für die Auswertung mit Makro-F1-Score pro Lemma ergänzten B&B-Code verwendeten. Die Ergebnisse sind in Tabelle 1 dargestellt. Wir haben festgestellt, dass die B&B-Daten im Vergleich mit den TLL-Daten gering sind (vgl. Abbildung 3) und daher statistisch nicht zuverlässige Ergebnisse liefern (d.h. die Standardabweichungen sind groß); diese Variabilität wird auch durch die whisker des boxplots (Abbildung 4) veranschaulicht.

Tabelle 1 zeigt die Score-Übersicht: Für beide Datensätze übertrifft Latin BERT das biLSTM-Vergleichsmodell. Abbildung 4 zeigt auch, dass der Median der Punktzahlen für TLL-Daten höher ist als für B&B-Daten.

Tabelle 1: Mittlere Leistungswerte über 115 Lemmata auf B&B sowie auf TLL-Daten. Wir weisen darauf hin, dass ein direkter Vergleich aufgrund des unterschiedlichen Labelings der Datensätze nicht aussagekräftig ist.

Datensatz	Modell	Mittleres F-Makro	Standardabweichung
B&B	BERT biLSTM	.635 .595	.25 .25
TLL	BERT biLSTM	.792 .701	.11 .12

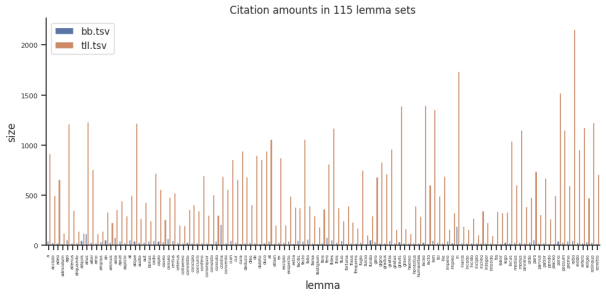


Abbildung 3: Datensatzgröße pro Lemma. Blau: L&S-Daten, Orange: TLL-Daten.

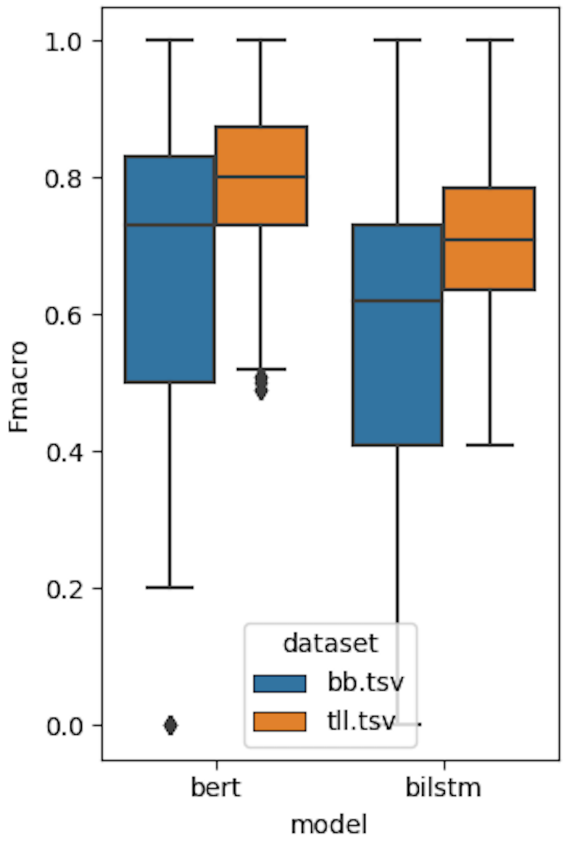


Abbildung 4: Leistung: mittleres F-Makro der Lemmata pro Datensatz mittels BERT bzw. mittels biLSTM-Baseline-Modell. Hinweis: Ein Vergleich der Leistung auf den Datensätzen ist aufgrund der unterschiedlichen Datenkonstruktion nicht aussagekräftig.

4 Schlussfolgerung

Unsere Studie bestätigt die Bedeutung von Latin BERT als Ressource für die WSD-Klassifizierungsaufgabe. Wir haben experimentell nachgewiesen, dass sich aus den TLL-Wortbedeutungen abgeleitete Trainingsdaten für diese Aufgabe gewinnen lassen. Die fine-tuning des kontextuellen Sprachmodells in einer Transformer-Architektur konnte als Methode zur experimentellen Validierung der Bedeutungsabbildung verwendet werden. Unsere WSD-Modelle, die auf TLL-Daten trainiert wurden, erbrachten eine gute Leistung und eine Verbesserung gegenüber dem Vergleichsmodell mit biLSTM-Architektur und statischen word embeddings. Die Resultate auf unserem Datensatz aus 115 Lemmata zeigen einen ähnlichen Trend und eine vergleichbare Leistung wie die Resultate auf dem kleineren Datensatz aus 40 Lemmata in Lendvai und Wick (2022).

Für die Studie haben wir einen Datensatz erstellt, der zwei grobkörnige Hauptbedeutungen pro Lemma enthält. Er beruht auf Daten, die uns freundlicherweise vom Verlag de Gruyter zur Verfügung gestellt wurden, wobei man noch klären sollte, ob er als Benchmark-Datensatz für Latin WSD veröffentlicht werden kann. Die grobkörnige Unterteilung in lediglich zwei Gruppen wird dem semantischen Spektrum der meisten Wörtern noch nicht gerecht. Dasselbe gilt allerdings auch für das Material von B&B, das ebenfalls Bedeutungen aus Untergruppen nicht unterscheidet. Innerhalb der Fachgemeinschaft sind uns keine weiteren gelabelten Datensätze aus Lateinwörterbüchern bekannt. Wir haben zwar versucht, aus einem verfügbaren digitalen Wörterbuch des Mittellateins gleichartige Daten (d.h. Zitate) zu extrahieren, doch deren Umfang war derzeit zu gering um als Trainingsmaterial verwendbar zu sein. Die beobachtete Nicht-Vergleichbarkeit der unterschiedlichen Lemma-Bearbeitungen in den verschiedenen Wörterbüchern hat gezeigt, dass weiterhin nach Wegen zu suchen ist, wie sich Datensätze vereinheitlicht repräsentieren und beschreiben lassen. Unsere Studie kann dazu wichtige Einsichten liefern.

Wir hoffen diese Studie erweitern zu können, um Erkenntnisse über die Organisation und maschinenlernbare Systematisierung von Bedeutungsinventaren zu gewinnen. In unserem Beitrag haben wir uns auf die Replikation des Experiments von Bamman & Burns sowie auf die sehr aufwendige Datenaufbereitung aus dem TLL konzentriert. Ein klarer Mehrwert unserer bisherigen Arbeit ist, dass wir qualitative sowie abstrakte Vergleiche zwischen den Quellen gemacht und auf unserem Datensatz gute Resultate erreicht haben. Ein künftiger Schritt kann darin bestehen, weitere Unterteilungen der Hauptgruppen für Klassifizierungen zu nutzen, vergleichbar mit dem hierarchischen Aufbau von Wörterbuchartikeln, sowie die Daten für die weiteren Hauptgruppen eines Lemmas zu nutzen, und den experimentellen Aufbau auszudehnen.

Bibliographie

Navigli, Roberto. 2009. „Word Sense Disambiguation: A Survey“. ACM Computing Surveys , 41(2)

Bamman, David, Patrick J. Burns. 2020. „Latin BERT: A Contextual Language Model for Classical Philology“. CoRR , abs/2009.10053. https://arxiv.org/abs/2009.10053

Bevilacqua, Michele, Tommaso Pasini, Alessandro Raganato, Roberto Navigli. 2021. „Recent trends in word sense disambiguation: A survey.“ In *Proceedings of*

the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21 .

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. „BERT: Pre-training of deep bidirectional transformers for language understanding“. arXiv preprint arXiv :1810.04805. https://arxiv.org/abs/1810.04805

Franzini, Greta, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura, Federica Zampedri. 2019. „Nunc est aestimandum: Towards an evaluation of the Latin WordNet“. In CLiC-it .

Lendvai, Piroska, Claudia Wick. 2022. „Finetuning Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae. In: *Proc. of the Workshop Cognitive Aspects of the Lexicon* , *Association for Computational Linguistics*, pp. 37-41.

Mikolov, Tomas, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. „Efficient estimation of word representations in vector space“. arXiv preprint arXiv :1301.3781.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. 2011. „Scikit-learn: Machine learning in Python“. Journal of Machine Learning Research 12: 2825-2830.

µEdition – Niedrigschwellige Digitale Editionen

Hall, Mark

mark.hall@work.room3b.eu
The Open University, Vereinigtes Königreich
ORCID: 0000-0003-0081-4277

Dass die Digital Humanities (DH) ein Kanonproblem haben, ist etwas, das schon länger diskutiert wird (vgl. Assman, 2010; Estill, 2019; Dziudzia und Hall, 2020; Estill et al., 2022). Dies obwohl das digitale Medium grundsätzlich eine Reduktion der Schwellen für Werke und Personen außerhalb des Kanons verspricht, da die Inhalte über das Internet zugänglich gemacht werden können, ohne dass zuerst die traditionellen Schleusenwärter (Verlage, Geldgeber, Gutachter:innen, Institutionen, …) vom Wert der Arbeit überzeugt werden müssen.

Die kritischen Edition, in seiner DH Umsetzung der digitalen Edition (Pierazzo, 2011; Sahle, 2016), sollte besonders von dieser Schwellenreduktion profitieren. Eine kritische Edition in Printform muss zwangsläufig im Vorhinein zeigen, dass ein hinreichendes Interesse an den Werken oder Personen der Edition besteht, um die Geldgeber oder Verlage davon zu überzeugen, die Edition zu finanzieren und publizieren. Die digitale Edition hat, theoretisch, wesentlich kleinere finanzielle und praktische Hürden, da das technische Minimum für die Veröffentlichung eine Webseite mit dem Inhalt der Edition ist. Trotzdem dominiert der Kanon die digitale Editionslandschaft (Estill, 2019). Dies hat verschiedene Gründe, der Fokus in dieser Arbeit liegt auf den Schwellen die die Editionssoftware aufwirft (vgl. Roselli del Turco, 2016).

Pierazzo (2019) definiert ein Spektrum an digitalen Editionen, von „Haute Couture“ Editionen, welche projektspezifisch und technologisch versiert sind, zu „Prêt-à-Porter“ Editionen, die komplett auf Standardsoftware basieren. In der Praxis sind der Großteil der sichtbaren, digitalen Editionen eher am „Haute Couture“ Ende des Spektrums zu finden (vgl. Dumont et al., 2023; Shakespeare et al., 2022). Der technische Unterbau für diese Editionen ist generell speziell für das Projekt entwickelt worden, was signifikante Kosten mit sich bringt. Diese Kosten werden generell über Geldgeber abgedeckt, wofür aber die Edition technische Innovation versprechen muss (Causer et al., 2012). Das Versprechen technischer Innovation führt generell dazu, dass die daraus resultierende technische Umsetzung sehr stark an die Spezifika des Werkes oder des Autors oder der Autorin angepasst werden, wodurch die Nachnutzbarkeit für andere Projekte erschwert wird.

Ein Fokus auf das Konzept der „Prêt-à-Porter“ Editionen bzw. Editionssoftware und den zugrundeliegenden Technologien wäre sinnvoll um die die Schwellen für die Erstellung einer Edition zu reduzieren. Der erste Schritt dazu ist grundlegend vorhanden, da sich TEI-XML als das primäre Datenformat für digitale Editionen durchgesetzt hat. Der große Vorteil von TEI-XML ist, dass das Format sehr flexibel ist (vgl. Burrows et al., 2021; Giovanetti and Tomasi, 2022) und dadurch fast das gesamte Spektrum an Annotationen, die für das Spektrum an digitalen Editionen notwendig ist, anbieten. Diese Flexibilität führt natürlich dazu, dass es sehr einfach ist, inkompatible TEI-XML Kodierungen zu erzeugen (Kudella and Jeffries, 2019). Das ist jedoch nur ein Problem wenn TEI-XML Daten verschiedener Projekte zusammengeführt werden sollen. Das zentralere Problem ist, dass das direkte Bearbeiten von TEI-XML aufwendig ist und dass das TEI-XML Format für die Präsentation im Internet in HTML umgewandelt werden muss, in der Terminologie von Roselli Del Turco (2016) benötigen die Nutzerin oder der Nutzer daher Werkzeugunterstützung in der *Production* und *Presentation* Phase.

Es existieren eine Reihe an Werkzeugen, die die „Prêt-à-Porter“ Idee umsetzen und Werkzeugunterstützung in diesen zwei Phasen bieten, unter anderem Ediarum (Dumont and Fechner, 2014), Hyper (Galka und Klug 2023), EVT[1], und TEI-Publisher[2]. Diese Werkzeuge sind über den Lauf vieler Jahre entwickelt worden und bieten nutzerfreundliche Umgebungen für die Erstellung und Präsentation digitaler Editionen. Das Problem ist, dass sie zwar wiederverwendbar sind, aber die Kosten für die Nutzung dieser Technologien signifikant sind. Ediarum erlaubt das einfache Edieren mit TEI-XML, baut aber auf der kommerziellen OxygenXML Software auf, deren Lizenzkosten effektiv nur mit institutioneller Unterstützung finanzierbar sind.

Die anderen Softwarekomponenten in Ediarum sind zwar Open-Source, benötigen aber Resourcen um die Edition online verfügbar zu machen. Ähnlich ist die Situation mit Hyper, EVT, und TEI-Publisher die vollständig auf Open-Source Software aufbauen und daher in der Erstellung minimale Kosten haben, aber signifikante Resourcen benötigen um die Edition verfügbar zu machen und diese Kosten bestehen für die gesamte Lebenszeit der Edition.

Diese Kostenanforderungen führen dazu, dass digitale Editionen für Doktoranden und Doktorandinnen, Einzelforscher und Einzelforscherinnen, und Miniprojekte als Projektform schwer zugänglich sind (Robinson, 2016), weil die Finanzierung für die Erstellung oder für die Verfügbarmachung fehlt.

Um diese Barrieren zu reduzieren wird oft vorgeschlagen, dass eine größere Standardisierung oder Formalisierung (Barabucci und Fisher, 2017) und bessere, generische Datenmodelle (Bürgermeister et al., 2023) dazu führen könnten, dass generische Platformen angeboten werden könnten (Fritze, 2019). Diese Ansätze nehmen aber alle an, dass großangelegte, DH-spezifische Infrastruktur notwendig ist. Van Zundert (2012) argumentiert jedoch, dass großangelegte Infrastruktur ineffizient ist, da sie generell schwerfällig und unflexibel ist. Der Aufwand um sich in die großangelegte Infrastruktur einzuarbeiten übersteigt dann auch oft was für Editionen mit kleinen Teams möglich ist.

Das µEdition Projekt

Eine Alternative für diesen Standardisierungsansatz ist es, die Anforderungen an eine digitale Edition aufzuweichen. Was wenn wir die Reduktion der Schwellen in den Mittelpunkt stellen und dafür akzeptieren, dass die Edition möglicherweise im Funktionalitätsumfang etwas eingeschrängt ist, das Interface stark standardisiert ist, TEI als Format optional ist und wir Langzeitarchivierung als „best-effort“ Aktivität sehen?

Das µEdition (Micro Edition – https://uedition.github.io) Projekt verfolgt genau diesen Ansatz, mit dem Ziel digitale Editionen für Projekte zu ermöglichen, die keinen Zugriff auf institutionelle oder finanzielle Ressourcen haben, die von einer einzelnen Person (oder einem Mini-Team) als Nebenprojekt erstellt werden, oder die Teil der forschungsgetriebenen Lehre sind. Das Projekt entstand aus der Arbeit an und Erfahrung mit zwei relativ unterschiedlichen Editionsprojekten dieser Art.

Das Editionsprojekt Karl Gutzkow[3] ist ein kritisches Editionsprojekt, das seit 1997 ohne externe Finanzierung an der Erstellung einer kritischen Edition zu Karl Gutzkow arbeitet (Lauster, 2020). Ursprünglich nutzte das Projekt eine manuell erzeugte Webseite für ihre digitale Edition. Seit 2018 wurde die digitale Edition auf eine moderne technische Basis gestellt und die Erfahrung aus dieser Arbeit hat die Entwicklung der µEdition maßgeblich beeinflusst. Insbesondere nutzt das Projekt TEI für die detaillierte Kodierung der Werke, aber einfachere Textformate für die restlichen Seiten der digitalen Edition.

Das zweite Projekt ist das Unter der Oberfläche[4] Projekt, welches das Ziel hat Autoren und Autorinnen sichtbarer zu machen, die aus verschiedenen Gründen von der Literaturgeschichte und dem Kanon vergessen wurde. In diesem Editionsprojekt geht es weniger um die Annotation der Werke, als darum die Autoren und Autorinnen zu präsentieren und auf ihre Werke zu verweisen. Daher setzt das Projekt primär auf Markdown als einfaches Textformat.

Projektziele

Innerhalb des Gesamtzieles die Schwellen für digitale Editionen zu reduzieren und basierend auf den Erfahrungen aus den zwei Projekten hat die µEdition die folgenden konkrete Ziele für die Umsetzung:

- Schnelle, einfache Erstellung und Veröffentlichung einer Edition.
- Mehrsprachigkeit als grundlegende Funktionalität.
- Statischer HTML Output.
- Nutzung existierenden Open-Source Werkzeuge.
- Markdown und TEI-XML als primäre Datenformate.
- Ein Editor für die einfache Bearbeitung der Edition.

Die technische Umsetzung der µEdition erfolgte in Python. Der Grund dafür ist, dass die µEdition Software möglichst einfach zu nutzen sein soll und Python im DH Bereich stark genutzt wird. Dies garantiert natürlich nicht, dass Python überall installiert ist, aber es reduziert für viele Editionsprojekte die Einstiegsschwelle. Um den Einstieg noch weiter zu erleichtern kommt die µEdition mit einer Projektschablone, die es ermöglicht innerhalb weniger Minuten eine erste, minimale Editionswebseite zu erzeugen.

Der Output der µEdition ist, wie in den Zielen aufgeführt, eine statische Webseite[5]. Der primäre Vorteil einer statischen Webseite ist, dass sie die Anforderungen an das Hosting, also der Verfügbarmachung im Internet, minimiert. Viele akademische Institutionen bieten ihren Mitarberiterinnen, Mitarbeitern und Studierenden gratis Webhosting an, welches aber oft auf statische Webseiten beschränkt ist. Anbieter wie GitHub[6] oder Read the Docs[7] bieten institutsunabhängig Gratishostinglösungen an und es gibt eine Fülle an anderen Anbietern. Eine statische Webseite beschränkt natürlich auch die Funktionalität, die dem Leser oder der Leserin angeboten werden kann. Für das Einstiegslevel, auf das die µEdition abzielt, ist dies jedoch eine akzeptabler Kompromiss.

Die µEdition baut auf existierender Software zur Erstellung von statischen Webseiten auf. Insbesondere wird das JupyterBook Werkzeug[8] zur genutzt. JupyterBook wurde aus zwei Gründen ausgewählt. Erstens basiert es auf einer Software, die seit 2008 dafür genutzt wird, statische Webseiten zu generieren, was eine langfristige Stabilität der Software garantiert. Zweitens werden Jupyter Notebooks im DH Bereich oft genutzt und das JupyterBook erlaubt es auch Teile der Edition direkt aus dem Jupyter Notebook zu generieren. Das JupyterBook kann auch leicht mit weiteren Annotationsmöglichkeiten erweitert werden, um editionsspezifische Annotationen hinzuzufügen.

Die µEdition erweitert das JupyterBook um zwei grundlegende Funktionen. Das erste ist Mehrsprachigkeit, da ohne Mehrsprachigkeit es schwierig ist, das Kanonproblem zu überwinden (Fiormonte, 2021; Spence and Brandao, 2021). Mehrsprachigkeit ist im JupyterBook relativ komplex umzusetzen. Um diese Komplexität zu umgehen nutzt die µEdition für jede Sprache ein eigenes JupyterBook. Die µEdition synchronisiert automatisch die Konfiguration und Inhaltsverzeichnisse über die Sprachen hinweg. Mit einem einzigen Befehl werden dann die statischen HTML Seiten für alle Sprachen generiert und in eine gemeinsame Struktur gebracht. Zusätzlich fügt die µEdition den HTML Seiten Interfaceelemente hinzu, damit der Nutzer oder die Nutzerin automatisch ihre bevorzugte Sprache sehen bzw. einfach zwischen den verschiedenen Sprachen wechseln können (Figure 1).

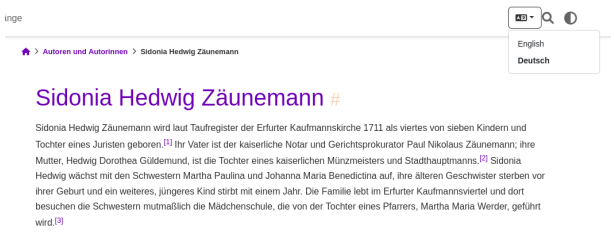


Abbildung 1 Screenshot aus der "Unter der Oberfläche" Edition. Rechts-oben sieht man die Funktionalität zum Wechseln zwischen den Sprachen.

Die zweite Funktion, die die µEdition dem JupyterBook hinzufügt, ist die Möglichkeit TEI-XML als Input- und als Output-format zu nutzen. Für beide Richtungen stellt die µEdition eine Reihe an Standardmappings bereit, um TEI-XML in HTML umzuwandeln, bzw. um Inhalte, die in Markdown erstellt wurden, als TEI-XML bereitzustellen. Diese Mappings können über die Konfiguration erweitert und überschrieben werden und erlauben es dem Editionsprojekt projekspezifische TEI-XML Strukturen zu nutzen (Figure 2).

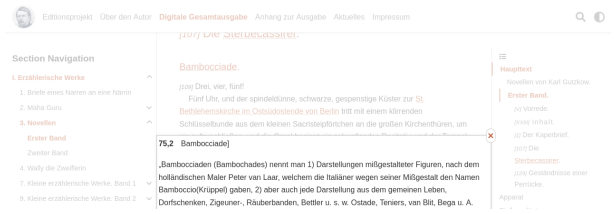


Abbildung 2: Screenshot der Gutzkow Edition. Demonstriert den Output eines mittels TEI-XML annotierten Werkes, inklusive der Funktionalität der Anzeige von Stellenkommentaren.

Wie schon oben beschrieben, generiert die µEdition statische HTML Seiten. Um den Publikationsflow noch weiter zu vereinfachen, stellt die µEdition Projektschablone die notwendige Funktionalität bereit um über zwei frei verfügbare Platformen die Edition zugänglich zu machen. Dies ist einerseits GitHub Pages[9], wozu die Edition nur in ein GitHub Repository eingespeist werden muss. Dann muss eine einzige Einstellung umgestellt werden und die Edition ist innerhalb weniger Minuten öffentlich verfügbar. Ähnlich funktioniert es über die Read the Docs Platform und das Projekt ist offen für die Unterstützung weiterer Platformen.

Die µEdition unterstützt sowohl Projekte von einer Einzelperson oder von Teams. Für Teams die mit Technologien wie Git umgehen wollen, kann die Edition leicht über ein derartiges Tool mit den Teammitgliedern geteilt werden. Andererseits kann die Edition auch einfach über Dokumentsharingplatformen wie Nextcloud oder Dropbox gemainsam bearbeitet werden. Die Nutzung von Git oder einer Dokumentsharingplatform ist auch wie in der µEdition die Archivierung umgesetzt wird. Dies ist natürlich nicht eine Langzeitarchivierung im formalen Sinn, aber in der µEdition sollen die Komplexität und Kosten einer formalen Langzeitarchivierung keine Schwelle für die Umsetzung der Edition sein.

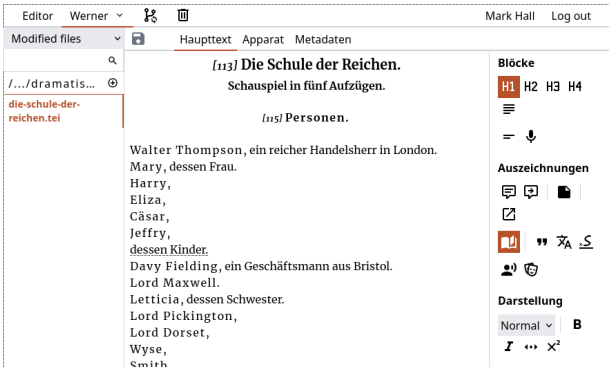


Abbildung 3: Screenshot des Editors. Zeigt den graphische TEI-XML Editor und rechts die Funktionalität für graphische Auszeichnungen.

Für Editionsprojekte die Git nutzen, wird im µEdition Projekt auch eine Editorumgebung entwickelt. Die Umgebung erleichtert den Umgang mit Git und bietet auch eine graphische Editorumgebung für die Bearbeitung von TEI-XML (Figure 3). Der Editor kann sowohl lokal genutzt werden, wie auch für das Team über das Internet bereitgestellt werden.

Diskussion

Der Minimalansatz der µEdition wirft natürlich auch ein paar interessante Fragen auf. Die größte Frage ist wie viel Edition braucht eine Edition um eine Edition zu sein? Zentral ist da besonders die Frage wie viel Annotation ist als Minimalanforderung notwendig. Die µEdition unterstützt sowohl TEI-XML, wie auch Markdown als Datenformate. Letzteres ist einfacher für den Einstieg, ist aber auch wesentlich weniger mächtig als TEI-XML. Trotzdem erlaubt es grundlegende Annotation der Textstruktur und über Fußnoten auch die kritische Annotation des Textes. Wenn also eine Edition Markdown als Anfangsformat nutzt und dann, im Laufe ihres Wachstums, die Möglichkeit der µE-

dition nutzt aus Markdown TEI-XML zu generieren und die ursprüngliche Markdownannotation durch die TEI-XML Version ersetzt, zu welchem Zeitpunkt wird die Edition eine „vollwertige“ Edition?

Die µEdition hat das Ziel den Einstig in das digitale Editionswesen zu vereinfachen und vertritt daher natürlich den Standpunkt, dass die Edition vom ersten Moment an „vollwertig“ ist. Das bedeutet aber nicht, dass das von der Wissenschaftscommunity auch so akzeptiert wird. Wenn die Schwelle für die „vollwertige“ Edition höher angelegt wird, dann stellt sich natürlich auch die Frage, ob die Nutzung der µEdition für Projekte überhaupt Sinn macht. Ich würde aber argumentieren, dass eine derartige, relativ willkürliche, Schwelle kontraproduktiv ist, wenn es darum geht, das Spektrum an Editionen zu erweitern.

Ein weiter Frage ist auch, ob die Annahmen der µEdition bezüglich der Schwellen ins Editionswesen überhaupt richtig liegt. Es ist natürlich möglich, dass das fehlen digitalisierter Inhalte oder die Schwierigkeiten diese zu finden die viel größeren Hindernisse zur Erweiterung der Editionslandschaft sind (Dziudzia und Hall, 2022). Diese Frage wird sich erst im Laufe der Zeit beantworten lassen und dann kommt natürlich auch die Frage auf, ab welchem Zeitpunkt man ein Projekt einstellt, um eine Fokussierung der Ressourcen der Community zu ermöglichen.

Ausblick

Es ist das Ziel der µEdition die Schwellen für den Einstieg in die Erstellung von Editionen zu vereinfachen. Dieses Ziel kann die µEdition nur erreichen, wenn sie auch von weiteren Projekten genutzt wird und durch deren Input verbessert wird. Zu diesem Zweck arbeitet das Projekt aktiv an der Weiterentwicklung der Software, insbesondere daran, die praktischen Schwellen zur Nutzung der µEdition weiter zu reduzieren und die Dokumentation für die einfache Nutzung der µEdition zu verbessern. Um dies erfolgreich umzusetzen ist das Feedback von Projekten, die die µEdition nutzen wollen, notwendig[10].

Die µEdition reduziert die Schwellen um Editionen zu erstellen und zu veröffentlichen. Um das Spektrum an Editionen zu erweitern ist das jedoch nur einer der notwendigen Schritte. Ein weiterer zentraler Schritt ist es, diese Editionen dann auch findbar zu machen. Da die Editionen als statische HTML Seiten verfügbar gemacht werden, ist es für die Internetsuchmaschinen natürlich auch einfach möglich, die Inhalte zu indizieren und findbar zu machen. Das verlässt sich aber darauf, dass der oder die Nutzer oder Nutzerin auch weiß, was die richtigen Schlüsselwörter sind, um die Edition zu finden. Das µEdition Projekt arbeitet daher auch daran, eine minimale Infrastruktur zu entwickeln, um einen zentralen Ausgangspunkt für das Entdecken neuer Editionen bereitzustellen.

Fußnoten

1. https://visualizationtechnology.wordpress.com/
2. https://teipublisher.com/
3. https://www.gutzkow.de
4. https://under-the-surface.research.room3b.eu
5. Die *statische Webseite* ist definiert als eine Webseite die primär nur aus HTML, CSS (für die Darstellung), und JavaScript (für die Interaktivität im Browser) Dateien besteht (Bilder, TEI Dateien, oder weitere Inhalte sind natürlich auch möglich). Was sie von einer dynamischen Webseite unterscheidet ist, dass insbesondere die HTML Inhalte nicht für jede Nutzerin oder jeden Nutzer dynamisch erzeugt wird, sondern dass die HTML Dateien einmal generiert werden und dann allen Nutzerinnen statisch zur Verfügung gestellt werden.
6. https://www.github.com
7. https://readthedocs.org
8. https://jupyterbook.org
9. https://pages.github.com
10. https://github.com/uEdition/uEdition/discussions

Bibliographie

Assmann, Aleida. "Canon and archive." *Cultural memory studies: An international and interdisciplinary handbook* (2008): 97-107.

Gioele, Barabucci, and Franz Fischer. "The formalization of textual criticism: Bridging the gap between automated collation and edited critical texts." In *Advances in Digital Scholarly Editing*, pp. 47-54. Sidestone Press, 2017.

Burrows, Toby, Matthew Holford, David Lewis, Andrew Morrison, Kevin Page, and Athanasios Velios. "Transforming TEI manuscript descriptions into RDF graphs." *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing* 15 (2021): 143.

Bürgermeister, Martina, Katharina Pektor, Christoph Steindl, and Johanna Eigner. "Offene Werkgenesen, Editionen und Archive. Versuch einer generischen Datenmodellierung." In *DHd*, p. 234. 2023.

Causer, Tim, Justin Tonra, and Valerie Wallace. "Transcription maximized; expense minimized? Crowdsourcing and editing the collected works of Jeremy Bentham." *Literary and linguistic computing* 27, no. 2 (2012): 119-137.

Dumont, Stefan, and Martin Fechner. "Bridging the gap: Greater usability for TEI encoding." *Journal of the Text Encoding Initiative* 8 (2014).

Dumont, Stefan, Tobias Kraft, Sabine Seifert, Christian Thomas, and Jan Wierzoch. "Die offene Edition. Vernetzung, Datenpublikation und Transparenz in der edition humboldt digital." In *DHd*, p. 121. 2023.

Dzidzia, Corinna und Hall, Mark. "Impulsvortrag: Einführung Zum Workshop "Repräsentativität in Digitalen Archiven". Zenodo, 2022, doi:10.5281/zenodo.6497680.

Estill, Laura. "Digital Humanities' Shakespeare Problem." *Humanities* 8, no. 1 (2019): 45.

Estill, Laura, Jennifer Guiliano, Élika Ortega, Melissa Terras, Deb Verhoeven, and Glen Layne-Worthey. "The circus we deserve? A front row look at the organization of the annual academic conference for the Digital Humanities." *DHQ: Digital Humanities Quarterly* 16, no. 4 (2022).

Fiormonte, Domenico. "Taxation against overrepresentation? The consequences of monolingualism for digital humanities." *alternative historiographies of the digital humanities* 2 (2021).

Fritze, Christiane. "Wohin mit der digitalen Edition? Ein Beitrag aus der Perspektive der Österreichischen Nationalbibliothek." *Bibliothek Forschung und Praxis* 43, no. 3 (2019): 432-440.

Galka, Selina, and Helmut W. Klug. "Minimal Editing: Die Hyperdiplomatische Transkriptionsplattform." In *DHd*, p. 217. 2023.

Giovannetti, Francesca, and Francesca Tomasi. "Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation." *Digit. Humanit. Q.* 16, no. 2 (2022).

Kudella, Christoph, and Neil Jefferies. "How Do We Model the Republic of Letters?." *Reassembling the Republic of Letters in the Digital Age* (2019).

Lauster, Martina. "Gutzkows Werke und Briefe, herausgegeben vom Editionsprojekt Karl Gutzkow: Ein Erfahrungs-und Werkstattbericht nach mehr als 20 Jahren." *Heine-Jahrbuch 2020* (2020): 207-224.

Pierazzo, Elena. "A rationale of digital documentary editions." *Literary and linguistic computing* 26, no. 4 (2011): 463-477.

Pierazzo, Elena. "What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter." *International journal of digital humanities* 1 (2019): 209-220.

Robinson, Peter MW. "Project-based digital humanities and social, digital, and scholarly editions." *Digital Scholarship in the Humanities* 31, no. 4 (2016): 875-889.

Rosselli Del Turco, Roberto. "The battle we forgot to fight: Should we make a case for digital editions?." In *Digital scholarly editing: theories and practices*, vol. 4, pp. 219-238. Open Book Publishers, 2016.

Sahle, Patrick. "What is a scholarly digital edition?." *Digital scholarly editing: Theories and practices* 1 (2016): 19-39.

Shakespeare, William, Stanley Wells, and Gary Taylor. *The oxford shakespeare: The complete works 2nd edition*. Oxford University Press, 2022.

Spence, Paul Joseph, and Renata Brandao. "Towards language sensitivity and diversity in the digital humanities." *Digital Studies/Le champ numérique* 11, no. 1 (2021).

Unsworth, John. "Computational work with very large text collections. Interoperability, Sustainability, and the TEI." *Journal of the Text Encoding Initiative* 1 (2011).

Van Zundert, Joris. "If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities." *Historical Social Research/Historische Sozialforschung* (2012): 165-186.

Epigraf – eine Plattform zur dokumenten- und datenorientierten Erfassung, Annotation, Vernetzung und Publikation von Textdaten

Jünger, Jakob
jakob.juenger@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0003-1860-6695

Gärtner, Chantal
chantal.gaertner@uni-muenster.de
Universität Münster, Deutschland

Herold, Jürgen
juergen.herold@uni-greifswald.de
Niedersächsische Akademie der Wissenschaften zu Göttingen, Deutschland

Michel, Maximilian
maximilian.michel@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Syring, Wolf-Dieter
wolf-dieter.syring@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Einleitung

Geisteswissenschaftliche Datenmodelle orientieren sich häufig daran, die vorgefundene Wirklichkeit zu modellieren: kulturelle Artefakte wie Schriftstücke, Bilder oder archäologisch aufbereitete Funde werden im Zuge der Erschließung strukturiert. Da entsprechende Bestände oft heterogen und lückenhaft sind, braucht es bei der Erfassung spezifisches Wissen über den jeweiligen Wissensbereich. Wenn die Gegenstände dabei umfassend abgebildet werden

sollen, erreichen Datenmodelle in der geisteswissenschaftlichen Forschung eine hohe Komplexität. Dennoch lassen sich drei typische Grundorientierungen bei der Modellierung unterscheiden: relationale Modelle zerlegen die Struktur in eine Vielzahl miteinander verbundene Datensätze, dokumentenorientierte Modelle stellen die Artefakte in den Mittelpunkt und erlauben auch gering schematisierte Modellierungen, graphorientierte Lösungen bilden die Inhalte über ein Netzwerk von Knoten und Kanten ab. Jede dieser Modellierungstechniken bringt nicht nur eine eigene Denkweise mit sich, sondern darüber hinaus Vor- und Nachteile bei der technischen Umsetzung. Datenbanken unterscheiden sich etwa in der Performanz von Lese- und Schreibzugriffen, der Flexibilität bei Strukturanpassungen, der Verbreitung für bestimmte Anwendungsfälle und nicht zuletzt in der Balance zwischen Komplexität und Nachvollziehbarkeit.

In unserem Beitrag stellen wir eine Plattform vor, die seit rund 20 Jahren in einem interakademischen Projekt zur Erschließung und Publikation von Inschriften eingesetzt und nun schrittweise für weitere Anwendungsfelder geöffnet wird. Der Beitrag fokussiert das über die Jahre konsolidierte Datenmodell, das die drei genannten Modellierungsperspektiven verbindet: Erstens ermöglicht es eine granulare, dokumentenorientierte Modellierung von Textdaten (Edition und Kollektion), zweitens basiert es auf einer relationalen Datenbank, was statistische, datenorientierte Analysen erleichtert, und schließlich lassen sich Netzwerkdaten ableiten, etwa um sie in Knowledge Graphs einzuspeisen. Für Nutzende verschwindet die Modellierung hinter der Anwendungsoberfläche. Sie spielt aber für die langfristige Weiterentwicklung eine wichtige Rolle, wenn es um die Analyseoptionen für fachwissenschaftliche Fragestellungen, die Wartbarkeit der Software, die Langzeitarchivierung der Daten und auch die institutionelle Einbindung geht. Wir verbinden die Vorstellung der Anwendung deshalb mit Details zum Datenmodell und verorten sie vor dem Hintergrund ähnlicher Anwendungen für die Erschließung von Inschriften und Texten.

Physische, logische und konzeptionelle Datenmodelle in den digitalen Geisteswissenschaften

Bei der Modellierung geisteswissenschaftlich relevanter Forschungsobjekte kann zwischen physischer, logischer und konzeptioneller Modellierung unterschieden werden (Flanders & Jannidis 2015, 3). Die physische Modellierung beschreibt, wie Daten auf Datenträgern abgespeichert werden und tritt in der Regel in den Hintergrund. Wichtig ist allerdings die Wahl des logischen Modells: „[The] choice of the logical model (e.g. a relational database instead of a markup language) determines the computational results or, better, the computational activities and operations on the data as based on the chosen model“ (Tomasi 2018, 174). Wenn Daten beispielsweise in tabellarischer Form vorgehalten werden, lassen sich darauf leicht statistische Analysen durchführen. Dagegen erlauben dokumentenorientierte Datenbanken, in denen Text mithilfe von Auszeichnungssprachen annotiert wird, eine flexiblere Erschließungstiefe und stärker auf das einzelne Objekt ausgerichtete Repräsentationen.

Auf dem logischen Datenmodell baut das konzeptionelle Modell auf, in welchem die formalen Datenstrukturen mit Bedeutung versehen werden. In den Digital Humanities sind auf der einen Seite universelle Ansätze zu finden, die beispielsweise auf standardisierte Vokabulare wie SKOS (W3C 2009) und das Resource Description Framework (W3C 2014) zurückgreifen. Auf der anderen Seite sind domänenspezifische Modelle auf den konkreten Untersuchungsgegenstand zugeschnitten, um etwa Inschriften (EpiDoc) oder andere Kulturgüter zu erfassen. Die Übertragung auf weitere Anwendungsgebiete ist dann nicht ohne Weiteres möglich – jedes Datenmodell repräsentiert den modellierten Gegenstand nicht nur, sondern reduziert ihn auf die für das Forschungsziel relevanten Eigenschaften (Stachowiak, 1973, 132-133). Bei der semantischen Modellierung werden Constraints eingeführt, die mögliche Datentypen und Beziehungen zwischen den Entitäten einschränken (Flanders & Jannidis, 2015, 5). Das resultierende Modell ist in der Regel auf eine einzelne Domäne – das heißt einen bestimmten Wirklichkeitsausschnitt – zugeschnitten. Datenmodellierung ist deshalb ein heuristischer Prozess (Diehr, 2021, 250), in dem Universalität (Interoperabilität) und Spezifität (Objektangemessenheit) ausbalanciert werden müssen.

Bei der konkreten Umsetzung sind in den Digital Humanities vor allem XML-Formate weit verbreitet, die sich etwa mit Universalsoftware wie Oxygen bearbeiten lassen. Das logische Modell ist damit festgelegt, das konzeptionelle Modell aber bleibt flexibel. Darauf aufbauend finden sich Softwarelösungen wie Ediarum (TELOTA, 2023), das eine Oxygen-Konfiguration für digitale Editionen und damit eine Spezifizierung des konzeptionellen Modells bietet, wobei die Datenhaltung in einer dokumentenbasierten Datenbank (eXist; Meier, 2003) erfolgt. Noch spezifischer auf eine konkrete Domäne angepasst sind Editoren wie EDEp (Horster et al., 2023), in dessen Oberfläche sich der in der Epigrafik etablierte TEI-EpiDoc-Standard widerspiegelt. Solche Anwendungen lassen die Datenformate zugunsten einfach bedienbarer Benutzeroberflächen in den Hintergrund rücken. Auch relationale Datenbanken werden für die Erfassung von Texten eingesetzt. So baut etwa die für Inhaltsanalysen etablierte Anwendung MAXQDA (VERBI, 2021) auf einer relationalen SQLite-Datenbank auf und das für Kulturgüter entworfene Cultural Heritage Framework wird in einer Vielzahl von Webanwendungen mittels der Typo3-Erweiterung Hisodat in SQL-Datenbanken implementiert (Schrade, 2017 & 2021). Ebenso werden Graphdatenbanken wie Neo4j in den digitalen Geisteswissenschaften zur Modellierung von Texten eingesetzt (Kuczera, 2016). Allerdings existieren soweit uns bekannt für

die Arbeit mit Texten bislang keine domänenspezifischen Nutzeroberflächen.

Die Anwendung Epigraf

Entstehungskontext und Anwendungsbereich

Mit der Anwendung Epigraf werden im Langzeitprojekt „Die Deutschen Inschriften“ historische Inschriften und deren Trägerobjekte erfasst und ediert. In den Arbeitsstellen der sechs beteiligten Akademien werden die Objekte fotografisch dokumentiert und die vorgefundenen Texte transkribiert, ediert und fachlich kommentiert. Die Ergebnisse der jeweils regional begrenzten Projekte werden in regelmäßiger Folge in gedruckter Form und im Web publiziert (DIO 2022). Epigraf ist darauf ausgerichtet, bei dezentraler Arbeitsweise einen konsistenten Datenbestand zu gewährleisten. Dazu wurden Module implementiert, die den gesamten Forschungsdatenlebenszyklus (Higgins 2008; Rümpel 2011) unterstützen (Abbildung 1):

- **Zusammenarbeit:** Jede Arbeitsstelle arbeitet in einer eigenen Datenbank. Auf diese Weise bleibt die organisatorische Autonomie der einzelnen, in unterschiedliche Akademien eingebundenen Arbeitsstellen, gewahrt. Durch ein einheitliches Domänenmodell bleiben die Daten aber vollständig kompatibel. Für die Koordination werden Wikis und ein Dateirepositorium eingesetzt.
- **Datenerfassung**: Der Kern der Anwendung besteht in Funktionen zur Erfassung von Objekten und Texten – für jedes Objekt wird ein Artikel erstellt oder importiert. Die Projektdatenbanken sind flexibel konfigurierbar, um unterschiedliche Gegenstände wie Inschriften, Regesten, Briefe oder Social-Media-Posts zu erfassen. Dazu enthält jede Datenbank eine eigene Konfigurationstabelle, in der das konzeptionelle Modell abgebildet wird.
- **Annotation:** Jeder Artikel setzt sich aus flexibel kombinierbaren Abschnitten zusammen, in denen der Text und alle relevanten Metadaten (Beschreibungen, Kommentare, Kategorisierungen über Vokabulare) sowie zugehörige Dateien (Bilder) enthalten sind. Für die Annotation von Texten steht eine projektspezifisch konfigurierbare Toolbar zur Verfügung (Abbildung 2).
- **Vernetzung**: Um die Daten später als Linked Open Data nach den FAIR-Prinzipien (Wilkinson et al. 2016) zu veröffentlichen, können für jeden Artikel und jede Kategorie Normdatenbezeichner (IRIs; W3C 2014) erstellt werden. Dadurch lassen sich Datenbestände zwischen verschiedenen Projektdatenbanken abgleichen. Das Datenmodell von Epigraf ist darüber hinaus mit dem Resource Description Framework (W3C 2014) kompatibel, so dass die Beziehungen zwischen Datenpunkten in der Form von Aussagen modellierbar sind.
- **Analyse:** Der Gesamtbestand ist im Volltext durchsuchbar und die zur Erschließung verwendeten Vokabulare können als Rechercheinstrument eingesetzt werden (Abbildung 3). Durch die tabellarische Erfassung lassen sich auch statistische Verfahren wie Clusteranalysen durchführen.
- **Publikation:** Über ein Pipeline-System werden die erfassten Daten mittels XSL-Stylesheets in Word-Dateien konvertiert, die nach der Endredaktion als Druckvorlage für einen Inschriftenband an den Verlag übergeben werden. Ein solcher Band umfasst mehrere hundert Artikel, Einleitungen und ein Register. Dieser klassische Publikationsworkflow wird zunehmend erweitert, so dass über eine Programmierschnittstelle CSV-, JSON- oder XML-Daten sowie standardisierte Dokumentformate wie TEI bzw. EpiDoc (TEI 2022; Elliott et al. 2020) ausgegeben werden.

Epigraf ist eine Webanwendung, die auf Standardtechnologien der Webentwicklung basiert. Das Projekt Die Deutschen Inschriften, in dessen Kontext die Anwendung entwickelt wird, strebt über die NFDI-Konsortien Text + und NFDI4Culture eine langfristige Integration in die Forschungslandschaft an. Für diesen Zweck werden Programmierschnittstellen etabliert, mit denen Abfragen zum Beispiel in föderierte Suchfunktionen integriert werden können. Weiterhin findet eine Öffnung der Anwendung für andere Projekte statt, sodass Epigraf ggf. zukünftig selbst als Datenzentrum fungieren kann, um Daten langfristig verfügbar zu halten.

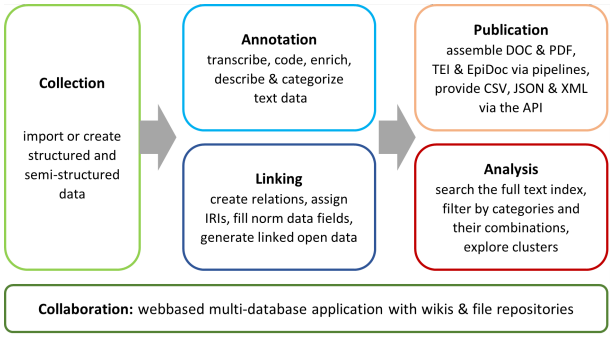


Abbildung 1: Module von Epigraf

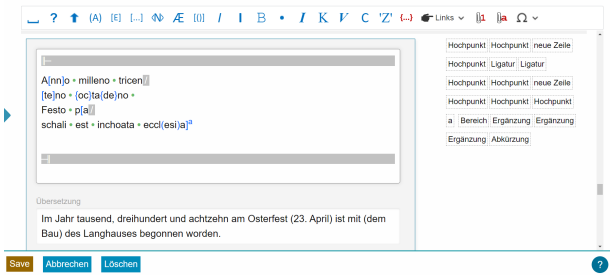


Abbildung 2: Beispiel für den Transkriptionseditor (Leidener Klammersystem)

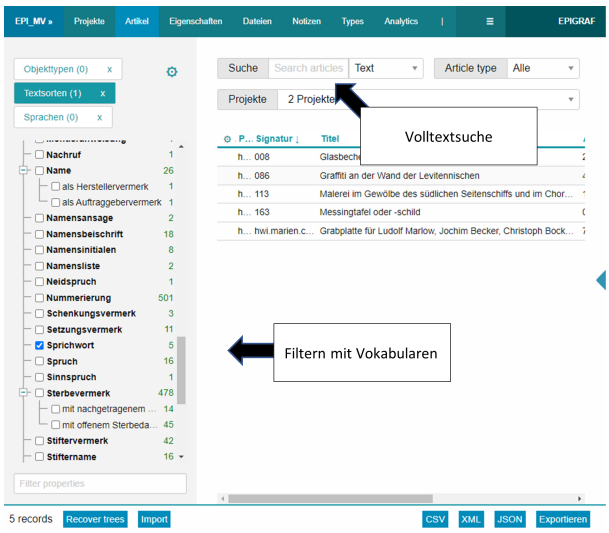


Abbildung 3:Beispiele für Recherchemöglichkeiten innerhalb der Korpora

Das Datenmodell von Epigraf

Epigraf war entsprechend der typischen dokumentenorientierten Herangehensweise in den Digital Humanities in den ersten Versionen ein auf den Gegenstand zugeschnittener XML-Editor. Im Verlauf der Jahre sind die eingesetzten Vokabulare, beispielsweise für Worttrenner, Personennamen oder Objekttypen, auf mittlerweile über 20.000 Kategorien für eine Projektdatenbank angewachsen. Abfragen über solche Vokabulare und auch Aktualisierungen lassen sich einfacher in einer relationalen Datenbank umsetzen, weshalb die Anwendung zu einer SQL-Datenbanken migriert wurde. Dabei wurden die Objekte zunächst in über 80 Tabellen vollständig normalisiert (d.h. redundanzarm) modelliert. Diese Komplexität wurde im Verlauf der Jahre Zug um Zug ohne Informationsverlust auf aktuell 10 inhaltliche Tabellen reduziert (Abbildung 4).

Die relationale Struktur folgt mittlerweile dem Entity-Relationship-Model (Chen, 1976), um eine Hierarchie aus Projekten, Artikeln, Abschnitten und Inhalten festzulegen. Das Epigraf-Datenmodell bietet lediglich Optionen an, mit denen textbasierte Daten strukturiert werden können. Welche konkreten Inhalte in dieser Struktur erfasst werden, und damit das Domänenmodell, lässt sich über eine Typisierung konfigurieren. Einzelne Datensätze enthalten unter anderem Felder für Text, Dateien, Datierungen oder Verlinkungen zu anderen Tabellen, ohne dass deren Bedeutung von vornherein festgeschrieben ist. Die Bedeutung wird in einer durch die Nutzer:innen bearbeitbaren Metatabelle vorgegeben – dort wird unter anderem definiert, welche Datensatztypen erlaubt sind und welche Beschriftung die Felder erhalten. Beispielsweise kann das Feld „content“ je nach Datensatztyp als „Beschreibung“ oder als „Transkription“ bezeichnet werden. Über die Typisierung wird ebenfalls festgelegt, welche Annotationen möglich sind. In der Bearbeitungsoberfläche wird daraus die Toolbar abgeleitet, sodass etwa in einer Beschreibung nur numerische Fußnoten und im Transkriptionsfeld zusätzlich alphabetische Fußnoten für textkritische Kommentare möglich sind.

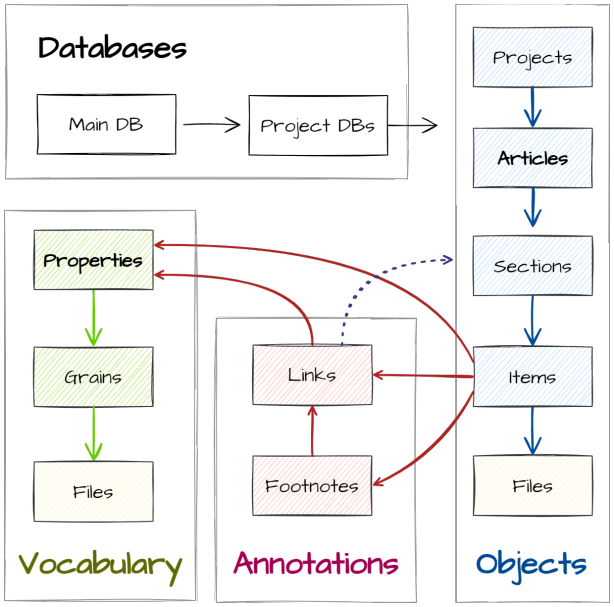


Abbildung 4: Das Datenmodell von Epigraf (vereinfachte Darstellung)

Beschreibungen komplexer Objekte werden entsprechend dem Entity-Attribute-Value-Model durch mehrere, in einem gemeinsamen Abschnitt zusammengefasste Datensätze umgesetzt. Das Entity-Attribute-Value-Model geht historisch auf medizinische Anwendungen zurück, in denen heterogene Datensorten strukturiert erfasst werden müssen (siehe Friedman et al., 1990; Nadkarni et al., 1999). Entstehen neue semantische Anforderungen, dann werden keine neuen Felder eingeführt – das logische Datenmodell bleibt konstant – sondern neue Datensatztypen definiert, die in einer Attributspalte der Datensätze referenziert werden. Um darüber hinaus eine flexible Auszeichnung von Texten und zusammengesetzte Daten wie Georeferenzierungen zu ermöglichen, erlaubt das Datenmodell in einzelnen Feldern die Verwendung von XML und JSON.

So ist eine tiefgehende Strukturierung von Forschungsdaten möglich, die zugleich anschlussfähig für eine Vielzahl von Anwendungsbereichen bleibt. Zwar sind Schreibvorgänge in eine solche Datenbank kostenintensiv, denn es sind bei der Bearbeitung eines einzigen Objekts mitunter Hunderte von Datensätzen mit den einzelnen Werten betroffenen. Da aber Speichervorgänge in wissenschaftlichen Editionen kaum zeitkritisch sind, kann dieser Nachteil zugunsten einer hohen Flexibilität in Kauf genommen werden. Die Entwicklung von Forschungssoftware auf der Basis eines solchen Datenmodells ist vor allem für Auswertungen effizient, da Analysefunktionen mit nur wenigen Anpassungen in verwandten Datenbeständen genutzt werden können.

Das konzeptionelle Datenmodell, bestehend aus Artikeln, Abschnitten, Inhalten und Kategorien, bleibt stets konstant. Das Domänenmodell spiegelt dagegen in Editionsprojekten

die jeweiligen Bearbeitungsrichtlinien eines Projekts wider, selbst wenn der Gegenstandsbereich ähnlich ist. Für die Inschriftenforschung liegen aktuell eine Konfiguration für die Erschließung im Projekt Die Deutschen Inschriften und eine Konfiguration für Inschriften im Projekt Corpus Inscriptionum Latinarum vor. Während beispielsweise im ersten Projekt vorrangig eine Transkription je Inschrift erarbeitet wird, sind bei den älteren lateinischen Inschriften mehrere Lesevarianten zu berücksichtigen.

Das in den vergangenen Jahren kontinuierlich weiterentwickelte Datenmodell von Epigraf lässt sich ebenso für die Erschließung und Analyse von Objekten außerhalb der epigrafischen Forschung (Abbildung 5) einsetzen. Eingeflossen sind in die Anwendung unter anderem Erfahrungen aus der Arbeit mit kritischen Editionen, qualitativen Inhaltsanalysen und (teil)automatisierten Inhaltsanalysen von Texten und Bildern. Das Modell ist dadurch so weit abstrahiert, dass es sich für eine Vielzahl an Kurztexten inklusive multimodaler Elemente eignet. Dazu werden die Artikel, Abschnitte, Inhalte, Annotationen, Fußnoten und Kategorien über die Metatabelle entsprechend typisiert, so dass ein Modell für die jeweilige Domäne entsteht. Eines der bereits erprobten Anwendungsfelder ist die Analyse von Social-Media-Daten. Hierzu können die verschiedenen Arten von Mitteilungen (Posts, Kommentare, Antwortkommentare) als Threads – das heißt eine Abfolge von Posts und Kommentaren – modelliert (Abbildung 6). Eine solche Zusammenstellung ist beispielsweise in der kommunikationswissenschaftlichen Forschung sinnvoll, um interpretative mit statistischen Methoden zu verbinden. Epigraf macht die Mitteilungen für Inhaltsanalysen im Kontext lesbar und durchsuchbar, abstrahiert dabei von der Darstellung einzelner Plattformen wie Facebook, Instagram oder Twitter und kann die Annotation von Themen, Sprechakten oder weiteren Konstrukten unterstützen. Die tabellarische Form (beim Importieren oder Exportieren als CSV-Datei) ermöglicht wiederum Auszählungen oder Regressionsanalysen mit gängigen Statistikprogrammen oder Programmiersprachen wie R und Python.

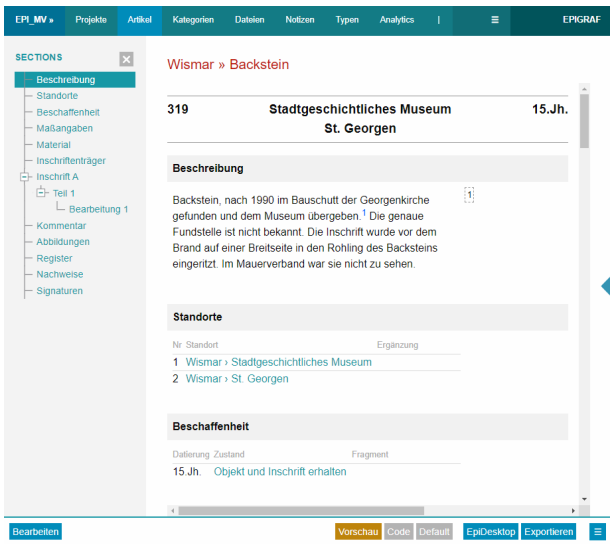


Abbildung 5: Beispiel für einen Inschriftenartikel in Epigraf (work in progress)

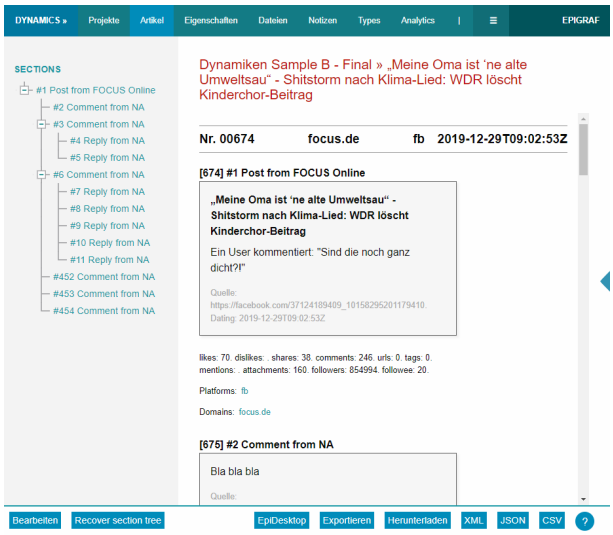


Abbildung 6: Beispiel für einen Facebook-Thread in Epigraf (work in progress)

Zusammenfassung und Ausblick

Epigraf ist eine Plattform zur Erfassung, Annotation, Vernetzung und Publikation von Textdaten, die seit rund 20 Jahren im Kontext eines historischen Editionsprojekts entwickelt wird. Die Modellierung ist gleichzeitig auf eine dokumentenorientierte, detaillierte sowie tiefgehende Aufbereitung des Quellenmaterials und auf eine datenorientierte Strukturierung für statistische Analysen ausgerichtet. Das konzeptionelle Datenmodell lässt sich über eine Konfiguration auf unterschiedliche Domänen anpassen, von Inschriften über Briefe bis zu Social-Media-Posts. Mittels Export-Pipelines werden Standardformate wie TEI oder netzwerkorientierte Modelle wie RDF unterstützt. Fortlaufend werden Funktionen aufgebaut, Analyse- und

Recherchefunktionen in umfangreichen Korpora getestet, Programmierschnittstellen eingerichtet und Packages zur Interaktion mit der Datenbank für R und Python entwickelt. Ein Open-Source-Release ist in Vorbereitung und eine Instanz für Wissenschaftler:innen außerhalb des Inschriftenprojekts befindet sich in der Testphase.

Bibliographie

Chen, Peter Pin-Shan. 1976. "The entity-relationship model—toward a unified view of data." ACM Transactions on Database Systems 1 (1): 9–36. 10.1145/320434.320440.

Diehr, Franziska. 2021. "Modelling in Digital Humanities: An Introduction to Methods and Practices of Knowledge Representation." In Music - Media - History. Bd. 44, hg. von Matej Santi und Elias Berner, 241–62. Bielefeld: transcript.

DIO. 2022. "Deutsche Inschriften Online: Projektbeschreibung." https://www.inschriften.net/projekt.html (zugegriffen: 13. Juli 2022).

Elliott, Tom, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, Scott Vanderbilt et al. 2020. "EpiDoc Guidelines: Ancient documents in TEI XML (Version 9)." https://epidoc.stoa.org/gl/latest/ (zugegriffen: 29. Juli 2022).

Flanders, Julia und Fotis Jannidis. 2015. "Knowledge Organization and Data Modeling in the Humanities." https://wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf (zugegriffen: 29. Juli 2022).

Friedman, Carol, George Hripcsak, Stephen B. Johnson, James J. Cimino und Paul D. Clayton. 1990. "A Generalized Relational Schema for an Integrated Clinical Patient Database." Proceedings of the Annual Symposium on Computer Application in Medical Care, 335–39.

Higgins, Sarah. 2008. "The DCC Curation Lifecycle Model." International Journal of Digital Curation 3 (1): 134–40. 10.2218/ijdc.v3i1.48.

Horster, Marietta, Francisca Feraudi-Gruénais, David Eibeck, Jörn Turner, Wolfgang Meier. 2023. "Editionstools für eine Digitale Epigraphik (EDEp)". https://edep.adw.uni-heidelberg.de (zugegriffen: 19.01.2023).

Kuczera, Andreas. 2016. "Graphbasierte digitale Editionen" In: Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte. https://mittelalter.hypotheses.org/7994 (zugegriffen 19. Juli 2023).

Meier, Wolfgang. 2003. "eXist: An open source native XML database." In Web, Web-Services, and Database Systems, hg. von Gerhard Goos et al., Bd. 2593, 169–183. Springer. https://doi.org/10.1007/3-540-36560-5_13

Nadkarni, Prakash. M., Luis Marenco, Roland Chen, Emmanouil Skoufos, Gordon Shepherd und Perry Miller. 1999. "Organization of Heterogeneous Scientific Data Using the EAV/CR Representation." Journal of the American Medical Informatics Association 6 (6): 478–93. 10.1136/jamia.1999.0060478.

Rümpel, Stefanie. 2011. "Der Lebenszyklus von Forschungsdaten." In Handbuch Forschungsdatenmanagement, hg. von Stephan Büttner, Hans-Christoph Hobohm und Lars Müller, 25–34. Bad Honnef: Bock + Herchen.

Schrade, Torsten. 2017. "Sammlungs- und Editionsportale mit dem Cultural Heritage Framework der Digitalen Akademie." https://digicademy.github.io/2017-editionsportale-jena (zugegriffen 19. Juli 2023)

Schrade, Torsten. 2021. "Historical Sources Online Database (hisodat)." https://github.com/digicademy/hisodat (zugegriffen 19.Juli 2023).

Stachowiak, Herbert. 1973. Allgemeine Modelltheorie. Wien: Springer.

TEI. 2022. "Text Encoding Initiative." https://tei-c.org/ (zugegriffen: 24. Mai 2022).

TELOTA. 2023. Ediarum. Berlin-Brandenburgischen Akademie der Wissenschaften. https://www.ediarum.org/ (zugegriffen 19. Juli 2023)

Tomasi, Francesca. 2018. "Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain." Historical Social Research, 170–79. 10.12759/hsr.suppl.31.2018.170-179.

VERBI Software. 2021. "MAXQDA. Version 2022." https://www.maxqda.com (zugegriffen: 19.07.2023)

W3C. 2009. "SKOS Simple Knowledge Organization System Reference: W3C Recommendation 18 August 2009." https://www.w3.org/TR/2009/REC-skos-reference-20090818/ (zugegriffen: 28.07.2022).

W3C. 2014. "RDF 1.1 Concepts and Abstract Syntax: W3C Recommendation 25 February 2014." https://www.w3.org/TR/rdf11-concepts/ (zugegriffen: 28.07.2022).

Wilkinson, Mark D., Michel Dumontier, I. Jsbrand 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 3:1–9. 10.1038/sdata.2016.18.

Erfahrungen aus dem Citizen Science-Projekt Itinera Nova als Taktgeber für Digital Humanities-Projekte

Bigalke, Jan
Jan.Bigalke@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0003-0721-9980

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-1438-379X

Drach, Sviatoslav
s.drach@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0003-3324-1655

Löbbert, Benedikte
b.loebbert@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0009-0002-6670-3692

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-9377-9492

Einleitung

Im April 2022 wurde das Weißbuch "Citizen-Science-Strategie 2030 für Deutschland" mit 94 Handlungsempfehlungen für 15 Handlungsfelder veröffentlicht, mit dem Ziel, die Citizen Science in Gesellschaft und Wissenschaft zu stärken, um deren Innovationspotenziale entfalten zu können (Bonn et al., 2021). Die formulierten hohen Erwartungen an Citizen Science treten aber nicht selten in Kontrast zu den tatsächlichen Projektergebnissen (Smolarski et al., 2023).

Ein positives und prominentes Beispiel für ein Citizen-Science-Projekt in den Digital Humanities ist das Projekt Gruß & Kuss (Rapp et al., 2022), das Bürger:innen mit Hilfe von entwickelten Tools und entsprechenden Veranstaltungen und Schulungen in den Prozess der Digitalisierung und Untersuchung einbindet.

In diesem Vortrag stellen wir Erfahrungen aus dem Kontext des Cologne Center for eHumanities (CCeH) und der Zusammenarbeit mit verschiedenen beteiligten Akteur:innen – Freiwillige, Archivmitarbeiter:innen, Fachwissenschaftlern:innen und Research Software Engineers – vor und diskutieren diese anhand dreier ausgewählter Handlungsfelder des international beachteten Weißbuchs, die aus Perspektive eines DH-Kompetenzzentrums besonders relevant sind: 1) Freiwilligenmanagement, 2) Datenqualität und Datenmanagement, 3) Sensorik und künstliche Intelligenz.

Ausgangspunkt bildet das Projekt *Itinera Nova*, das mit großem Erfolg wesentlich auf Freiwilligenarbeit bei der Transkription und Erschließung von Quellen setzt und schon seit 2009 technisch-methodisch vom CCeH begleitet wird. Diesem bereits fest etablierten Citizen-Science-Projekt stellen wir zwei sehr unterschiedlich verfasste Projekte gegenüber. Es handelt sich dabei um die jeweils seit 2022 geförderten Projekte *DigiByzSeal* (ANR-DFG, Gepris #469385434) und *Beginen in Köln* (DFG, Gepris #491803989; Böhringer und Weber, 2022; Bigalke et al., 2023). Anders als im Fall von *Itinera Nova* handelt es sich hier um "klassische" DH-Kooperationsprojekte mit Fachwissenschaftler:innen, bei denen keine öffentliche Beteiligung im Sinne der Citizen Science vorgesehen ist. Entlang der oben genannten Handlungsfelder stellen wir heraus, wie Vorgehensweisen und Erfahrungen aus dem Bereich der Citizen Science in der Zusammenarbeit mit Fachwissenschaftler:innen Anwendung finden und skizzieren ein gemeinsames infrastrukturelles Modell, das sich dabei bewährt hat, jeweils zugeschnittene Werkzeuge zur Verfügung zu stellen, um nicht-technisch versierten Bearbeiter:innen schnell in die Lage zu versetzen, an der Datenbearbeitung mitzuwirken.

Ausgangslage

Das Projekt *Itinera Nova* hat die Digitalisierung und Erschließung der Schöffenregister und Rechnungsbücher der belgischen Stadt Löwen zum Ziel. Der betreffende Bestand umfasst weit über 1.300.000 Seiten aus dem Zeitraum 1361–1795. Die Bearbeitungspipeline reicht vom Erstellen der Digitalisate über das Anlegen von Metadaten bis zur Archivierung der Rohdaten. Digitalisate und Metadaten werden auf der Projektwebseite zunächst intern den Bearbeiter:innen zugänglich gemacht und dort indexiert, d.h. einzelne Akte werden auf den Digitalisaten identifiziert und wiederum mit Metadaten versehen. Die einzelnen Akte werden transkribiert und nach abschließender Prüfung veröffentlicht.

Das Ziel des Projekts *DigiByzSeal* ist es, byzantinische Bleisiegel zu digitalisieren, zu beschreiben und damit ein umfassendes Corpus zur weiterführenden Forschung in der Byzantinistik zur Verfügung zu stellen. Die betreffenden Sammlungen umfassen ca. 4000 Siegel. Die Auszeichnung

und Annotation der Siegel erfolgt in XML, in dem auf EpiDoc basierenden SigiDoc-Schema.

In dem Projekt *Beginen in Köln* soll eine über Jahrzehnte erstellte und gepflegte Textdatenbank mit rund 3000 Einträgen mit semantischen Kategorien erschlossen werden. Gegenstand der Datenbank sind Regesten zu den mittelalterlichen Schreinsbüchern der Stadt Köln, die eine detaillierte prosopographische, soziale und sozial-topographische Auswertung zu Frauen vorlegen, die die frommen Lebensweise der Beginen wählten.

Handlungsfelder

Handlungsfeld Freiwilligenmanagement

„Im Jahr 2030 zeichnen sich Citizen-Science-Projekte durch ein professionelles Freiwilligenmanagement aus.“ (Bonn et al., 2021)

Im Mittelpunkt des *Itinera Nova*-Projektes am Stadtarchiv Leuven steht eine Freiwilligencommunity von rund 50 Personen, die im Wesentlichen von einer/m hauptamtlichen Archiv-Mitarbeiter:in betreut wird. Freiwillige sind an allen oben genannten Schritten der Bearbeitung nicht nur beteiligt, sondern führen diese maßgeblich aus. Der Arbeitsschwerpunkt liegt bei der zeitintensiven Transkription. Die dauerhafte Finanzierung der Erschließung eines so umfangreichen Bestands in dieser Tiefe wäre ohne das Modell des zivilen Engagements kaum denkbar.

Die Freiwilligencommunity muss dauerhaft aufrechterhalten und ausgebaut werden. Die Rolle des Community-Managements ist dabei entscheidend (so auch Vohland et al., 2021). Das in Leuven bereits langfristig gepflegte “Vereinsleben” umfasst verschiedene Events wie Meetings in Kleingruppen, Workshops und Feiern im Stadtarchiv, bei denen die Akteur:innen die aktuellen Projektaufgaben besprechen, sich über die spezifischen Themen austauschen sowie die Kompetenzen ausbauen, die offenen Fragen klären oder einfach nur sozial zusammenkommen. Die Folgen der Corona-Pandemie, Vielfalt von Aufgaben und Kompetenzen, der Austausch mit vielen Akteur:innen intern und auf internationaler Ebene und das Bereitstellen von Tools und Räumlichkeiten (sowohl digitale als auch analoge) stellen dabei große Herausforderungen dar.

Diese Tools umfassen sowohl am CCeH bereitgestellte, wie die eigene Projekt-Webplattform, als auch externe, wie zum Beispiel Transkribus[1]. Durch Feedbackrunden wird sichergestellt, dass die Anmerkungen und Wünsche von Citizen Scientists (vor allem im Bereich UX/UI) ermittelt und berücksichtigt werden können.

In dem Projekt *DigiByzSeal* sind rund 10 Siegelkundler:innen verteilt auf die deutschen und französischen Projektstandorte unmittelbar eingebunden. Der Adressatenkreis des Vorhabens ist jedoch weiter, denn schließlich soll der international verteilten Forschungscommunity mit dem SigiDoc-Profil ein auf EpiDoc/TEI basierendes Modell für die Beschreibung von Siegeln an die Hand gegeben werden. Das Projekt veranstaltet Workshops und Schulungen, bei denen schnell der Bedarf nach einem Tool ermittelt wurde, mit dem Fachwissenschaftler:innen, auch ohne technische Kenntnisse, Siegel annotieren können. Moderation des Interessent:innenkreises und Qualitätsmanagement werden aus dem Projekt geleistet. Die Erfahrungen aus dem *Itinera Nova*-Projekt zeigen jedoch, dass die Rolle des Community-Managements auf Dauer gestellt sein muss.

Obwohl das Vorhaben *Beginen in Köln* als verhältnismäßig kleines DFG-Projekt nicht für die Organisation eines Citizen-Science-Ansatzes ausgelegt ist, wird die Idee regelmäßig von außen an das Projekt herangetragen. Dies liegt zum einen an dem Forschungsgegenstand, der eine an Frauen- und Lokalgeschichte interessierte Öffentlichkeit anspricht, zum anderen an der vielschichtigen Quellenbasis, den Schreinsbüchern, deren umfängliche systematische Erschließung ein vielfach formuliertes Desiderat darstellt. Die Erfahrungen aus *Itinera Nova* zeigen jedoch, dass Crowdsourcing/Citizen Science sich nicht in der Bereitstellung einer technischen Eingabe- und Moderationsplattform erschöpfen.

Handlungsfeld Datenqualität und Datenmanagement

„Im Jahr 2030 existieren wiederverwendbare, flexible Methoden und Werkzeuge für die Erhebung, die Qualitätssicherung und -kontrolle, die Analyse, die Archivierung und die Veröffentlichung von Citizen-Science-Daten.“ (Bonn et al., 2021)

Das *Itinera Nova*-Projekt verfolgt den Anspruch wissenschaftlich hochwertige Daten vorzulegen und stützt sich daher auf etablierte Standards wie XML-TEI. Für Freiwillige ohne entsprechende Kenntnisse stellt dies eine große Hürde dar. Dank dem einfach zu benutzenden Editor, der keine Kenntnisse von XML-TEI voraussetzt, wurde es den Freiwilligen ermöglicht, Transkriptionen anzulegen und sie zu annotieren. Die aktuell erstellten Metadaten umfassen zum großen Teil nur die wichtigsten Angaben wie Name, Datum und Sprache der Akten. Wünschenswert wären umfassendere Metadaten gewesen, auf die aber bewusst zur Reduktion der Komplexität verzichtet wurde. Die transkribierten Texte können festgelegte Annotationen wie Abbreviatur, hochgestellt, unklar, gestrichen, neue Zeile und neue Seite umfassen. Auch hier wurde aus denselben Gründen auf die Auszeichnung von Personen, Orte und weitere Entitäten verzichtet.

Mit Hilfe der in die *Itinera Nova* Plattform implementierten Redaktions- und Moderationssysteme wurde die Infrastruktur für Revision und Qualitätsmanagement geschaffen. Die von Citizen Scientists erfassten Daten werden dann durch erfahrene Moderator:innen (die ebenso Citizen Scientists sind) überprüft und erst dann veröffentlicht.

Solche Infrastrukturen für Erhebung und Qualitätssicherung der Daten, die den Projektbeteiligten ohne technische Kenntnisse bzw. Kenntnisse von XML-TEI oder anderen Standards, die Mitarbeit an Digital-Humanities-Projekten

ermöglichen, sind essentiell. Dies bestätigt sich in zahlreichen Digital-Humanities-Projekten, auch in denen, die nicht im engeren Sinne der Citizen Science zuzuordnen sind.

Das Projekt *DigiByzSeal* setzt genau auf diese Erfahrungen auf. Hier können die Fachwissenschaftler:innen die Daten zu den Siegeln in eine auf das Datenmodell angepasste Vorlage eintragen. Eine Balance zwischen Pragmatismus und Präzision zu finden, ist dabei zentral. Reduktion von Redundanzen und Abstriche in der Tiefenerschließung führen im Gegenzug dazu, dass die Benutzeroberfläche einfacher und übersichtlicher gestaltet werden, effektiver genutzt werden kann und mehr Akzeptanz findet.

Erfahrungen, die in *Itinera Nova* mit einer Community ohne IT-Kenntnisse gemacht wurden, gelten auch für Projekte, in denen die IT-Unterstützung auf eine einzelne Person abzielt. Bei dem Projekt *Beginen in Köln* wurde statt eines komplexen, auf TEI basierenden Modells ein stark vereinfachtes Datenmodell gewählt. Austauschformate in TEI und RDF/XML können dennoch verlustfrei generiert werden. Für dieses wird ein Formulareditor erstellt, der die beteiligte Fachwissenschaftlerin bei einer konsistenten semantischen Erschließung unterstützt.

Handlungsfeld Sensorik und künstliche Intelligenz

„Im Jahr 2030 sind Sensorik und künstliche Intelligenz etablierte Werkzeuge für Citizen-Science-Aktivitäten.“ (Bonn et al., 2021)

Die aktuellen KI-Ansätze sind im Bereich der Texterkennung schon jetzt sehr vielversprechend. Im Projekt *Itinera Nova* werden Handwritten Text Recognition-Tools (HTR) wie Transkribus für automatische Handschrifterkennung eingesetzt. Mit Hilfe der erstellten Modelle[2] lassen sich gute Ergebnisse[3] erzielen. Doch dafür sind eine große Menge von Trainingsdaten notwendig. Da die Quellen aus einer großen zeitlichen Spanne kommen, müssen idealerweise für alle paar Jahrzehnte die Trainingsdaten bzw. Modelle produziert werden. Außerdem ist geplant, Named Entity Recognition-Verfahren (NER) zu verwenden. Für das Training dieser Modelle gibt es momentan allerdings sehr wenige Trainingsdaten. Abgesehen davon, müssen die Ergebnisse von Menschen kontrolliert und ggf. korrigiert werden. Die Bedienung von KI-Tools setzt teilweise spezifische technische Kompetenzen voraus und kann in den Workflow für alle Beteiligten (noch) nicht unmittelbar eingebunden werden.

Künstliche Intelligenz im Bereich der Sigillographie ist noch am Anfang, ein Projekt an der Sorbonne in Paris beschäftigt sich mit der Erkennung und Klassifikation von Ikonographien auf Siegeln (Eyharabide et al., 2023). Hier können die im Projekt zum ersten Mal systematisch digital erhobenen Siegel einen Mehrwert beim Training bieten und zukünftig von den Ergebnissen profitieren, wenn weitere Siegel digitalisiert und die Ikonografien mithilfe des Modells klassifiziert werden sollen.

Bei den *Beginen in Köln* würden sich Methoden der künstlichen Intelligenz anbieten. Vor allem NER zum automatischen Erkennen und Auszeichnen von Personen und Orten sowie die Extraktion von semantischen Verknüpfungen sind hier vielversprechende Ansätze. Erste Versuche, die hier durchgeführt wurden, lieferten keine schlechten Ergebnisse, doch um wirklich gute Ergebnisse erzielen zu können, müssten Modelle speziell auf die Regesten trainiert werden. Dazu fehlte uns leider die Trainingsgrundlage. Daher scheint die Kosten-Nutzen-Rechnung keine Vorteile gegenüber der manuellen Annotation zu bieten.

Gemeinsame Architektur

Bei allen drei Projekten, *Itinera Nova*, *DigiByzSeal* und *Beginen in Köln*, ist es von großem Vorteil, dass wir dieselbe Basis-Infrastruktur für die ähnlich gelagerten Anwendungsfälle nutzen können. Die Datenmodelle sind jeweils in XML umgesetzt, auch wenn hier in allen Projekten unterschiedliche XML-Schemata verwendet werden. Als Datenbank wird die XML-Datenbank BaseX verwendet, auf Basis derer APIs entwickelt und bereitgestellt werden. Diese fungieren als “Middleware” zwischen der Datenbank und dem Frontend. Das Frontend ist in allen Projekten mit VueJS entwickelt worden. Diese Basis-Infrastruktur ermöglicht es, in relativ kurzer Zeit Tools und Benutzeroberflächen zu entwickeln, die auf die individuellen Bedarfe im Projekt angepasst sind.

Zusammenfassung

Die Citizen-Science-Konzepte bieten viele Chancen für Digital Humanities und bei manchen Projekten (hier: *Itinera Nova*) geradezu alternativlos. Die Erfahrungen in der Beteiligung an einem etablierten Citizen-Science-Projekt als DH-Kompetenzzentrum mit einer Vielzahl von forschungsgetriebenen Vorhaben haben aufgezeigt, dass engagiertes und dauerhaftes Freiwilligenmanagement der Schlüssel zu einer erfolgreichen Einbindung der Öffentlichkeit oder Community darstellen. Die Vorstellung, dass Citizen Science zur “Rekrutierung” von freiwilliger Arbeitskraft alleine durch die Bereitstellung von Crowdsourcing-Plattformen implementiert werden kann, ist vor diesem Hintergrund kritisch zu beurteilen.

Auf der Ebene des Datenmanagement und der Datenqualität ist eine erprobte Infrastruktur, die den Workflow abbildet und Editor, Redaktions- und Moderationssystem umfasst, ein zentraler Baustein. Hierbei verwendete Infrastrukturbausteine haben sich ebenfalls bewährt bei Projekten mit kleinen oder Kleinstgruppen, bei denen der Bedarf besteht, technisch nicht geschulte Bearbeiter:innen in die Datenaufbereitung einzubeziehen.

Machine Learning und KI bieten in den hier vorgestellten Projekten mittelfristig das Versprechen, den Bearbeitungsaufwand durch bspw. OCR/HTR-Vorerfassung, NER oder semantischer Verknüpfung zu reduzieren und damit die Ar-

beit für die hier beschriebene Zielgruppe attraktiver zu machen.

Im Vortrag wollen wir die projektspezifische praxisorientierte Citizen-Science-Strategie im *Itinera Nova*-Projekt und den Anteil des CCeH daran vorstellen und mit der DH-Fachcommunity diskutieren. Insbesondere möchten wir vorstellen, wie die gemeinsame Projektarchitektur mit den weiteren vorgestellten Projekten positive Effekte auch in klassischen DH-Projekten entfaltet.

Fußnoten

1. Transkribus ist ein KI-gestützte Plattform für Texterkennung, Transkription und das Durchsuchen von historischen Dokumenten. Weitere Informationen dazu s. https://readcoop.eu/de/transkribus/ (zugegriffen: 19. Juli 2023).
2. Weitere Informationen zu den Modellen s. https://readcoop.eu/de/glossary/model-training/ (zugegriffen: 18. Juli 2023).
3. CER (Character Error Rate) deutlich unter 10%.

Bibliographie

Bigalke, Jan, Jonathan Blumtritt und Tessa Gengnagel. 2023. „Beginen in Köln: Von der Textdatenbank zur zeitgemäßen digitalen Auszeichnung und Analyse“. In Book of Abstracts der 9. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2023). Trier, Luxemburg: Zenodo. https://doi.org/10.5281/zenodo.7715269.

Böhringer, Letha und Barbara Weber. 2022. „Beginen - Frauengemeinschaften im Mittelalter Interv. Dr. Letha Böhringer geführt von Barabara Weber“. Deutschlandfunk. https://www.deutschlandfunk.de/beginen-frauengemeinschaften-im-mittelalter-interv-dr-letha-boehringer-dlf-ca4a8b47-100.html.

Bonn, Aletta, Wiebke Brink, Susanne Hecker, Thora M. Herrmann, Christin Liedtke, Matthias Premke-Kraus, Silke Voigt-Heucke, et al. 2021. “Weißbuch Citizen Science Strategie 2030 Für Deutschland.” SocArXiv. August 7. doi:10.31235/osf.io/ew4uk.

Eyharabide, Victoria, Béatrice Caseau, Jean-Claude Cheynet, Lucia Orlandi, Qijia Huang, et al. 2023. “Byzantine Sigillography meets Artificial Intelligence: The BHAI Project. Numismatics, Sphragistics and Epigraphy”, In press. hal-03901611.

Rapp, Andrea, Stefan Büdenbender, Nadine Dietz, Lena Dunkelmann, Birte Gnau-Franké, Nina Liesenfeld, Stefan Schmunk, u. a. 2022. „Mein liebster Schatz! Das Citizen Science-Projekt Gruß & Kuss stellt sich vor“. In Book of Abstracts der 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2022). Potsdam: Zenodo. https://doi.org/10.5281/zenodo.6328189.

Smolarski, René (Hg.), Carius Hendrikje (Hg.) und Prell Martin (Hg.). 2023. “Citizen Science in den Geschichtswissenschaften: Methodische Perspektive oder perspektivlose Methode?.“ V&R unipress. DH&CS. 1. Auflage.

Vohland, Katrin (Hg.), Anne Land-Zandstra (Hg.), Luigi Ceccaroni (Hg.), Rob Lemmens (Hg.), Josep Perelló (Hg.), Marisa Ponti (Hg.), Roeland Samson (Hg.) und Katherin Wagenknecht (Hg.). 2021. “The Science Of Citizen Science”. Cham: Springer International Publishing. doi:10.1007/978-3-030-58278-4.

FAIR/CARE Principles as Normative Ethics in Digital Musicology MEI, Metadata, and Minimising Invisible Labour

Neumann, Joshua
joshua.neumann@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0000-0001-7970-5553

Richts-Matthaei, Kristina
Kristina.Richts@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0000-0001-8569-1995

Given the highly collaborative nature of (digital) scholarship, crediting those who contribute to the creation or lifecycle of any project is an ethical imperative. As such, labor visibility is and remains a critical topic, especially as digital scholarship continues to evolve and define itself. To ask questions of musical creative or intellectual outputs, analogue or digital, necessitates an understanding of both their nature and the circumstances of their creation. Filmmaker Hollis Frampton’s assessment of photographic snapshots applies to multi-faceted scholarly undertakings. Each one “purports to be an ideal…wholly static cross section through a four-dimensional solid, or tesseract, of unimaginable intricacy.” (Frampton, 2009, 27) Critical editions, collected works, and other similar projects are such printed tesseracts in musicology. Often in these projects, praise for a managing editors approaches apotheosis as singularly visionary titans. Martin Staehelin’s review of Edwin Lowinsky’s three-volume edition of The Medici Co-

dex for the Journal of the American Musicological Society lauds the editor profoundly.

> At every step it reveals the characteristic expertise and the superior synthetic powers of a scholar intimately familiar with his subject, committed to it in a deeply personal way, and sparing no effort in his musicological work. Without a doubt Edwin Lowinsky has in this critical edition of the Medici Codex lived up to his motto of achieving only the best and,...has…actually achieved the best that present-day research in the music of the Renaissance can hope to attain. (Staehelin, 1980, 587)

In addition to its historiography of the AMS, Staehelin's review highlights that he was acting as an arbiter of quality within a community of scholars, thereby evincing Pierre Bourdieu's stipulation of taste as a cultural construct moderated by gatekeepers who are also often competitor-colleagues. (Bourdieu, 1984) Anyone who engages with artistic/intellectual works operates within **art worlds**, which generate art as entity, process, and experience. (Becker, 1982) These communities enact a myriad of collaborative and individual processes, the absence of any of which alters the final output. Considering their nature and operations as the miniature societies they are is invaluable.

Normative Ethics

Social contractualism thus applies, as it offers tools for assessing and constructing societal normative ethics. T.M. Scanlon's reasonable rejectability is a useful barometer for demarcating right from wrong. "An act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement." (Scanlon, 1998, 153) In Scanlon's work, moreover, "One reason for focusing on wrong is to draw attention to the domain that contractualism is concerned to map, concerning what it is for one person to have been wronged by another." (Ashford and Mulgan, 2018)

For creative or intellectual projects, invisible labour is just such a wrong, even as it has proliferated across disciplines, especially in the space of single author or editor publications. Even in multi-author work, the full accounting of roles of non-authorial contributors is often absent for various reasons, resulting in invisible labour. As such, any accurate and thus honest intellectual history of research outputs is often not possible. One way to a solution adopts John Rawls's **original position** behind the **veil of ignorance**. By not knowing one's potential role in (academic) society, one is more likely to pursue the goal of en-framing justice as Heideggerian Gestell through developing fair governing principles to which everyone agrees. "…Fair terms of cooperation specify an idea of reciprocity or mutuality:..The idea of rational advantage specifies that it is that those engaged in cooperation are seeking to advance from the standpoint of their own good." (Rawls, 2001, 6) Rawls leverages self-interest from the position of ignorance to epitomise justice in the form of fairness to all. (Ashford and Mulgan, 2018)

Both Scanlon and Rawls seek to enframe societies along the lines of fairness and justice where the two concepts co-define each other. The FAIR and CARE principles for data offer a tandem approach for normative ethics of research community governance. FAIR's emphases on metadata under Findable and Accessible are notable. ("FAIR Principles, 2021) Metadata must be rich and accessible even when data they describe might no longer exist. The CARE principles focus more on the rights of people involved in research, both subjects and practitioners. ("CARE Principles," 2019) In many ways, the focus here is on constructive collaboration, namely honest attribution and equitably shared governance. The FAIR principles thus seek to create a technological enframement in which the CARE principles are readily attainable, eventually becoming the norm in research praxis.

Metadata and MEI

Metadata is the appropriate locus for crediting contributions associated with intellectual labour, especially taking the FAIR and CARE principles as normative ethics. The Music Encoding Initiative offers numerous affordances in this regard. Because MEI has become the standard format for high quality musical data and metadata only in the past ten-to-fifteen years, is community-driven, and dedicated to principles of fairness evinced by (and derived from) Rawls and Scanlon, it is also ideally situated for future projects.

One significant tool emergent from the MEI community is the Metadata editor and repository for MEI data – MerMEId, which was developed primarily for use in the Catalogue of Carl Nielsen's Works. (Catalogue of Carl Nielsen's Works, 2018) Here, one can create and enrich metadata for a work's history in terms of people, places, etc., along with an intellectual history of engagements with it. Making these designations possible are the options in the contributor panels in the work and file tabs of the editor. A person's role is fully customisable via the attribute editor, as are other descriptors. Furthermore, linking to authority data sources is readily available for works histories. Adapting Hollis Frampton's analysis of images as both objects in their own right and as semiotic icons or indices to this aesthetic space reveals how this flexibility matters. A musical work or source is "like language, doubly identified: once with itself, and once again with its referent." (Frampton, 2009, 71) Ongoing and planned further developments for the MerMEId aim to make it more adaptable and easier to enrich for project-specific needs.

In addition to role attributes or ORCID IDs as authority data, one can also easily leverage the MEI schema to offer plain text descriptions of contributions to a project in either the file description (Figure 1) or the revision description (Figure 2).

The flexibility and economy of digital metadata, especially in MEI, to make labour more visible reflects one

of Frampton's strongest appreciations of work in photography, where he focused on collections and inventories in a manner similar to constructing both narrowly focused projects or broadly-focused collected or complete editions. "The artist reaffirms his own existence through gradually replacing the space of the given world with the inventory of spaces of all the photographs he has made." (Frampton, 2009, 75) Changing 'artist' for 'researcher' and 'photograph' for 'project' or 'Ausgabe' does not undermine Frampton's point.

Figure 1. <fileDesc> contributor accounting in <meiHead> for contributions at/embedded in file creation (beginning of file).

Figure 2. <change>, <persName>, and <changeDesc> in <revisionDesc> for changes made after file creation (end of file).

Musicological praxis is moving in this direction, albeit slowly, especially in the space of corpus-study based projects such as critical editions and works catalogues. Major canonical composers have heretofore been the overwhelming focal point for such projects, many of which still feature single-editor acknowledgement. Additional contribution acknowledgements almost always reside in a preface or a footnote, namely the data contents, rather than metadata. Mentions in the object data, whether primary or supplementary material, are the norm, and are useful means of crediting contributions from various persons attached to a project. However, locating such mentions solely in the data content creates a credit differential with scholars mentioned in the metadata.

The exclusion of these contributors from the metadata creates an incomplete intellectual history that falls short of FAIR adherence for: 1) Findability's "Rich metadata," 2) Accessibility's "Metadata outliving data," 3) Interoperability's "Qualified references to other data and metadata," and 4) Reusability's "rich and accurate descriptions of data," "detailed provenance," and "domain-relevant community standards." Through understanding benefits and pitfalls of different eras of project inception and variable external guidelines (expectations of funding agencies), and learning from them, envisaging and working toward a more equitable operational model is possible.

Ways Forward

Obstacles persist and are intertwined with ongoing misunderstandings about the nature of digital tools and infrastructures—namely that, for many, digital work exists as a kind of machina-ex-deus. As such, and because it is not always as integral to musicological projects, it can often seem unworthy of the credit it is due. (Luhmann and Burghardt, 2021, 149) Community labour, especially digital, either fades or is pushed to the background too easily, creating an opportunity, if not obligation, for education about the nature of digital work and the rigour and creativity required to engage it.

Even as the digitality of some projects enhances the presence of FAIR and CARE principles in labour acknowledgement, the reality is that until the thinking of disciplinary gatekeepers evolves, the political, social, and disciplinary capital of print editions reigns supreme. As such, teaching and research activities are often negatively affected, especially in the case of tenure and promotion processes where review committees in music need guidance on evaluating digital work, and the guidance that exists needs more frequent updating than that for analogue work. These guidelines also, in turn, affect training opportunities for students, and the nature of capstone projects.

Digital scholars (including musicologists) do well, then, to adopt Hans Van Maanen's consideration of art worlds with foci on equal access and opportunity while emphasising the need to decentralise cultural power, i.e. – gatekeeping. (Maanen, 2010, 238) What is true as this discussion focuses on art making processes that perhaps centre aesthetic performance, remains true if one shifts the focus to research activity, whether analogue or digital. Rawls's difference principle, which dictates that the most fair and just distribution is the one where the least-advantaged group does best, is a useful metric. Adapting van Maanen's considerations of interactions between art worlds and their social-political contexts to the scope of (digital) musicology, especially with the need to educate academic gatekeepers might results in a three-fold approach:

1. Recognizing music scholarship as aesthetic and scholarly communication. This recognition in turn requires guaranteeing access to contributor status, within appropriate guidelines for academic rigour and technological engagement.

2. Shifting paradigms toward greater FAIR/CARE inclusion as fundamental principles via ethical and technological education. Beyond the inherent need for technological education in the making phase consumption skills (how to interact with and value) digital editions that are FAIR/CARE adherent is also necessary.

3. Continuing to decentralise the loci of cultural, educational, and publishing power away from the rapidly-becoming-antiquated model of for-profit print publishing by committing to open science, open source, and open access editions, at least as much as is feasible.

This paper thus advocates for increasing access to engaging scholarly praxis, and for receiving fair recognition in the metadata, while enhancing standards of scholarly rigour. Rather than suggesting a unilateral equivalency of attribution among contributors, since not all contributions are the same, it proposes that anyone who contributes intellectual labour deserves credit for their work. MEI's affordances for documenting metadata in musicological work make this credit easily giveable and customizable with only a few lines of code.

To be fully FAIR, any edition's data must be as rich and accurate as possible. Additionally, its metadata must be so, especially as it concerns recognizing the art world of its creation. Only then can a community begin to engage in the formal training necessary to contribute high quality data that is both interoperable and reusable at a technical level. These collaborative communities exist already, though to varying degrees of visibility. FAIR-ifying their metadata makes them more visible and, in so doing, also makes the normative ethics of FAIR and CARE more standard praxis than thought experiment. This kind of scholarly model, based on Becker's Art Worlds and infused with Rawls's and Scanlon's senses of social contractualism, exhorts ongoing assessment and adjustment of praxes—in musicological work, training, and pedagogy. Historiographical efforts thus also merit reconsideration. In the context of music scholarship, authors, editors, and all contributors are inextricable from their context. Leveraging MEI's metadata affordances for detailed and accurate intellectual histories of these communities and the works they engage enables more visibility for labour within them, an undergirding tenet of the FAIR and CARE principles. In sum, then, this paper offers perhaps no sweeping paradigm shift, but instead a few small steps toward a more just, community-based, and digital model of scholarly work in musicology.

Bibliographie

Ashford, Elizabeth, and Tim Mulgan. "Contractualism." Stanford Encyclopedia of Philosophy, April 20, 2018. https://plato.stanford.edu/entries/contractualism/#Bib.

Becker, Howard Saul. Art Worlds / Howard S. Becker. Berkeley, CA: University of California Press, 1982.

Bourdieu, Pierre. Distinction: A social critique of the judgement of taste. Routledge & Kegan Paul, 1984.

"Care Principles." Global Indigenous Data Alliance, September 2019. https://www.gida-global.org/care.

"Fair Principles." GO FAIR, January 21, 2022. https://www.go-fair.org/fair-principles/.

Foltmann, Niels Bo, Axel Teich Geertinger, Peter Hauge, Niels Krabbe, Bjarke Moe, and Elly Bruunshuus Petersen, eds. Catalogue of Carl Nielsen's Works, 2018. https://www.kb.dk/dcm/cnw/navigation.xq.

Frampton, Hollis, and Bruce Jenkins. On the Camera Arts and consecutive matters: The writings of Hollis Frampton. Cambridge, MA: The MIT Press, 2009.

Luhmann, Jan, and Manuel Burghardt. "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 (2021): 148–71. https://doi.org/10.1002/asi.24533.

Maanen, Hans van. How to study art worlds: On the societal functioning of aesthetic values. Amsterdam: Amsterdam University Press, 2010.

Rawls, John, and Erin Kelly. Justice as fairness: A restatement. Cambridge, MA: Belknap Press of Harvard University Press, 2001.

Scanlon, T. M. What we owe to each other. Cambridge, MA: Harvard University Press, 1998.

Staehelin, Martin. "Review: The Medici Codex of 1518. A Choirbook of Motets Dedicated to Lorenzo de'Medici, Duke of Urbino by Edward E. Lowinsky." Journal of the American Musicological Society 33, no. 3 (1980): 575–87. https://doi.org/10.2307/831307.

Fanfictions – Literatur von Frauen über Männer? Korpusbasierte Analyse der Geschlechterrollen bei Texten und Autor*innen deutschsprachiger Fanfictions

Schmidt, Thomas
thomas.schmidt@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg, Deutschland
ORCID: 0000-0001-7171-8106

Sasse, Jonathan
jonathan.sasse@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg, Deutschland

Wolff, Christian
christian.wolff@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg, Deutschland
ORCID: 0000-0001-7278-8595

Einleitung

Fanfictions sind literarische Texte, erstellt von Fans und Hobby-Autor*innen, die Figuren und Geschichten aus bereits bestehenden Medien wie Filmen oder Büchern nutzen, um neue Geschichten über diese zu schreiben und auf Online-Plattformen zu veröffentlichen (Dym et al., 2018). Dieses spezielle literarische Genre wurde in den letzten Dekaden mit dem Aufstieg des Internets immer populärer und deswegen auch vielseitig in den Geistes- und Kulturwissenschaften in Hinblick auf Geschichte und kulturellen Einfluss untersucht (siehe Hellekson und Busse, 2006; Jamison, 2013). Die Verfügbarkeit von großen narrativen Textmengen mit detaillierten Metadaten macht Fanfictions auch zu einer beliebten Quelle für verschiedene Aufgaben im *Natural Language Processing* (NLP) (z.B. Muttenthaler et al., 2019; Kim und Klinger, 2019; Zhang et al., 2019). In den Digital Humanities werden mit computergestützten Methoden Fragen wie der Einfluss sozialer Aspekte auf Sprache (Frens et al., 2018; Rebora et al., 2021), Intertextualität (Milli und Bamman, 2016; Kleindienst und Schmidt, 2020; Cuntz-Leng et al., 2023), kulturelle Entwicklung (Pianzola et al., 2020), nationale Besonderheiten von Fanfictions (Schmidt et al., 2021b), Fan-Kultur (Yin et al., 2017; Kleindienst et al., 2022), Autor*innen und Leser*innen-Netzwerke (Carvallo und Parra, 2020) sowie Charakter-Netzwerke (Schmidt et al., 2022) untersucht.

Geschlechtsspezifische Fragestellungen spielen eine wichtige Rolle im Kontext von Fanfictions. Bisherige Analysen für mehrheitlich englischsprachige Texte deuten auf eine erhöhte weibliche Autorschaft in diesem Genre hin (Barnes, 2015; Duggan, 2020). Motivation und Bedeutung für die Popularität von Slash-Fanfictions (Fanfictions mit Fokus auf homo-romantischen Beziehungen zwischen Männern) und damit die Dominanz von männlichen und Vernachlässigung weiblicher Figuren wurden vielfach anhand begrenzter Mengen von Texten diskutiert (Jung, 2002; Tosenberger, 2008; Rossdal, 2015; Busse und Lothian, 2017). Dem entgegen argumentiert andere Forschung mit ähnlichem methodischem Zugang („close reading“, vgl. Busse, 2009; Leow, 2011; Handley, 2012; Duggan, 2017; 2020; 2022), dass die Autorschaft wesentlich diverser ist und weibliche Charaktere eine wichtige und nicht-stereotype Rolle spielen. Derartige Analysen werden computergestützt in größeren Rahmen auch von den Untersuchungen von Milli und Bamman (2016) getragen, während Fast et al. (2016) eine stereotype und negative Repräsentation von weiblichen Figuren identifizieren.

Wir präsentieren im Folgenden die Ergebnisse eines Projekts, dass die bisher vorliegenden computergestützten Analysen mit Fokus auf den deutschsprachigen Bereich weiterführt. Unsere Forschungsbeiträge sind (1) die Akquise und Bereitstellung eines strukturierten Korpus speziell für die Analyse deutschsprachiger Texte und Communites, (2) allgemeine Korpus- und Metadatenanalysen und (3) erste Analysen zur Verteilung von Geschlechtern bezüglich Figuren und Autor*innen in diesem Korpus.

Korpusakquise

Als Plattformen für die Korpusakquise wurden Fanfiktion.de (FF.de)[1] und Archive of Our Own (AO3)[2] gewählt. FF.de gilt als die größte deutschsprachige Fanfiction-Plattform und AO3 als eine der populärsten internationalen Plattformen, die allerdings nur eine verhältnismäßig kleine Menge deutschsprachiger Inhalte aufweist. Dennoch ist AO3 eine bedeutende Plattform für die Fanfictions-Forschung und wir planen in Zukunft die Akquise von weiterem deutschsprachigem Material für vergleichsbasierte Analysen.

Die Inhalte beider Plattformen (Texte, Metadaten, Kommentare/Reviews, Nutzer*innen-Profile) wurden mittels *Scrapy*[3] , einem Python-Framework für Web-Crawling im August 2022 akquiriert (im Fall von AO3 zusätzlich mit Hilfe von Metadaten-Filter auf den deutschsprachigen Inhalt reduziert). Die relevanten Inhalte der HTML-Seiten wurden in das JSON-Format umgewandelt und in einer *MongoDB*-Instanz gesichert.[4] Die grundlegende Repräsentation von Fanfictions auf beiden Plattformen ist ähnlich: Fanfictions setzen sich aus Kapiteln, ihren Texten sowie Metadaten zusammen. Einige Metadaten beider Plattformen sind identisch und lassen sich direkt aufeinander abbilden. Der Umgang mit nicht-kongruenten Metadaten wird im Folgendem explizit angesprochen. Für die nachfolgende Korpusanalyse wurden Fanfictions ausgesondert, die entweder keinen oder nur vernachlässigbar wenig Inhalt enthielten sowie solche, die ausschließlich aus Bildern oder URLs bestanden.

Allgemeine Korpusanalyse

Tabelle 1 illustriert die allgemeinen Statistiken des Korpus in Summe und aufgeteilt nach Plattform. Mit 394.848 einzelnen Fanfictions liefert FF.de im Vergleich zu AO3 (18.075) deutlich mehr deutschsprachigen Inhalt. In beiden Plattformen muss man sich mittels eindeutigem Nutzernamen anmelden, um Texte zu posten oder damit in Kommentaren und Bewertungen zu interagieren – hierauf beziehen sich die Nutzerstatistiken. Als Reviews bezeichnen wir im Folgenden Kommentare. Token-Statistiken wurden mittels *Spacy*[5] berechnet. Im Durchschnitt besteht eine Fanfiction aus 8.487 Tokens, die längste Fanfiction hat eine Länge von 4.106.713 Tokens.

Plattform	Fanfictions	Kapitel	Tokens	Nutzer	Reviews
FF.de	394.848	1.885.066	3.351.074.976	135.726	4.849.646
AO3	18.075	70.857	153.402.525	14.249	37.721
Total	412.923	1.955.923	3.504.477.501	149.975	4.887.367

Tabelle 1: Allgemeine Korpusstatistiken.

Als Fandom wird die mediale Referenz, also das fiktionale Universum bzw. Thematik bezeichnet, in dem eine Fanfiction spielt. Tabelle 2 und 3 illustrieren die Top 10 Fandoms für FF.de und AO3 respektive.

Fanfiktion.de

Häufigkeit	Fandom	Anteil
54.405	Harry Potter	13,78 %
38.568	Musik	9,77 %
27.303	Naruto	6,91 %
17.830	Internet-Stars	4,52 %
13.734	Twilight	3,48 %
11.443	One Piece	2,9 %
11.440	Sport	2,9 %
8.199	J.R.R. Tolkien	2,08 %
6.045	Supernatural	1,53 %
5.311	Marvel	1,35 %

Tabelle 2: Top 10 Fandoms in Fanfiktion.de.

Archive of Our Own

Häufigkeit	Fandom	Anteil
2.978	Tatort	16,48 %
1.793	Harry Potter	9,92 %
1.234	Marvel	6,83 %
671	Kanon	3,71 %
663	Sport	3,67 %
654	Musik	3,62 %
509	The Three Investigators - Die drei ???	2,82 %
506	Supernatural	2,8 %
493	Sherlock BBC	2,73 %
487	J.R.R. Tolkien	2,69 %

Tabelle 3: Top 10 Fandoms in Archive of our Own.

Die grundsätzlichen Fandom-Verteilungen verhalten sich konform zu Analysen auf größeren englischsprachigen Plattformen mit Fandoms wie Harry Potter, Marvel und Supernatural als besonders häufigen Fandoms. Im Fall von FF.de wird die besondere historische Bedeutung von Anime (Naruto, One Piece) für die deutsche Fanfiction-Community deutlich (siehe auch Cuntz-Leng und Meintzinger, 2015). Für AO3 kristallisieren sich spezielle nationale Besonderheiten heraus wie die Häufigkeit von Tatort-, Die drei ???[6]- sowie Sport- (vor allem Fußball-) Fanfictions. Auf geschlechtsspezifischer Ebene dominieren Fandoms mit mehrheitlich männlichen Hauptcharakteren.

Ein weiteres wichtiges Metadatum im Kontext dieser Forschung sind Beziehungstypen, die für beide Plattformen äquivalent vorliegen. Dadurch wird markiert, ob eine romantische/erotische Beziehung zwischen Figuren eine wichtige Rolle spielt und welcher Geschlechternatur diese ist. Tabelle 4 zeigt die kumulierte Verteilung für beide Plattformen. Dabei ist zu beachten, dass eine Fanfiction im Fall von AO3 auch mehrere Angaben bezüglich Beziehungstypen haben kann.

Beziehungstyp	Häufigkeit/Anteil (AO3)	Häufigkeit/Anteil (FF.de)	Häufigkeit/Anteil (gesamt)
Generisch	3.572 (17,5%)	272.616 (69,0%)	276.188 (66,5%)
M/M	10.129 (49,5%)	106.922 (27,1%)	117.051 (28,2%)
W/M	3.668 (17,9%)	11.116 (2,8%)	14.784 (3,6%)
W/W	953 (4,7%)	1.873 (0,5%)	2.826 (0,7%)
Multi	648 (3,2%)	2.054 (0,5%)	2.702 (0,7%)
N/A	1.059 (5,2%)	0 (0%)	1.059 (0,3%)
Diverse	426 (2,1%)	267 (0,1%)	693 (0,2%)

Tabelle 4: Beziehungstypen-Verteilungen auf FF.de und AO3 (M = Männlich, W = Weiblich). Prozentzahlen sind gerundet und beziehen sich auf den jeweiligen Sub-Korpus.

Obwohl der Großteil der Geschichten als Generisch (Generic) (66,5%) gekennzeichnet ist und damit definitionsgemäß keinen spezifischen Beziehungstyp fokussiert, spielen Slash-Fanfictions (M/M) eine bedeutende Rolle, da sie den größten Teil der verbleibenden Fanfictions ausmachen. Im Vergleich dazu sind Beziehungen, welche weibliche Figuren beinhalten, eher selten.

Geschlechtsbasierte Nutzer*innen- und Figuren-Analyse

Methodik

Zur vertieften Analyse wurde eine Geschlechtsklassifikation genannter Personennamen in den Fanfictions durchgeführt. Dazu wurde zunächst eine *Named Entity Recognition* (NER; Eigennamen-Erkennung) zur Erkennung von Personennamen in den Texten durchgeführt. Die NER wurde mit FLAIR[7] und dem vortrainierten Modell *ner-multi-fast*[8] implementiert (Akbik et al., 2018; 2019). Das Modell ist multilingual und wurde auf vier verschiedenen Korpora (für jede Sprache) trainiert (Englisch, Deutsch, Holländisch, Spanisch). Stichprobenartige Analysen zeigten, dass dieses Modell die besten Ergebnisse und Effektivität im Vergleich zu anderen FLAIR-NER-Modellen erzeugt. Experimentelle Vergleiche mit transformerbasierten Sprachmodellen zeigten keine wesentlichen Verbesserungen aber erhöhte Performanzprobleme aufgrund der Größe unseres Datensatzes, was auch ein Grund für die Entscheidung dieses Modells war. Die Häufigkeiten von Personenerkennungen in den Texten (andere NER-Entitäten wurden hierbei vernachlässigt) wurden im JSON-Format gespeichert und dem Korpus hinzugefügt.

Nach ersten Experimenten mit vortrainierten Modellen zur geschlechtsbasierten Namenserkennung[9] haben wir uns entschlossen ein eigenes Modell zur Namenserkennung zu trainieren. Grund hierfür waren wiederum Performanzprobleme und die Identifikation von größeren Problemen vortrainierter Modelle mit fiktionalen Namen aus Fantasy und Science-Fiction. Es wurde also ein eigenes Modell basierend auf folgenden Datensätzen aus Namen und Geschlechtsangaben trainiert:

- *NLTK name corpus*[10]
- Ein öffentlicher Datensatz von Namen von Data.gov[11]
- Eine öffentliche Liste von Babynamen von babynames.com

Insbesondere die Liste von babynames.com hat einen besonderen Mehrwert da hier fiktionale Namen aus Kunst und Kultur enthalten sind.[12] Die Datensätze wurden verknüpft, Duplikate entfernt und auf eindeutige männliche und weibliche Namen reduziert. Der finale Trainingsdatensatz besteht aus 106.000 Namen-Geschlechts-Angaben und wurde in einem 5x5 Kreuzvalidierungssetting mittels *Keras*[13] und *TensorFlow* in einem neuronalen Netz mit LSTM-Architektur genutzt, um das Klassifikationsmodell zu trainieren. Das finale Modell wurde nach Analyse von Hyperparameter-Experimenten mit Adam-Optimizer, einer Batch-Größe von 64 und für 16 Epochen trainiert und erreicht eine durchschnittliche Erkennungsgenauigkeit von 90%. Dieses Modell wurde auf die zuvor erkannten Eigennamen und ihre Verteilungen angewandt. Das Modell ist nur auf die Erkennung von geschlechtsidentifizierenden Namen trainiert (also keine Koreferenzen oder Pronomen).

Für Analysen, die das Autor*innen-Geschlecht verwenden, wird das Korpus auf den FF.de-Anteil beschränkt, da nur in diesem Nutzer*innen über ihr Profil freiwillige Geschlechtsangaben machen können. Eine Geschlechtserkennung auf Nutzernamen ist aufgrund ihrer Beliebigkeit and Abstraktheit in diesem Kontext nicht sinnvoll.

Ergebnisse

Tabelle 5 zeigt das Verhältnis von männlichen zu weiblichen Namen bezüglich der fiktionalen Charaktere in den Fanfictions auf. Diejenigen Namen, die keine Erkennungssicherheit von mindestens 80% erreicht haben, wurden als unsicher markiert.[14]

Plattform	Männlich	Weiblich	Unsicher
FF.de	36.000.856 (60,85%)	19.471.225 (32,91%)	3.695.846 (6,25%)
AO3	579.925 (71,63%)	189.421 (23,40%)	40.212 (4,97%)
Total	36.580.781 (60,99%)	19.660.646 (32,78%)	3.736.058 (6,23%)

Tabelle 5: Verteilung von männlichen und weiblichen Eigennamen.

Insgesamt zeigt sich, dass die Nennungen von männlichen Eigennamen überwiegen, im Schnitt in einem Verhältnis von 61% zu 33% mit ca. 6% Namen, die nicht eindeutig klassifiziert werden konnten. In Tabelle 6 wird die Verteilung der freiwilligen Selbstangaben von Nutzer*innen auf FF.de aufgezeigt, wobei eine Gesamtübersicht sowie eine Unterteilung nach Autor*innen und Reviewer*innen gegeben ist. Unter letzteren werden die Geschlechtsangaben der Poster*innen von Reviews/Kommentaren verstanden. Jede Autor*in und jede Reviewer*in werden dabei einmal gezählt, unabhängig von der Zahl der veröffentlichten Geschichten oder Reviews. Es ist bei der Interpretation der Zahlen zu beachten, dass viele Autor*innen gleichzeitig auch als Reviewer*innen aktiv sind und diese beiden Kategorien Duplikate enthalten, die Gesamt-Information bezieht sich aber auf die tatsächliche Gesamtzahl aller eindeutig differenzierbaren Nutzer*innen.

Geschlecht	Gesamt	Autor*innen	Reviewer*innen	Alter
weiblich	87.784 (64,68%)	72.959 (68,04%)	51.084 (67,89%)	26,89
männlich	7.834 (5,77%)	6.291 (5,87%)	4.521 (6,01%)	27,98
diverse	671 (0,49%)	511 (0,48%)	450 (0,60%)	23,12
N/A	39.437 (29,06%)	27.463 (25,61%)	19.185 (25,50%)	27,10
Total	135.726	107.224	75.240	26,97

Tabelle 6. Demographie-Statistik für FF.de.

Ein hoher Teil der Nutzer*innen gibt kein Geschlecht an (29%). Bezogen auf Nutzer*innen, die ein Geschlecht angeben, zeigt sich jedoch eine deutliche Dominanz von weiblichen Personen. Abstrahiert man von den Nicht-Angaben (N/A), ist das Verhältnis sogar ca. 92% zu 8%. Es gibt keinen wesentlichen Unterschied beim Vergleich von Autor*innen und Reviewer*innen. Die Plattform wird basierend auf Selbstauskunft also primär von weiblichen Personen genutzt. Die Altersinformationen dienen lediglich der demographischen Vertiefung und sind nicht Fokus dieses Beitrags. Sie zeigen aber eine durchschnittlich eher junge Nutzer*innen-Gruppe auf (etwa 27 Jahre).

In Tabelle 7 werden die beiden Analysen in ein Verhältnis zueinander gesetzt und der Anteil weiblicher und männlicher Figurennamen in den einzelnen Autor*innen-Geschlechtsgruppen untersucht. Es zeigt sich kein wesentlicher Unterschied im Vergleich zu männlichen und weiblichen Autor*innen in der Nutzung von weiblichen oder männlichen Figurennamen. Weibliche Autor*innen nutzen männliche Figuren in einem Verhältnis von 64% zu 36% weiblichen Figuren. Der Anteil von weiblichen Figuren verringert sich lediglich um 1% für männliche Autoren.

Autor*innen-Geschlecht	Weibliche Figurennamen	Männliche Figurennamen
weiblich	35,82%	64,18%
männlich	34,52%	65,48%
sonstige	31,03%	68,97%
N/A	31,85%	68,15%

Tabelle 7. Anteil weiblicher und männlicher Figurennamen in den einzelnen Geschlechtsgruppen.

Diskussion

In diesem Beitrag wurden die Ergebnisse der Korpusakquise einer Sammlung von deutschsprachigen Fanfictions aufgezeigt. Es ist zu beachten, dass wir dabei einige wesentliche Bestandteile noch nicht vertieft präsentieren konnten, wie z.B. weitere Metadaten und Review-Analysen. Das Korpus ist eine relevante Ressource für den Bereich der *Computational Literary Studies* und *Fan Studies*, das noch mit zahlreichen Methoden exploriert werden kann. Wir beabsichtigen, insbesondere explorative Netzwerkanalysen, lexikalische Frequenzanalysen und Sentiment/Emotionsanalysen in Analogie zu anderen DH-Projekten anzuwenden (Schmidt et al., 2020a; Moßburger et al., 2020; Schmidt et al., 2020c; Dennerlein et al., 2023; Schmidt et al., 2023). Folgende Limitationen sind zu adressieren: So ist die hier gewählte NER und Geschlechtserkennung eine allgemeine Lösung, die nicht auf den Anwendungsfall hin optimiert wurde. Durch Annotationsstudien mit einer Auswahl von Texten sowie anschließender Anwendung von Machine Learning soll sich diesem Problem angenähert werden. Auch basieren die Geschlechtsstatistiken aus Selbstangaben und ca. 30% der Angaben fehlen. Die Daten zu dieser Analyse basieren auf lediglich einer Plattform (FF.de), was die Aussagekraft dieser Statistiken limitiert.

Dennoch konnten bereits durch allgemeine Metadatenanalysen nationale Besonderheiten eines deutschsprachigen Fanfiction-Korpus wie z.B. die Bedeutung von Fandoms wie Tatort oder Die drei ??? in AO3 sowie die Bedeutung von Anime in FF.de (Cuntz-Leng und Meintzinger, 2015) herausgearbeitet werden. Dies verdeutlicht die Notwendigkeit der Analyse nicht-englischer Texte nicht nur für die lokalen Wissenschafts-Communities, sondern auch für ein angemessenes Verständnis des Phänomens an sich. Im Kontext der geschlechtsspezifischen Fragestellungen konnten Analysen und Behauptungen, wonach Frauen Fanfictions nutzen, um Geschichten über unterrepräsentierte weibliche Figuren zu schreiben (Busse, 2009; Leow, 2011; Handley, 2012; Duggan, 2017; 2020; 2022; Milli und Bamman, 2016) nicht bestätigt werden. Im Gegensatz bestätigen sich bisherige Annahmen (Jung, 2002; Busse und Lothian, 2017; Tosenberger, 2008; Rossdal, 2015; Fast et al., 2016), die Fanfictions als Literatur von Frauen primär über männliche Figuren mit Fokus auf homo-romantischen Beziehungen verstehen auch für die deutschsprachige Fanfiction-Community. Auch eine Reduktion auf das Harry Potter-Fandom analog zu Duggan (2017; 2020; 2022) zeigt dieselben Verhältnisse auf. Es sei hier jedoch auch zu beachten, dass dieses Phänomen auch als Spiegelung der allgemeinen Überrepräsentation von Männern in kulturellen Medien betrachtet werden kann was jedoch bisher in mehrheitlich qualitativen Studien untersucht wurde (Collins, 2011; Bretthauer et al., 2007; Garcia et al., 2015; Jia et al., 2015; Neville und Anastasio, 2019; Schmidt et al., 2020b). Wir halten es auch für eine sehr spannende Idee, die hier vorliegende binäre Geschlechtsauffassung durch weitere Geschlechtsgruppen wie non-binär oder androgyn zu erweitern, wie dies teilweise schon in Computer Vision-Projekten in den DH gemacht wurde (Schmidt et al. 2021a; Schmidt und Kurek, 2022). Annotation und Akquise von non-binären Namen wäre hier für weitere Studien notwendig. Insgesamt ist mehr Forschung die *Close Reading*, computergestützte Analysen und empirische Forschung in einem Mixed-Methods-Ansatz verknüpft, für ein vertieftes Verständnis von Fanfictions erforderlich. Das Korpus wird für die Forschungs-Community auf Anfrage zur Verfügung gestellt und soll dadurch dem Forschungsfeldern der *Fan Studies*, *Internet Studies* und *Computational Literary Studies* neue Impulse liefern.[15]

Fußnoten

1. https://www.fanfiktion.de/
2. https://archiveofourown.org/
3. https://scrapy.org/
4. https://www.mongodb.com
5. https://spacy.io
6. Obschon die „Die drei ???“-Bücher ihren Ursprung in den USA haben, wurden Sie aufgrund der Popularität speziell in Deutschland weitergeführt.
7. https://github.com/flairNLP
8. https://huggingface.co/flair/ner-multi-fast
9. z.B. https://pypi.org/project/gender-guesser/
10. https://www.nltk.org/howto/corpus.html
11. https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national- data
12. Siehe z.B. https://babynames.com/names/fictional-character-names.php , https://babynames.com/names/water-names.php
13. https://keras.io/
14. Die Wahrscheinlichkeit würde über die Sigmoidfunktion im Output-Layer berechnet.
15. Das Korpus ist auf Anfrage an thomas.schmidt@ur.de erhältlich.

Bibliographie

Akbik, Alan, Tanja Bergmann und Roland Vollgraf. 2019. “Multilingual Sequence Labeling with One Model.” In *NLDL 2019, Northern Lights Deep Learning Workshop*.

Akbik, Alan, Duncan Blythe und Roland Vollgraf. 2018. “Contextual String Embeddings for Sequence Labeling.” In *Proceedings of the 27th International Conference on Computational Linguistics*, herausgegeben von Emily M. Bender, Leon Derczynski, and Pierre Isabelle, 1638–1649. Santa Fe, New Mexico, USA: Association for Computational Linguistics. https://aclanthology.org/C18-1139.

Barnes, Jennifer L. 2015. “Fanfiction as Imaginary Play: What Fan-Written Stories Can Tell Us about the Cognitive Science of Fiction.” *Poetics* 48 (February): 69–82. https://doi.org/10.1016/j.poetic.2014.12.004.

Bretthauer, Brook, Toni Schindler Zimmerman und James H. Banning. 2007. “A Feminist Analysis of Popular Music.” *Journal of Feminist Family Therapy* 18 (4): 29–51. https://doi.org/10.1300/J086v18n04_02.

Busse, Kristina. 2009. “In Focus: Fandom and Feminism: Gender and the Politics of Fan Production: Introduction.” *Cinema Journal* 48 (4): 104–107.

Busse, Kristina und Alexis Lothian. 2017. “Debating Queer Sex, Gay Politics and Media Fan Cultures.” *The Routledge Companion to Media, Sex and Sexuality*.

Carvallo, Andrés und Denis Parra. 2020. “Analyzing Network Effects on a Fanfiction Community.” *ArXiv:1909.02886 [Cs]*, August. http://arxiv.org/abs/1909.02886.

Collins, Rebecca L. 2011. “Content Analysis of Gender Roles in Media: Where Are We Now and Where Should We Go?” *Sex Roles* 64 (3): 290–298. https://doi.org/10.1007/s11199-010-9929-5.

Cuntz-Leng, Vera und Jacqueline Meintzinger. 2014. “A Brief History of Fan Fiction in Germany.” *Transformative Works and Cultures* 19 (July). https://doi.org/10.3983/twc.2015.0630.

Cuntz-Leng, Vera, Christof Leng, and Michael Ellsworth. 2023. “Lord of the Words.” *Zeitschrift für Literaturwissenschaft und Linguistik* 53 (3): 711–728. https://doi.org/10.1007/s41244-023-00312-3.

Dennerlein, Katrin, Thomas Schmidt, and Christian Wolff. 2023. “Computational Emotion Classification for Genre Corpora of German Tragedies and Comedies from 17th to Early 19th Century.” *Digital Scholarship in the Humanities*, fqad046. https://doi.org/10.1093/llc/fqad046.

Duggan, Jennifer. 2017. “Revising Hegemonic Masculinity: Homosexuality, Masculinity, and Youth-Authored Harry Potter Fanfiction.” *Bookbird: A Journal of International Children’s Literature* 55 (2): 38–45. https://doi.org/10.1353/bkb.2017.0022.

Duggan, Jennifer. 2020. “Who Writes Harry Potter Fan Fiction? Passionate Detachment, ‘Zooming out,’ and Fan Fiction Paratexts on AO3.” *Transformative Works and Cultures* 34 (September). https://doi.org/10.3983/twc.2020.1863.

Duggan, Jennifer. 2022. “‘Worlds. . .[of] Contingent Possibilities’: Genderqueer and Trans Adolescents Reading Fan Fiction.” *Television & New Media* 23 (7): 703–720. https://doi.org/10.1177/15274764211016305.

Dym, Brianna, Cecilia Aragon, Julia Bullard, Ruby Davis und Casey Fiesler. 2018. “Online Fandom: Boldly Going Where Few CSCW Researchers Have Gone Before.” In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 121–124. Jersey City NJ USA: ACM. https://doi.org/10.1145/3272973.3274542.

Frens, Jenna, Ruby Davis, Jihyun Lee, Diana Zhang und Cecilia Aragon. 2019. “Reviews Matter: How Distributed Mentoring Predicts Lexical Diversity on Fanfiction.Net.” *arXiv:1809.10268v3 [Cs.SI]*. https://doi.org/10.1184/R1/7793804.v1.

Garcia, David, Ingmar Weber und Venkata Garimella. 2014. “Gender Asymmetries in Reality and Fiction: The Bechdel Test of Social Media.” *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 131–140. https://doi.org/10.1609/icwsm.v8i1.14522.

Handley, Christine. 2012. “‘Distressing Damsels’: Narrative Critique and Reinterpretation in Star Wars Fanfiction.” *Fan Culture: Theory/Practice, Newcastle upon Tyne: Cambridge Scholars Publishing*, 97–118.

Hellekson, Karen und Kristina Busse. 2006. *Fan Fiction and Fan Communities in the Age of the Internet: New Essays*. McFarland.

Jamison, Anne. 2013. *Fic: Why Fanfiction Is Taking Over the World*. Dallas, Texas: Smart Pop.

Jia, Sen, Thomas Lansdall-Welfare und Nello Cristianini. 2015. "Measuring Gender Bias in News Images." In *Proceedings of the 24th International Conference on World Wide Web*, 893–898. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2740908.2742007.

Jung, Susanne. 2002. "Queering Popular Culture: Female Spectators and the Appeal of Writing Slash Fan Fiction." In *Gender Forum*, 2:30–50.

Kim, Evgeny und Roman Klinger. 2019. "Frowning Frodo, Wincing Leia, and a Seriously Great Friendship: Learning to Classify Emotional Relationships of Fictional Characters." 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)*, 647–653. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1067.

Kleindienst, Nina und Thomas Schmidt. 2020. "Investigating the Transformation of Original Work by the Online Fan Fiction Community: A Case Study for Supernatural." In *Digital Practices. Reading, Writing and Evaluation on the Web*. Basel, Switzerland. https://epub.uni-regensburg.de/50828/.

Kleindienst, Nina, Thomas Schmidt und Christian Wolff. 2022. "Analysis and Exploration of Supernatural Fanfictions from the Platform Archive of Our Own." In *Responding to Asian Diversity. Digital Humanities 2022 Conference Abstracts.*, herausgegeben von DH2022 Local Organizing Committee, 649–653. Tokyo, Japan: Alliance of Digital Humanities Organizations (ADHO).

Leow, Hui Min Annabeth. 2011. "Subverting the Canon in Feminist Fan Fiction." *Transformative Works and Cultures* 7. https://doi.org/10.3983/twc.2011.0286

Milli, Smitha und David Bamman. 2016. "Beyond Canonical Texts: A Computational Analysis of Fanfiction." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, herausgegeben von Jian Su, Kevin Duh, and Xavier Carreras, 2048–2053. Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1218.

Moßburger, Luis, Felix Wende, Kay Brinkmann und Thomas Schmidt. 2020. "Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum." In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, herausgegeben von Graciela Gonzalez-Hernandez, Ari Z. Klein, Ivan Flores, Davy Weissenbacher, Arjun Magge, Karen O'Connor, Abeed Sarker, et al., 70–81. Barcelona, Spain (Online): Association for Computational Linguistics. https://aclanthology.org/2020.smm4h-1.11.

Muttenthaler, Lukas, Gordon Lucas und Janek Amann. 2019. "Authorship Attribution in Fan-Fictional Texts given Variable Length Character and Word N-Grams.". In Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. Lugano, Switzerland. https://ceur-ws.org/Vol-2380/paper_49.pdf

Neville, Conor und Phyllis Anastasio. 2019. "Fewer, Younger, but Increasingly Powerful: How Portrayals of Women, Age, and Power Have Changed from 2002 to 2016 in the 50 Top-Grossing U.S. Films." *Sex Roles* 80 (7): 503–514. https://doi.org/10.1007/s11199-018-0945-1.

Pianzola, Federico, Simone Rebora und Gerhard Lauer. 2020. "Wattpad as a Resource for Literary Studies. Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins." *PLOS ONE* 15 (1): e0226708. https://doi.org/10.1371/journal.pone.0226708.

Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J Berenike Herrmann, Maria Kraxenberger, Moniek M Kuijpers, et al. 2021. "Digital Humanities and Digital Social Reading." *Digital Scholarship in the Humanities* 36 (Supplement_2): ii230–50. https://doi.org/10.1093/llc/fqab020.

Rossdal, Maria. 2015. "All of the Greek and Roman Classics. Antikerezeption in Fanfiction." *thersites. Journal for Transcultural Presences & Diachronic Identities from Antiquity to Date* 1. https://doi.org/10.34679/thersites.vol1.5.

Schmidt, Thomas, Marlene Bauer, Florian Habler, Hannes Heuberger, Florian Pilsl und Christian Wolff. 2020a. "Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte." In *DHd 2020: Spielräume; Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts.*, herausgegeben von Christof Schöch und Patrick Helling, 296–300. Paderborn, Germany. https://doi.org/10.5281/zenodo.4621928.

Schmidt, Thomas, Isabella Engl, Juliane Herzog und Lisa Judisch. 2020b. "Towards an Analysis of Gender in Video Game Culture: Exploring Gender Specific Vocabulary in Video Game Magazines." In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*, 333–341. Riga, Latvia. http://ceur-ws.org/Vol-2612/short20.pdf.

Schmidt, Thomas, Florian Kaindl und Christian Wolff. 2020c. "Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit." In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*, 157–172. Riga, Latvia. http://ceur-ws.org/Vol-2612/paper11.pdf.

Schmidt, Thomas, Alina El-Keilany, Johannes Eger und Sarah Kurek. 2021a. "Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies." In *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*. Krasnoyarsk, Russia. https://epub.uni-regensburg.de/50867/

Schmidt, Thomas, Johanna Grünler, Nicole Schönwerth und Christian Wolff. 2021b. "Towards the Analysis of Fan Fictions in German Language: Exploration of a Corpus from the Platform Archive of Our Own." In *2nd International Conference of the European Association*

for Digital Humanities (EADH 2021). Krasnoyarsk, Russia. https://epub.uni-regensburg.de/50829/

Schmidt, Thomas, Johannes Hoffmann und Christian Wolff. 2022. "Analyzing Character Networks in Crossover Fan Fictions of Archive of Our Own." In *Workshop on Computational Methods in the Humanities 2022 (COMHUM 2022)*. Lausanne, Switzerland: Laboratoire laussannois d'informatique et statistique textuelle. https://wp.unil.ch/llist/files/2022/06/COMHUM_2022_paper_14.pdf.

Schmidt, Thomas und Sarah Kurek. 2022. "Der Einsatz von Computer Vision-Methoden Für Filme - Eine Fallanalyse Für Die Kriminalfilm-Reihe Tatort." In *DHd 2022 Kulturen Des Digitalen Gedächtnisses. 8. Tagung Des Verbands "Digital Humanities Im Deutschsprachigen Raum" (DHd 2022)*, herausgegeben von Michaela Geierhos, Peer Trilcke, Ingo Börner, Sabine Seifert, Anna Busch und Patrick Helling, 65–72. Potsdam, Germany: Zenodo. https://doi.org/10.5281/zenodo.6328169.

Schmidt, Thomas, Fabian Schiller, Matthias Götz und Christian Wolff. 2023. "A Corpus of Memes from Reddit: Acquisition, Preparation and First Case Studies." In *INFORMATIK 2023 - Designing Futures: Zukünfte Gestalten*, herausgegeben von Maike Klein, Daniel Krupka, Cornelia Winter, and Volker Wohlgemuth, 795–804. Bonn: Gesellschaft für Informatik e.V. https://doi.org/10.18420/inf2023_89.

Tosenberger, Catherine. 2008. "Homosexuality at the Online Hogwarts: Harry Potter Slash Fanfiction." *Children's Literature* 36 (1): 185–207. https://doi.org/10.1353/chl.0.0017.

Yin, Kodlee, Cecilia Aragon, Sarah Evans und Katie Davis. 2017. "Where No One Has Gone Before: A Meta-Dataset of the World's Largest Fanfiction Repository." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6106–6110. Denver Colorado USA: ACM. https://doi.org/10.1145/3025453.3025720.

Zhang, Weiwei, Jackie Chi Kit Cheung und Joel Oren. 2019. "Generating Character Descriptions for Automatic Summarization of Fiction." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July): 7476–7483. https://doi.org/10.1609/aaai.v33i01.33017476.

HermeneuTopic. Ein Workflow zur interaktiven mixed-methods Exploration (philosophie-)historischer Textkorpora.

Reiners-Selbach, Stefan
stefan.reiners-selbach@hhu.de
Heinrich-Heine-Universität Düsseldorf, Deutschland
ORCID: 0000-0002-4763-4348

Baedke, Jan
jan.baedke@rub.de
Ruhr-Universität Bochum, Deutschland
ORCID: 0000-0003-2138-785X

Böhm, Alexander
alexander.boehm@rub.de
Ruhr-Universität Bochum, Deutschland

Fábregas-Tejeda, Alejandro
Alejandro.FabregasTejeda@rub.de
Ruhr-Universität Bochum, Deutschland
ORCID: 0000-0002-1797-5467

Straetmanns, Vera
vera.straetmanns@rub.de
Ruhr-Universität Bochum, Deutschland

Einleitung

Methoden der Digital Humanities halten Einzug in die Philosophie. In der Fachdiskussion, insbesondere der historisch informierten Wissenschaftstheorie, finden sich mehr und mehr Arbeiten, die mit Text Mining-Methoden historische Korpora untersuchen (Malaterre, 2019; Noichl, 2021; Pence, 2022; Bertoldi et al., 2023; Böhm et al., 2022; Lean et al., 2023; Malaterre und Lareau, 2022). Für solche und ähnliche Herangehensweisen prägt sich derzeit der Begriff *empirical philosophy of science* aus – mit dem Desiderat, sich nicht mehr nur mit willkürlicher Auswahl scheinbar kanonischer Texte zu begnügen, sondern Wissenschaft in der Breite historisch zu untersuchen, „to detect features that would evade unaided examination" (Lean et al., 2023).

So konnte auch auf der DHd 2023 das erste Panel zur „Digitalen Philosophie- und Pädagogikgeschichte“ stattfinden. Die Probleme, die hier unter anderem identifiziert worden sind, beziehen sich auf die Herangehensweisen an multilinguale Korpora (Noichl, 2023) sowie auf die Lückenhaftigkeit und Undurchschaubarkeit der Datengrundlagen der digitalen Philosophie und unsere Zugriffe auf diese (Heßbrüggen-Walter, 2023). Wir versuchen im Folgenden, auf beide aufgewiesenen Problemfelder zu reagieren, indem wir ein neues methodisches Konzept als Workflow vorstellen: HermeneuTopic.

Multilinguale Korpora stellen hinsichtlich Analyse und Modellierung in vielen Bereichen der Digital Humanities eine Herausforderung dar (Dombrowski, 2020; Noichl, 2023). Ein ausschließlicher Fokus auf einen Diskurs in nur einer Sprache ist allerdings nicht zufriedenstellend (Galina Russel, 2014; Noichl, 2023); und die maschinelle Übersetzung eines Korpus, um sprachliche Homogenität herzustellen (s. Bertoldi et al., 2023; Böhm et al., 2022; Malaterre und Lareau, 2022), ist ebenfalls nur eine Behelfslösung (Noichl 2023). Wir stimmen Noichl zu, dass multilinguale Sprachmodelle hierfür die sinnigste Lösung darstellen, um sprachliche Diversität zu wahren, gleichzeitig aber Arbeiten unterschiedlichster Sprachen miteinander vergleichen und verarbeiten zu können. Es ist allerdings trotz sprachlicher Diversität schwer möglich, dem Anspruch, eine Disziplin in Gänze zu erfassen, gerecht zu werden (Heßbrüggen-Walter, 2023). Heßbrüggen-Walter mahnt in diesem Zusammenhang, stärker fragengeleitet vorzugehen und einzugrenzen, dabei Herangehensweisen, Provenienzen, aber auch Grenzen und Biases bei der Korpusbildung transparent darzustellen (2023). Eine mögliche Lösung hierfür stellt ein stärker hermeneutisches, mixed-methods Vorgehen dar: Die Modellierung und computationelle Analyse des Untersuchungsgegenstands wird nicht ans Ende der Arbeit, sondern als Element der Exploration an den Beginn gestellt. Während die Diskussion um die Rolle der Hermeneutik in den Digital Humanities ebenfalls vielfältig geführt wird und sich vielerorts die Frage nach der Anpassung der hermeneutischen Methode an die neuen Gegebenheiten der digitalen Textanalyse findet (s. z.B. Beyen, 2013; Dobson, 2019; Fickers, 2020; Ramsay, 2011; van Zundert, 2016), könnte man allerdings ebenso mit der Anpassung oder der Einbindung der digitalen Methoden an das hermeneutische Vorgehen reagieren.

Wir stellen im Folgenden einen Prototyp für einen solchen Workflow vor. Ziel des Workflows ist es, sowohl möglichst sprachagnostisch als auch (meta-)datenagnostisch zu sein. Texte sollen möglichst unabhängig von ihrer Sprache, der Menge ihrer Metadaten oder ihrer Länge analysiert und erkundet werden können. Dazu nutzen wir ein multilinguales Sprachmodell (s. auch Noichl, 2023), mit dem wir exemplarisch Textvektoren von einem kleineren Korpus von 322 Texten generieren. Dasselbe Sprachmodell nutzen wir, um ein multilinguales, vektorbasiertes Topic Modeling durchzuführen. Aus den Resultaten erzeugen wir daraufhin eine interaktive Karte, mit deren Hilfe das Korpus exploriert werden kann. Der resultierende Vektorraum, in den wir die Textvektoren als Streudiagramm projizieren, stellt die Ähnlichkeit der Texte zueinander dar. Daneben annotieren wir die Texte mit ihren Metadaten (sofern vorhanden) und ihren Topics. Aus dieser Visualisierung kann wiederum per Klick auf die Punkte im Streudiagramm zu den durch sie repräsentierten Texten navigiert werden. Damit stellen wir einen Workflow an einem Beispiel vor, der es erlaubt, basierend auf digitalen Analysen Fragen an auch kleineren Textkorpora zu explorieren und über eine Visualisierung als Navigationstool in die Close-Reading-Analyse überzugehen.

Datensatz

Der von uns genutzte Datensatz soll nur als Beispiel dienen, um den Workflow zu zeigen. Dabei handelt es sich um einen von uns bereits vorgestellten Datensatz (Böhm et al., 2022). Dieser besteht aus 322 Texten aus den Journals Acta Biotheoretica und Bios, der Schriftenreihe Bibliotheca Biotheoretica sowie einer von Julius Schaxel herausgegebene Schriftenreihe und weiteren Monografien ohne Reihe (1901-1971, s. Fig. 1).

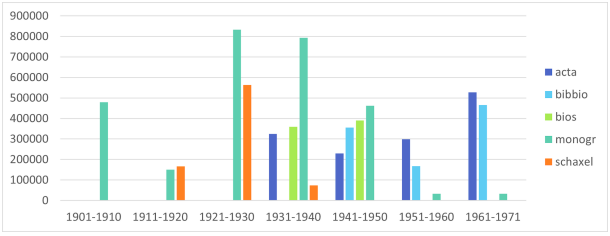


Abbildung 1: Zusammensetzung des Beispielkorpus (in Wortanzahl per Serie per Jahrzehnt)

Es ist also ein vergleichsweise kleines, teilweise handselektiertes Korpus, das einige ‚kanonische‘ Werke sowie Journals zu deren Ergänzung in der Breite enthält. Das Korpus setzt sich aus deutsch-, englisch- und französischsprachigen Texten zusammen. Maßgebend war dabei die Forschungsfrage: Wie entwickelt sich die Theorie der Biologie im Laufe des 20. Jahrhunderts von einer philosophischen zu einer mathematisch-naturwissenschaftlichen Disziplin? Die Textauswahl ist vorläufig und wird derzeit stetig ergänzt; sie reicht allerdings aus, um den Workflow vorzustellen.

Methode

Ziel des Workflows ist es, aus Texten (wenn vorhanden, ergänzt mit Metadaten) ein interaktives, analysebasiertes Tool zur Navigation dieser Textdaten zu generieren: HermeneuTopic. Dazu werden die eingegebenen Texte zunächst in Chunks von 500 Wörtern zerlegt, sodass aus 322 Texten 13662 Chunks entstehen. Die ursprünglichen Dokumente werden aus ihren Chunks wieder zusammen-

gesetzt und als XML-Dateien abgespeichert, wobei die einzelnen Chunks als Paragraphs (p-Element) mit xml:id referenziert werden, damit aus der resultierenden Visualisierung heraus einzelne Chunks im Kontext ihrer Texte aufgerufen werden können. Ein solches Chunking ist einerseits für das folgende Topic Modeling und das Embedding sinnvoll, die kleinere Textabschnitte für die Texteingabe vorsehen. Andererseits dient es dazu, Homogenität hinsichtlich der Textlänge herzustellen. Dies ermöglicht einen ebenfalls agnostischen Umgang mit der Textlänge, beziehungsweise die Verarbeitung von Texten mit unterschiedlichsten Textlängen (im Fall der Beispieldaten Monographien und Journalartikel) in derselben Analyse. Nach dem Chunking werden die Chunks mithilfe des multilingualen Sprachmodells *Sentence-BERT* (*distiluse-base-multilingual-cased)*[1] (Reimers und Gurevych, 2019; Reimers und Gurevych, 2020) in einen 512-dimensionalen Vektorraum projiziert. Dieser 512-dimensionale Vektorraum wird mit UMAP (McInnes et al., 2020) auf zwei Dimensionen reduziert, in die die Chunks als Punktvektoren projiziert werden. Dasselbe multilinguale Sprachmodell wird genutzt, um mit Top2Vec (Angelov, 2020) ein multilinguales Topic Modeling durchzuführen. Im Gegensatz zu herkömmlichen LDA-Topic Models (Blei et al., 2003) spricht für Top2Vec nicht nur die Möglichkeit, multilinguale Sprachmodelle einzubinden. Während bei LDA-Verfahren die Zahl der Topics im Vorfeld festgelegt werden muss, optimiert Top2Vec die Anzahl der Topics selbst. Daneben ist für Top2Vec-Analysen auch weniger Vorverarbeitung wie Lemmatisierung oder Stemming notwendig, was bei mehrsprachigen Korpora zusätzliche Herausforderungen bieten würde. Außerdem handelt es sich bei Top2Vec um ein vektorbasiertes Verfahren, das gleichzeitig als Clustering dienen kann und somit für die Visualisierung hilfreich ist. Zuletzt wird Bokeh (Bokeh Developement Team, 2018) genutzt, um aus einem Python-Programm heraus eine interaktive Java-Script-gestützte Visualisierung zu generieren, die im Browser als HTML-Dokument dargestellt werden kann: Die von UMAP generierten Punktvektoren werden als Streudiagramm geplottet. Dabei werden, wenn vorhanden, Metadaten sowie die Topic-Indizes und Topic-Keys des Top2Vec-Topic Modeling als Informationen für ein Mouseover-Pop-Up übergeben; die Top2Vec-Topics werden ebenfalls als thematische Cluster genutzt, indem die Punkte nach Topic eingefärbt werden. Um diese Cluster besser visualisieren zu können, nutzen wir Top2Vecs Funktion zur hierarchischen Topic-Reduktion, wenn die Anzahl der Topics 20 überschreitet. Die Annotation erfolgt mit den Keys und Indizes der nicht-reduzierten Topics. Bei Klick auf einen der Punkte gelangt man im entsprechenden XML-Dokument zum jeweils referenzierten Chunk, der durch den Punkt im Streudiagramm repräsentiert wird. Die resultierende Visualisierung kann als HTML mit eingebettetem Java-Script exportiert werden und ist somit – gemeinsam mit den XML-Dateien des Korpus – portabel und kann ebenfalls als Website zur Nachnutzung zur Verfügung gestellt werden. Grundsätzlich sind damit alle Daten in ein Standard-Datenformat überführt und können ohne weitere Umstände nachgenutzt werden. Daneben ist sämtlicher Programmcode, der genutzt wurde, quelloffen und kann ebenfalls nachgenutzt werden.

Vorläufige Ergebnisse

Die Anwendung des beschriebenen Workflows auf den Beispieldatensatz resultiert in einem interaktiven Bokeh-Plot, der wie eine Landkarte des Textkorpus gelesen und als Navigationstool genutzt werden kann. Die unterschiedlichen Topic-Cluster (58, reduziert auf 20) bilden Themenregionen, die erkundet werden können (s. Fig. 2).

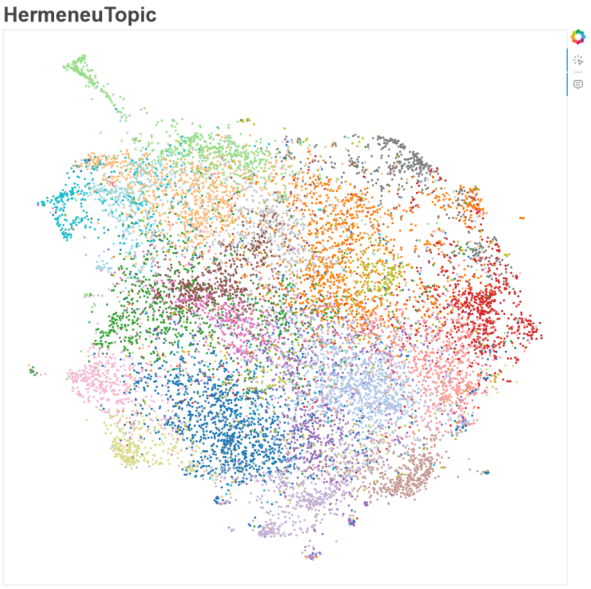


Abbildung 2: Ansicht des interaktiven Bokeh-Plot. Jeder Punkt repräsentiert einen Text-Chunk. Die Dimensionen geben Textähnlichkeit wieder. Farben stellen hierarchisch reduzierte Top2Vec Topics dar.

Es ist am Scatterplot ersichtlich, dass die unterschiedlichen Sprachen unseres Datensatzes (Deutsch, Englisch und Französisch) aligniert sind und nicht getrennt clustern, ohne dass mit Übersetzung sprachliche Homogenität hergestellt worden wäre. Dies lässt sich ebenfalls an einigen Beispielen nachvollziehen (s. Fig. 3)

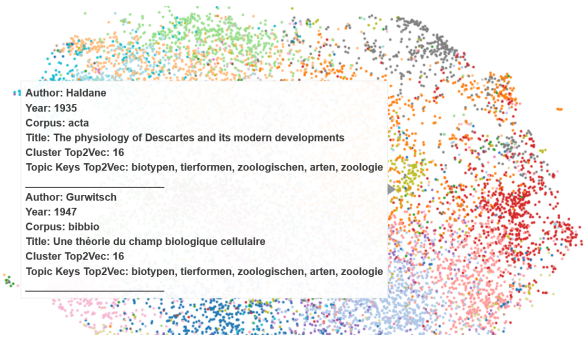


Abbildung 3: Beispiel für ein Pop-Up-Fenster, das bei Mouseover Metadaten sowie die Ergebnisse des Topic Modeling anzeigt; außerdem ein englisch- sowie ein französischsprachiger Text, die nicht sprachlich getrennt clustern.

Outlier bilden wiederum thematisch randständige Texte, wie etwa Texte zur Psychologie (der Tiere). Dies ist zu erwarten in einem Korpus, das sich mit theoretischer Biologie befasst und dessen Zentrum die Diskussion über Gesetzmäßigkeiten und deren Mathematisierbarkeit bildet (s. Fig. 4).

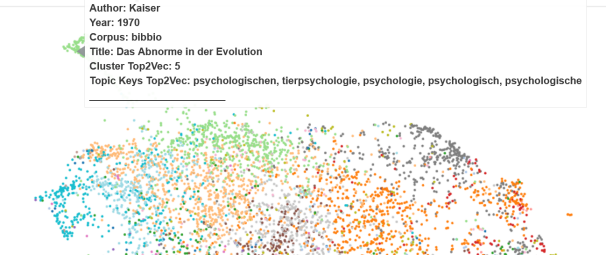


Abbildung 4: Blick auf einen Outlier: Psychologie bzw. Tierpsychologie

Will man nun einen Text weiterverfolgen, gelangt man mit Klick auf den entsprechenden Punkt im Scatterplot zu dem durch diesen repräsentierten Chunk, welcher im dazugehörigen Gesamtdokument als p-Element im XML referenziert ist (s. Fig. 5).

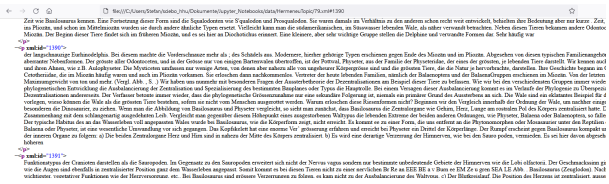


Abbildung 5: Beispiel für die Navigation zu einem Text-Chunk in einem XML-Dokument.

Werfen wir einen kurzen Blick auf das Topic Modeling, sehen wir, dass auch hier sinnvolle Ergebnisse entstehen – beim Top2Vec-Topic Modeling sind die Topics mit den niedrigsten Indizes diejenigen mit dem höchsten Anteil am Korpus:

Topic 0: 'biosysteme', 'organsysteme', 'organogenese'…
Topic 1: 'biologen', 'biologie', 'biotheoretica'…
Topic 2: 'kompliziert', 'kompliziertheit', 'komplizierte'…

So sind die Topics 0 und 1 wenig überraschend bei einem Korpus zur Theorie der Biologie, weil hier scheinbar, zumal in Topic 1, genau dies verhandelt wird. Interessant ist Topic 2, da hier die Komplexität von biologischen Systemen im Zentrum steht. In den Keys folgen ebenfalls die Wörter „Zusammensetzung“ und „Kombination“, die weiter darauf hindeuten. Doch an dieser Stelle kommen Fragen an das Korpus auf, denen weiter im Detail nachgegangen werden muss - was mithilfe des resultierenden Navigationstools in Auseinandersetzung mit der topologischen Struktur sowie den Ergebnissen des Topic Modelings und den ursprünglichen Texten geschehen kann.

Diskussion

HermeneuTopic stellt einen sprachagnostischen Workflow dar, der es ermöglicht, selbst ohne umfassende Metadaten, Textkorpora unterschiedlichster Textlänge zu erkunden, da die Visualisierung gänzlich aus Textdaten und nicht aus Metadaten generiert wird. Dabei ermöglicht die interaktive Visualisierung, ähnlich einer Landkarte, ein heuristisches und exploratives Vorgehen: Die Strukturen und Topics, die die digitalen Analysen aufgedeckt haben, ermöglichen eine stärker fragengeleitete und hypothesengestützte Close Reading-Analyse in deren Anschluss. Die digitalen Analysen werfen Fragen auf, die unmittelbar an den Texten selbst untersucht werden können. Damit stellt der vorgeschlagene Workflow einen mixed-methods Ansatz dar, der die Vorzüge von herkömmlichen hermeneutischen Herangehensweisen und digitalen Analysen kombinieren soll, indem die Analysen – durch interaktive Bereitstellung der Daten – sich an den hermeneutischen Herangehensweisen orientieren. Hiermit reagieren wir auf Heßbrüggen-Walter (2023), indem der hier vorgestellte Workflow nicht auf Ganzheit der Daten angewiesen ist oder abzielt, sondern dezidiert im Vorfeld eingrenzt, auf kleinere Korpora zugeschnitten ist, um auch im zweiten Schritt den Texten händisch-lesend zu begegnen. Auch wenn es sich bei den von uns gewählten Daten nur um Beispieldaten handelt, konnten wir hiermit gleichzeitig auf die von Noichl (2023) geäußerte Kritik an unserem bisherigen Approach mit Machine Translation eingehen und seine Forderung, multilingual zu arbeiten, einlösen.

Ausblick

Wir stellen mit diesem Beitrag nur einen ersten Prototyp vor, der weiter ausgebaut werden kann. Das multilinguale Sprachmodell *distiluse-base-multilingual-cased* konnte in dieser Studie auch ohne weiteres Fine Tuning sinnvoll eingesetzt werden. Es ist aber zu erwarten, dass durch ein gezieltes Fine Tuning auf dem jeweils zu untersuchenden Korpus noch bessere Ergebnisse zu erzielen wären. Insbesondere das Fachvokabular der (theoretischen) Biologie sollte dem vortrainierten Sprachmodell noch vergleichsweise unbekannt sein. Es wäre daher zu erwarten, dass

nach einem Fine Tuning spezifischere Topics und schärfer umrissene Cluster erkennbar wären. Allerdings stellt die Qualitätssicherung der Ergebnisse noch grundsätzlich ein Problem dar. Top2Vec bietet keine Option für statistische Qualitätsmaße, wie dies bei LDA-Topic Modeling möglich ist, sodass bisher händische Kontrolle und Beurteilung durch folgendes Close Reading die einzigen Möglichkeiten bieten, die Qualität der Analysen zu überprüfen. Daneben ist das vorgenommene Chunking basierend auf Wortanzahl nur eine pragmatische Lösung. Hier wäre es ein Desiderat, auch etwa bei fehlender Auszeichnung von Textabschnitten solche natürliche Textunterteilungen automatisiert erkennen und im XML abbilden zu können. Außerdem könnte der hier vorgestellte Workflow neben anderen Möglichkeiten der Filterung und Suche durch eine Semantic Search-Funktion ergänzt werden: Das Embedding, das mit dem Sprachmodell erstellt wird, wäre auch für eine Symmetric Semantic Search-Funktion geeignet.

Fußnoten

1. Steht keine GPU zur Verfügung oder ist eine Optimierung der Laufzeit relevant, kann mit Top2Vec auch eine Version des mutlilingualen Universal-Sentence-Encoder (z.B. universal-sentence-encoder-multilingual-large) genutzt werden.

Bibliographie

Angelov, Dimo. 2020. „Top2Vec: Distributed Representations of Topics“. arXiv. https://doi.org/10.48550/arXiv.2008.09470.

Bertoldi, Nicola, Francis Lareau, Charles H. Pence, Christophe Malaterre. 2023. A quantitative window on the history of statistics: topic-modelling 120 years of Biometrika. *Digital Scholarship in the Humanities* 2023. *https://doi.org/10.1093/llc/fqad072*

Beyen, Marnix. 2013. „A Higher Form of Hermeneutics? The Digital Humanities in Political Historiography.“ BMGN - Low Countries Historical Review. 128 (4): 164–70.

Blei, David M., Andrew Y. Ng, und Michael I. Jordan. 2003. „Latent Dirichlet Allocation“. *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Böhm, Alexander, Stefan Reiners-Selbach, Jan Baedke, Alejandro Fábregas Tejeda, und Daniel J. Nicholson. 2022. „What was Theoretical Biology? A Topic-Modelling Analysis of a Multilingual Corpus of Monographs and Journals, 1914-1945”. In: Geierhos, M. (ed.). *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*. https://doi.org/10.5281/ZENODO.6304590

Bokeh Development Team. 2018. *Bokeh: Python library for interactive visualization*. https://bokeh.pydata.org/en/latest/.

Dobson, James E. 2019. *Critical Digital Humanities: The Search for a Methodology*. University of Illinois Press. https://doi.org/10.5406/j.ctvfjd0mf.

Dombrowski, Quinn. 2020. „What’s a ‚Word‘: Multilingual DH and the English Default“. 15. Oktober 2020. https://www.quinndombrowski.com/blog/2020/10/15/whats-word-multilingual-dh-and-english-default/.

Fickers, Andreas. 2020. „Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?“ *Zeithistorische Forschungen/ Studies in Contemporary History* 17: 157–68. https://doi.org/10.14765/zzf.dok-1765.

Galina Russell, I. 2014. „Geographical and Linguistic Diversity in the Digital Humanities“. *Literary and Linguistic Computing* 29 (3): 307–16. https://doi.org/10.1093/llc/fqu005.

Heßbrüggen-Walter, Stefan. 2023. „Offene Daten für die digitale Philosophie: Anforderungen an eine Datensammlung zur Philosophie und ihrer Geschichte“. In *DHd2023: Open Humanities, Open Culture* (First Edition). Zenodo, hg. v. Anna Busch und Peer Trilcke. https://doi.org/10.5281/zenodo.7688632

Lean, Oliver M., Luca Rivelli, und Charles H. Pence. 2023. „Digital Literature Analysis for Empirical Philosophy of Science’“. *British Journal for the Philosophy of Science* 74. https://doi.org/10.1086/715049.

Malaterre, Christophe, Jean-François Chartier, und Davide Pulizzotto. 2019. „What Is This Thing Called Philosophy of Science? A Computational Topic-Modeling Perspective, 1934–2015“. *HOPOS: The Journal of the International Society for the History of Philosophy of Science* 9 (2): 215–49. https://doi.org/10.1086/704372.

Malaterre, Christophe, und Francis Lareau. 2022. „The Early Days of Contemporary Philosophy of Science: Novel 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. 2020. „UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction“. arXiv. https://doi.org/10.48550/arXiv.1802.03426.

Noichl, Maximilian. 2021. „Modeling the Structure of Recent Philosophy“. *Synthese* 198 (6): 5089–5100. https://doi.org/10.1007/s11229-019-02390-8.

Noichl, Maximilian. 2023. „PhilroBERTa: Ein multilinguales Sprachmodell zur Beantwortung philosophiehistorischer Fragestellungen“. In *DHd2023: Open Humanities, Open Culture* (First Edition). Zenodo, hg. v. Anna Busch und Peer Trilcke. https://doi.org/10.5281/zenodo.7688632

Pence, Charles H. 2022. „Testing and Discovery: Responding to Challenges to Digital Philosophy of Science“. *Metaphilosophy* 53 (2–3): 238–53. https://doi.org/10.1111/meta.12549.

Ramsay, Stephen. 2011. *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press. https://www.jstor.org/stable/10.5406/j.ctt1xcmrr.

Reimers, Nils, und Iryna Gurevych. 2019. „Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks“. arXiv. https://doi.org/10.48550/arXiv.1908.10084.

Reimers, Nils, und Iryna Gurevych. 2020. “Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation.” arXiv. https://doi.org/10.48550/arXiv.2004.09813.

Zundert, Joris J. van. 2016. „Screwmeneutics and Hermenumericals: The Computationality of Hermeneutics.“ In *A New Companion to Digital Humanities*, hg. v. Susan Schreibman, Ray Siemens und John Unsworth ., 331–47. Blackwell Companions to Literature and Culture 93. Chichester: John Wiley & Sons. https://doi.org/10.1002/9781118680605.ch23.

Katalog und Textkorpus zu Diskettenmagazinen der 1980er und 1990er (Re-)Digitalisierung frühen digitalen Kulturerbes

Roeder, Torsten
dh@torstenroeder.de
Universität Würzburg, Deutschland
ORCID: 0000-0001-7043-7820

Yannik, Herbst
yannik.herbst@uni-wuerzburg.de
Universität Würzburg, Deutschland
ORCID: 0000-0002-6547-9599

Johannes, Leitgeb
johannes.leitgeb@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland
ORCID: 0009-0006-2058-9133

Madlin, Marenec
madlin@mrsmuseum.de
Universität Würzburg, Deutschland
ORCID: 0009-0003-7434-827X

Tomash, Shtohryn
tomash.shtohryn@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland
ORCID: 0009-0000-4597-603X

Überblick

Dieser Vortrag präsentiert die Ergebnisse eines Drittmittelprojekts, das sich die Erschließung von Diskettenmagazinen der 1980er und 1990er Jahre zum Ziel gesetzt hat. bei den sogenannten „Diskmags“ handelt es sich um digitale multimediale Zeitschriften, die auf elektromagnetischen Floppy-Disks verbreitet wurden und auf klassischen Heimcomputersystemen wie z.B. Apple II, C64, Sinclair ZX Spectrum, Schneider CPC 464 und vielen anderen lesbar waren. Anfangs vor allem als kommerzielle Magazine für die frühen, hochpreisigen Systeme produziert (Beispiel dazu in Abbildung 1), entstanden mit der Markteinführung günstigerer Geräte aber auch bald Magazine aus der rapide wachsenden Heimcomputer-Community (Beispiel dazu in Abbildung 2). Diskmags stellten in der Zeit vor dem Breitbandanschluss und dem Web 2.0 ein relevantes „born digital“-Medium dar, in dem sich unterschiedlichste frühe digitale Kulturszenen hauptsächlich durch Text, aber auch mittels Bild, Animation und Ton untereinander austauschten. Inhaltlich waren Diskmags aber nicht auf die Heimcomputerszene beschränkt, sondern es existierten beispielsweise auch Fanzines und Literaturmagazine (Beispiel dazu in Abbildung 3).

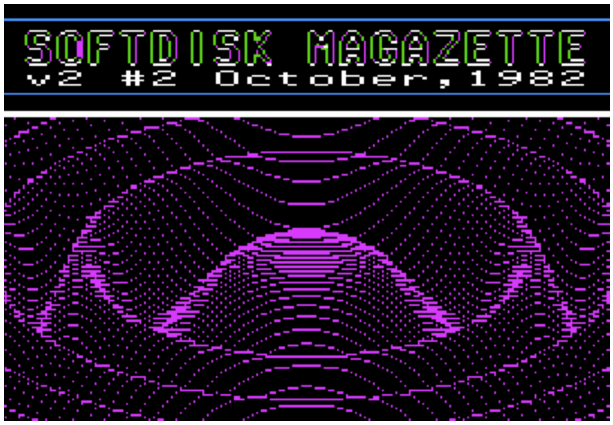


Abbildung 1: Das kommerzielle Magazin »Softdisk« (1981-1985, System: Apple II) war das erste Diskmag, das auf 5.25"-Diskette publiziert wurde; davor gab es bereits ähnliche Magazine auf Datasette.

Abbildung 2: Das Szene-Magazin »Bad Mag« (1992-1993, System: Amstrad/Schneider CPC 464) wurde auf der ungewöhnlichen 3-Zoll-Diskette verbreitet und versammelte Beiträge zur der Demoszene in Großbritannien, Irland und Westdeutschland.

Abbildung 3: Das Magazin »Warp« (1995-1996, System: Atari ST) richtete sich ausschließlich an Star Trek Fans und funktionierte über eine gamifizierte Menüsteuerung.

Die Überlieferungslage der Diskmags ist prekär, da sie nicht durch Bibliotheken oder Archive gesammelt wurden (vgl. Roeder 2020). In dem hier vorgestellten Projekt entstand ein Katalog, durch den jetzt mehrere tausend Diskettenmagazine und weit über zehntausend Ausgaben erstmals systemübergreifend und nach wissenschaftlichen Kriterien recherchierbar sind (Zwischenergebnisse und illustrierte Beispiele siehe Roeder et al. 2023). Mithilfe von Textmining-Methoden wurde aus den Binärdateien deutschsprachiger Magazine ein Textkorpus erstellt, das nicht nur für die Volltextsuche, sondern auch für die Erforschung der Sprache und der Thematiken der frühen digitalen Kultur nachgenutzt werden kann. Dieser Beitrag wertet die bisherigen Projektergebnisse aus und stellt die angewendeten Methoden zur Diskussion.

Gefördert wurde das einjährige Projekt in erster Linie als Kooperationsprojekt des NFDI-Konsortiums Text + (siehe https://www.text-plus.org/forschungsdaten/kooperationsprojekte/), in dessen Datenservices sowohl die Katalog- als auch Textdaten einfließen. Außerdem erhielt das Projekt Zuschüsse durch den Unibund der Universität Würzburg und durch die Vogel Stiftung Dr. Eckernkamp. Gehostet wird das Projekt am Zentrum für Philologie und Digitalität der Universität Würzburg unter der Domain diskmags.de.

Hintergrund

In den 1980er Jahren etablierten sich Heimcomputer nach und nach als Unterhaltungskonsolen und Arbeitsgeräte gleichermaßen. Die Basis, die diese Entwicklung vorantrieb, bestand aus einer Subkultur von Computer-Enthusiasten, in der sich Anwender, Computerspieler, Programmierer, Spieleentwickler, Democoder sowie Hacker und Cracker zusammenfanden. Das authentische Erlebnis jener frühen digitalen Kultur ist uns heute jedoch nicht mehr präsent. Definiert wurde es zum einen durch die Hardware, die einerseits durch ein typisches „look and feel“, andererseits durch eine spezifische Ausstattung (Geschwindigkeit, Speicher, Soundkanäle, Farben, Bildauflösung, Peripherieanschlüsse) oft sehr charakteristisch war. Zum anderen war die Bedienung von Programmen mehr hybrid als „born digital“: Computern lagen dicke Handbücher bei, Anwendungsprogramme wurden oft mit umfangreichen Bedienungsanleitungen ausgeliefert, und auch in Spielen war das Druckbeiwerk manchmal unerlässlich. Programme wurden oft nicht nur digital kopiert, sondern händisch aus Printmagazinen abgetippt. Software erforderte Einarbeitungszeit und eine hohe Frustrationsschwelle, bei vergleichsweise geringer Stabilität der Betriebssysteme.

Die intensive Selbstdokumentation, die von der heutigen digitalen Massenkultur durch zahlreiche Medienkanäle geleistet wird, steht für die damalige Zeit nicht in derselben Dichte zur Verfügung. Eine besondere Rolle fällt deshalb den zahlreichen Periodika zu, die in jener Zeit produziert wurden und Einblicke in die Vielfalt der Heimcomputerszene geben. Einen guten Teil decken hier die professionellen Printmagazine ab, die allerdings das Geschehen innerhalb der diversen Subszenen nur selten fokussierten. Daher lohnt die nähere Betrachtung von Diskettenmagazinen: eine Art multimedialer Born-Digital-Journale, die mit dem Beginn der Heimcomputerkultur entstanden sind und nicht nur die charakteristischen multimedialen Ansätze jener Zeit spiegelt, sondern auch Dokumente der vielfältigen Kreativität und Lebendigkeit der einstigen Heimcomputer-Community darstellen.

Diskettenmagazine

Wie einleitend erläutert und veranschaulicht, handelte es sich bei Diskettenmagazinen um digitale Zeitschriften, die ausschließlich auf einem dafür geeigneten Computersystemen rezipiert werden konnten. Ein vereinheitlichtes System gab es nicht: Jedes Diskmag beinhaltete seine eigene Reader-Software.

Verbreitet wurden Diskettenmagazine stets auf ihrem namengebenden Medium, aber auf ganz unterschiedlichen Vertriebswegen. Einige konnte man regulär am Kiosk be-

ziehen, andere wurden gegen Einsendung von Retourporto und Leerdiskette oder einen angemessenen Unkostenbeitrag per Post versendet, manche verbreiteten sich ausschließlich durch Privatkopien.

Primär wurden sie in Gegenden mit hoher Heimcomputerdichte hergestellt. Prinzipiell gilt: Wo es Heimcomputer gab, gab es Diskmags. Durch die Bindung an bestimmte Computersysteme war ihr Verbreitungsgrad außerdem weniger durch sprachliche Grenzen, sondern primär durch technische Hürden limitiert.

Inhaltlich boten Diskettenmagazine sowohl Informationen zu digitalen Technologien, Geräten, Spielen und Anwendungsprogrammen als auch Programmier- und Bastelanleitungen. Einige Diskmags widmeten sich intensiv einem Bereich aus Grafik, Musik, Games, Literatur oder Fandom. Fast immer wurde Software (Anwendungen, Spiele, Demos) mitgegeben. Diskettenmagazine konkurrierten kaum mit den textbasierten Bulletin Board Systems (BBS) der späten 1980er, jedoch übernahm das World Wide Web im Verlauf der 1990er, spätestens mit dem Aufkommen von Breitbandverbindungen, die Funktion der multimedialen Informationsverteilung, und viele der Diskettenmagazine wurden zusammengelegt oder beendet. Nur wenige wurden online weitergeführt, manche führten dabei die Bezeichnung „Diskmag“ im Titel weiter; in der Demoszene ist dies bis heute üblich.

Katalogisierung

Bibliotheken sammelten Diskettenmagazine nur, wenn ein wesentlicher Printanteil vorlag, was nur selten der Fall war. Verlässliche Informationen und digitale Dokumente (vor allem Disk-Images als Binärdateien, aber auch Screenshots sowie Scans von zusätzlichem Printmaterial) finden sich am ehesten auf einschlägigen Fansites, die zudem häufig auf einzelne Systeme, Sprachen oder Themen begrenzt sind. Selbst die englischsprachige Wikipedia verzeichnet lediglich einen Bruchteil.

In der ersten Phase des Erschließungsprojektes stand deshalb die umfassende Katalogisierung im Vordergrund. Dazu wurden mehrere bestehende Verzeichnisse ausgewertet und zusammengeführt. Aktuell sind dies fünf Datenquellen: Demozoo (https://demozoo.org), Pouet (https://www.pouet.net), C64 Scene Database (https://csdb.dk), Internet Archive (https://archive.org, mehrere Collections) und ZXpress (https://zxpress.ru); weitere könnten zukünftig folgen. Parallel dazu wurde Sekundärliteratur ausgewertet (Volko 2012).

Diese Datenquellen wurden mithilfe von Scraping-Techniken geharvestet und in einem GitHub-Repository abgelegt (https://github.com/zpd-digital-editions/Diskmags). Für den Abgleich der Daten kam die bewährte Software OpenRefine (https://openrefine.org) zum Einsatz. Die Zuordnung der Einzelausgaben zu den jeweiligen Titeln war aufgrund vieler ähnlich lautender Titel und vielen Titelvarianten keine triviale Aufgabe; widersprüchliche Angaben wurden entsprechend nachrecherchiert. Insbesondere die Angaben zu Sprachen und Ursprungsländern stellten sich als problembehaftet heraus. Die vereinheitlichte Datensammlung wurde anschließend in ein Semantic MediaWiki exportiert und steht unter diskmags.de für die Recherche und Nachbearbeitung zur Verfügung. Mithilfe des Wikis können die existierenden, teils sehr aktiven Communities bei der Datenkuratierung direkt einbezogen werden. Es wird angestrebt, die erhobenen Katalogdaten über die Infrastruktur von Text+ in Normdatenkataloge und ins Linked Open Data Network einzuspielen.

Ergab deshalb die erste Schätzung vor Projektstart noch, dass vermutlich mit etwa 200 bis 300 Diskettenmagazinen zu rechnen sei, musste diese Zahl selbst nach Bereinigung zahlreicher Doppeleinträge erheblich nach oben korrigiert werden. Aus den genannten Datenbanken wurden Nachweise zu ca. 2.500 unikalen Titeln mit weit über 10.000 Einzelausgaben extrahiert. Die rein quantitative Relevanz dieses ungewöhnlichen Mediums wurde somit um den Faktor 10 unterschätzt.

Die Auswertung der Datensammlung ergab mehrere Erkenntnisse. Nach dem aktuellen Stand der Datensammlung ist der Höhepunkt der Diskmag-Kultur zwischen den späten 1980ern und dem Ende der 1990er anzusetzen (vgl. Abbildung 4). Davor existierten Diskmags vor allem für eher exklusive und hochpreisige Geräte wie Apple II. Die vor allem durch Commodore eingeleitete Preissegmentierung des Heimcomputermarktes beförderte die Herausbildung einer breiteren Heimcomputer-Community und einer entsprechenden Diskmags-Kultur. Mit dem Erscheinen leistungsstärkerer Computer setzte sich dies in mehreren Wellen z.B. auf Systemen wie Commodore Amiga und Atari fort, überraschenderweise aber auch auf dem eigentlich deutlich älteren Spectrum ZX, der in den postsowjetischen Ländern durch Nachbauten eine immense Rezeption erfuhr. In der späten Zeit dominierten dann MS-DOS-basierte Diskmags, bevor durch CD-ROMs und die Möglichkeiten des Web 2.0 das Diskmag als Medienformat weitgehend obsolet wurde. Es existiert allerdings weiterhin als Community-Produkt in einigen bis heute aktiven Retrocomputing-Szenen. Dies ist ein Glücksfall für die Erschließung, da so das teils obskure Wissen um die Funktionsweise der Hard- und Software durch Zeitzeugen zugänglich ist.

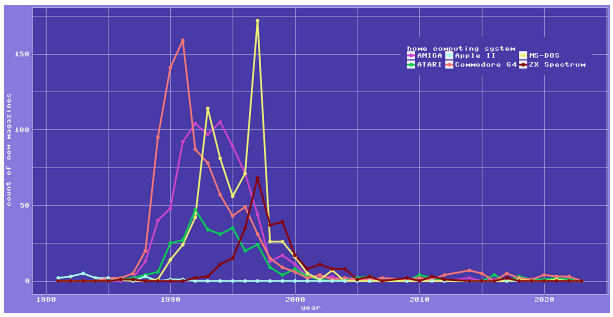


Abbildung 4: Anzahl neuer(!) Diskettenmagazine nach Jahr und System (Auswahl). Grafik: Johannes Leitgeb. Stand: Juli 2023.

Textkorpus

Das Textkorpus umfasst Plaintext aus mehreren vollständigen Jahrgängen deutschsprachiger Diskettenmagazine. Die Textextraktion aus mittlerweile obsoleten Datenträgern bzw. Datenformaten birgt eine Reihe von antizipierbaren Problemstellungen: So ist bereits die Zeichenkodierung stark herstellerabhängig und konnte zudem softwareseitig manipuliert werden, so dass die angezeigten Zeichen nicht immer dem jeweiligen Kodierungsstandard entsprachen. Ein Unicode-Mapping ist deshalb nicht immer eindeutig möglich, insbesondere wenn selbsterstellte Symbolzeichen verwendet wurden, deren Semantik manchmal über Blockgrafik-Elemente hinausging. Zudem wurde gerade Text aufgrund der relativ geringen Speicherkapazität von Floppys gerne komprimiert, wie auch eine entropische Analyse nahelegte. Zum Teil konnte hier jedoch auf Community-Initiativen zurückgegriffen werden, durch welche bereits teilweise eine Re-Digitalisierung vorliegt. Als entscheidend für die Textextraktion erwies sich jedoch die Erkenntnis, dass Texte häufig als Screencodes gespeichert wurden, die von den üblichen Zeichenkodierungen deutlich abweichen konnten. So z.B. lautet der ASCII- und Commodore-PETSCII-Code für den Buchstaben »A« einheitlich 65, der Screencode des Commodore 64 jedoch 1.

Im Unterschied zum Katalog sind bei der Erstellung des Textkorpus aber vor allem urheber- und personenschutzrechtliche Fragen zu berücksichtigen, wofür aufgrund der Komplexität der Rechtslage (z.B. nicht ermittelbare Urheber, nicht eindeutige Rechtefreigaben, ggf. sensible Daten) zum Projektabschluss eine entsprechende Handreichung vorgelegt werden wird.

Ziel ist die Publikation eines Textkorpus in einem flach hierarchischen Format wie DTABf (siehe https://www.deutschestextarchiv.de/doku/basisformat/), möglichst mit Named Entity Recognition und Artikelgrenzen, abhängig von der Rechtslage zugangsbeschränkt, ggf. zumindest mit der Möglichkeit eines Stichwortindexes und einer Wortstatistik. Für die Auswertung des Textkorpus ist beabsichtigt, sowohl das spezielle Vokabular als auch die besondere Stilistik der Heimcomputer-Szene mit Auswertungsverfahren der DH zu untersuchen. Mittels Named Entity Recognition lassen sich ggf. Netzwerke aus Personen und Gruppen erschließen, Softwaretitel auffinden und hinsichtlich ihrer Rezeption untersuchen oder Diskussionen um Computersysteme, deren Leistungsspektren und deren spezifischer Auslotung verfolgen. Ab einer größeren Textmenge erscheint auch die Anwendung von Topic-Modeling-Verfahren sinnvoll.

Ausblick

In den Digital Humanities steht die Auseinandersetzung mit digitalem Kulturerbe als Erhaltungsgegenstand noch relativ am Anfang. Im Umgang mit älteren Computersystemen existiert in den Digital Humanities vor allem im europäischen Raum, von einzelnen Projekten abgesehen, noch keine allgemeine Methodik. Standardformate wie TEI berücksichtigen digital überliefertes Material und Medienformate bislang nicht oder nur unzureichend. Die bereits hohe technologische Distanz erfordert jedoch geradezu eine verstärkte Auseinandersetzung mit digitalen Überlieferungsformen, sowohl bezogen auf einzelne Objekte als auch auf die historischen digitalen Ökosysteme als Ganzes. Insbesondere zu den 1980er und 1990er Jahren besteht eine erhebliche Lücke hinsichtlich der wissenschaftlichen Erschließung. Das Erschließungsprojekt zu Diskettenmagazinen stellt in dieser Hinsicht eine kleine Pionierleistung dar und bietet zahlreiche Anknüpfungspunkte für zukünftige Forschungsprojekte, die sich mit älterer Software, Hardware, Datenträgern und Datenformaten befassen.

Bibliographie

Roeder, Torsten: Rescuing Diskmags: Towards Scholarly [Re-]Digitisation of an Early Born-Digital Heritage, in: Magazén 1,3, 2022, S. 139–58.

Roeder, Torsten; Herbst, Yannik; Leitgeb, Johannes; Marenec, Madlin; Shtohryn, Tomash: Preserving the Early Born-Digital Heritage of Floppy Disk Magazines. Zenodo, 2023.

Ruan, Jianhai; P. McDonough, Jerome: , in: IEEE International Symposium on IT in Medicine & Education, 2009, S. 745–48.

Volko, Claus-Dieter: Enzyklopädie der Diskmags. Norderstedt: BoD, 2012.

Konzepträume: Ein Vorschlag zur besseren Abstimmung von Theoriehintergrund und digitalen Datenanalysen

Kremer, Dominik
dominik.kremer@fau.de
Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Deutschland

Lang, Sabine
sab.lang@fau.de
Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Deutschland

Wechselseitige Bezogenheit von Objekten und gesellschaftlichen Produktionsbedingungen

Wenngleich Humanities und Sozialwissenschaften oft als getrennte Forschungscommunities gedacht werden, sind beide Bereiche im materiellen Artefakt aufeinander bezogen. Die integrative Anwendung beider Perspektiven auf Fragestellungen der Digital Humanities birgt also ein erhebliches Potential. Dies werden wir im Folgenden zunächst (1) an Beispielen aus dem Bereich der Kunstgeschichte und der Kulturgeographie herausarbeiten, in der von uns postulierten Methode der Konzeptraumanalyse verdeutlichen, (2) dass bestimmte konzeptionelle Blickwinkel auch unterschiedliche digitale Repräsentationen erfordern und im Anschluss aufzeigen, (3) wie eine komplementäre Betrachtung geeigneter digitaler Daten vor dem Hintergrund verschiedener Erklärансätze als Chance verstanden werden kann, etablierte Deutungen im Arbeitsprozess produktiv zu hinterfragen.

Bereits die traditionelle Kunstgeschichte z.B. hinterfragt Objekte (i.B. zeremonielle Gegenstände) auf ihre gesellschaftlichen Produktionsbedingungen hin. Besonders bei Werken der frühen Neuzeit stehen Fragen nach Auftraggeber*in, Zweck, Werkgenese oder Werkstatttraditionen (Müller, 2023) im Vordergrund. Das Werkverstehen fordert deshalb nicht nur eine formale Analyse, sondern auch immer den Einbezug gesellschaftlicher Faktoren. Auch in der Bewertung digitaler Artefakte und im Besonderen digitaler Kunstwerke werden Faktoren, wie der Herstellungsprozess, die beteiligten Akteure*innen oder verwendete Technologien berücksichtigt. Die Herausforderungen potenzieren sich schließlich, wenn generative Verfahren, z.B. durch Bildgeneratoren wie DALL-E oder Stable Diffusion, durch kunsthistorische Theorien zum Wesen der Kunst, künstlerischem Stil oder zu Rezeptionsprozessen erfasst und hinterfragt werden sollen, insbesondere bezüglich der reproduzierten gesellschaftliche Vorurteile oder Stereotypen (Maganga, 2023). Es geht also nie ausschließlich um das (digitale) Objekt an sich, sondern um die Prozesse seiner Produktion im gesellschaftlichen Kontext und zwar unabhängig davon, ob diese Prozesse behutsam-moderierend (durch wissenschaftliche Befragung der Objekte) oder disruptiv (durch Deutungsmacht) gesetzt werden. Im Kontext ihrer digitalen Erschließung müssen schließlich auch Metadaten zu den Objekten nach ihren Produktionsbedingungen hin hinterfragt werden: Wer hat Metadaten erstellt, in welchem Kontext und zu welchem Zweck? Und welche Informationen über das Objekt wurden (bewusst) ausgelassen?

Die Betrachtung gesellschaftlicher Produktionsbedingungen hat eine lange Tradition in der Kunstgeschichte. Bereits die von Erwin Panofsky (1892-1968) vorgestellte ikonographisch-ikonologische Methode berücksichtigt den Entstehungskontext und zeigt, dass auch die Produktionsbedingungen – neben dem Artefakt – einen wichtigen Zugang zum Objekt bieten (Locher, 2007, 70-71). Die 3. Stufe, die „ikonologische Interpretation“, ermittelt z.B. die “»[…] Prinzipien […], die die Grundeinstellung einer Nation, einer Epoche, einer Klasse, einer religiösen oder philosophischen Überzeugung enthüllen […]«“ (Noll, 2011, 195) und die sich schließlich im Objekt widerspiegeln (Noll, 2011, 195). Das Beispiel Street Art verdeutlicht dies; nicht zuletzt spielen gesellschaftliche und räumliche Faktoren sowie neue Medientechnologien für die Entstehung, Gestalt und Aushandlung von Street Art eine wichtige Rolle (Glaser, 2017; Ross und Ferrell, 2016; Street Art & Urban Creativity, 2015).

In Ergänzung dazu verweisen Sozialwissenschaften bei der kritischen Analyse der Verheißungen technologischer Lösungen für gesellschaftliche Probleme (z.B. bei Smart Cities (Ahmad et al., 2022)), aber auch bei KI-Imaginationen (Zhou und Nabus, 2023) auf die transformative Wirkung und die damit verbundenen Risiken dieser Technologien auf Gesellschaft. Eine wichtige Rolle spielen dabei Science and Technology Studies (STS) (Hackett et al., 2008). Zur Beantwortung dieser Fragen stehen neben der Analyse räumlich und situativ gebundener Praktiken und deren kontinuierlicher Veränderung (Schatzki, 2002) auch die zugrunde liegenden Diskurse und deren Veränderung über die Zeit (Keller, 2008) im Mittelpunkt. In der Suche nach geteilten “Imaginaries” (Taylor, 2004), den dominanten Leitmetaphern (Lakoff und Johnson, 2003) und Erzählungen (Viehöver, 2001) von Gesellschaft, richtet sich der Blick wie in den Digital Humanities auf multimodale Bild- (Rose, 2001) und Textdaten (Mayring, 2016). Es ist offensichtlich, wie stark Social Studies in ihrer Umsetzung, selbst bei einem originär kritischen Blick auf (digitale) Technologien, bei der Analyse hegemonialen, konkurrierenden und sich ständig weiterentwickelnden (Kommunikations-)Strukturen von Gesellschaft (Jannidis et al., 2017) vom distant reading (Moretti, 2013) der Digital Humanities profitieren können. Kenntnisse über den Einsatz digitaler Methoden der Text- und Bildanalyse und zugleich über die Risiken ihres Gebrauchs bieten dabei erhebliches Potential für einen kundigen und dabei gleichzeitig reflektierten (digitalen) Methodeneinsatz (Kremer und Walker, 2022).

Counter-Modelling durch Konzepträume

Durch digitale Analysemethoden werden somit bestimmte strukturelle Befunde überhaupt erstmals möglich (Jannidis et al., 2017; Moretti, 2013). Im Zuge der Critical Data Studies (Dalton und Thatcher, 2014; Kitchin und Lauriault, 2014) hat es sich aber etabliert, auch solche forschenden digitalen Repräsentationen und Datensammlungen als unpolitische Räume anzuzweifeln (Iliadis und Russo, 2016), systematisch auf Macht, Überwachung und Steuerung in der vermeintlichen Entscheidungsunterstützung zu hinterfragen und ggf. durch Counter-Data (Dalton und Thatcher, 2014) unterrepräsentierten Interessen zu mehr Sichtbarkeit zu verhelfen (Iliadis und Russo, 2016).

Dadurch verzahnt sich ein kritischer, sozialwissenschaftlicher Blickwinkel mit konkreter eigener Arbeit auf digitalen Daten. Der Fokus lag dabei anfangs vor allem auf automatisch prozessierten Big Data (Boyd und Crawford, 2012).

In den Digital Humanities sehen wir eine Reihe weiterer digitaler Artefakte teils deutlich kleineren Umfangs, in die sich ebenso bestimmte fachliche Perspektiven von Anfang an in digitale Repräsentationen einschreiben, die es schwer machen, bei der Betrachtung der Daten einen ergänzenden oder vergleichenden Blickwinkel einzunehmen. Die dominanten Perspektiven (Denksysteme im Sinne von Kitchin und Lauriault, 2014) manifestieren sich dann – entweder unbewusst durch die organisatorische Struktur der Daten oder bewusst als Teil einer vorab erarbeiteten Informationsarchitektur eines Projekts – als Daten- und Informationsmodellierung. Drei Beispiele:

- Datenmodellierungen, die kuratierten Datensammlungen zugrunde liegen
- Datenmodellierungen, die (multi-modalen) Datenanalysen zugrunde liegen
- Informationsmodellierungen, die sich in Softwarebibliotheken und Softwarewerkzeuge einschreiben

Diese Modellierungspraxis erschließt natürlich primär den Zugang zu den in den Daten repräsentierten Phänomenen und Prozessen, verstellt aber dadurch effektiv und meist auch dauerhaft ihre durchaus mögliche Betrachtung aus einem anderen Blickwinkel. Daten sind also immer Ausdruck bestimmter deutender Machtverhältnisse (Selwyn, 2021). In Analogie zu *Counter-Data* sehen wir somit durch unseren Ansatz der Konzeptraumanalyse die Möglichkeit, Datenmodellierungen und ihre sozialen Produktionsbedingungen nicht erst ex-post, sondern komplementär dazu im Sinne eines *Counter-Modellings* bereits bei ihrer Erarbeitung systematisch auf ihre Anschlussfähigkeit an hilfreiche alternative Erklärsansätze zu prüfen. Eignen sich Kundendaten (Alter, Geschlecht, Umsatz) wirklich, um später Wahlverhalten vorherzusagen? Oder Daten aus WikiArt, um die Entwicklung von Stil zu studieren, obwohl signifikante Lücken hinsichtlich Zeit oder Herkunft der Künstler*innen bestehen? Unter welchen Einschränkungen darf eine Softwarebibliothek der kognitiven Linguistik dazu verwendet werden, um valide Ergebnisse für soziokulturell variantenreiche Diskursfragmente bereitzustellen?

Der Konzeptraum meint also den Möglichkeitsraum an vorstellbaren Übersetzungen zwischen Denksystemen und den aus ihnen heraus produzierten Datenmodellierungen, die bestimmte Lesarten technisch verarbeitbar und analysierbar machen und gleichzeitig implizit alternative Lesarten behindern. Unser Ansatz des *Counter-Modellings* zielt nicht auf *general-purpose* Datenmodelle, wie sie als Top-Level-Ontologien bereits seit Jahrzehnten sehr sorgfältig ausgearbeitet werden (vgl. CIDOC CRM; Bekiari et al., 2021) und auch nicht auf eine Kombination beliebiger Fachtraditionen. Wir plädieren vielmehr dafür, jede spezifische Anwendungsontologie (vgl. Timpf, 2002) bereits in der Ausarbeitungs- und Erprobungsphase sorgsam auf Konzepte aus unterschiedlichen Denksystemen hin zu prüfen, die den Erkenntniswert bezüglich einer an die Anwendung gestellten Frage substantiell erweitern und steigern können. Mit der Konzeptraumanalyse schlagen wir dazu einen ersten einfachen Workflow vor, der diesen Prozess unterstützen kann. Dies werden wir im Folgenden anhand folgender Beispiele näher erläutern:

1. Lebensweltliches Raumwissen über Stadträume äußert sich zumeist räumlich sehr vage und tageszeitlich äußerst dynamisch. Dennoch sind statische Kartenrepräsentationen nach dem Spatial Turn in den Humanities noch immer ein Hauptanalysewerkzeug für räumliche Datenanalysen (Moura de Souza et al., 2022). Im Sinne der Kritischen Kartographie (Glasze, 2009) muss es also der Anspruch sein, andere als kartesische Repräsentationen zu entwickeln, um Raumwissen sichtbar und analysierbar zu machen, z.B. durch ortsbezogene Modellierungen (Westerholt et al. 2020). Lebensweltliches Raumwissen dient dabei als Beispiel für sozialwissenschaftliche Fragen, deren Gegenstand teils materiell gebunden ist.
2. Kulturwissenschaftliche Fragen, die soziale Produktionsbedingungen von Objekten und deren Daten hinzuziehen, werden z.B. in der Provenienzforschung gestellt und beziehen sich u.a. auf Lücken in Provenienzinformationen bestimmter Objekte. In diesem Kontext sind Lücken oft problematisch, da sie auf Unrechtskontexte hinweisen können. Deshalb sollten sie entsprechend markiert und erläutert werden (Lang, 2023). Im Moment existieren aber noch keine Datenmodelle, die eine semantische Modellierung von Lücken erlauben. Um Lücken aussagekräftig zu modellieren, muss ein konzeptionelles Verständnis für Arten und Ursachen und Wissen über existierende Methodologie zur Datenmodellierung vorhanden sein.

Zur Prüfung möglicher unterschiedlicher Erkläransätze im Sinne eines *Counter-Modellings* schlagen wir folgende **Analysekategorien** vor:

- Denksystem: Wie oben beschrieben kann ein Denksystem, also ein entsprechend gerahmter Blick auf Welt durch einen Theorieansatz oder einen beruflichen Perspektive gewonnen sein, oder auch anhand einer Fragestellung aus Weltwissen ad hoc entwickelt werden.
- Konzeptraum: Der Konzeptraum umfasst die zentralen Begriffe unterschiedlicher Denksysteme. Im Konzeptraum kann geprüft werden, ob und auf welchen Daten diese Begriffe sichtbar gemacht werden können. Es kann außerdem geprüft werden, ob verschiedene Denksysteme an einer gemeinsamen Datenmodellierung beprobt werden können, oder ob unabhängige Datenmodelle nötig sind.
- Datenmodellierung: In der Datenmodellierung werden Daten so organisiert (und falls nötig genau für diesen

Zweck erhoben), dass ihre Struktur bestimmte Teile des Konzeptraums lesbar macht.

Diese verzahnen sich zu folgendem **Workflow** (vgl. Figure 1), der explorativ mehrfach durchlaufen werden kann, um komplementäre Antworten auf eine bestimmte Fragestellung ableiten zu können:

1. **Denksysteme identifizieren:** Welche unterschiedlichen Denksysteme können auf die Frage angewendet werden? Mit welchen Begriffen wird aus fachlicher Sicht auf die Fragestellung Bezug genommen? Welche Grundannahmen liegen dabei vor?
2. **Konzeptraum analysieren:** Welche zentralen Begriffe aus den Denksystemen sollen auf den Daten sichtbar gemacht werden? Welche abweichenden Anforderungen ergeben sich daraus für Daten, Datenarten und abgeleitete Informationen, die verarbeitet werden sollen?
3. **Entsprechende Datenmodellierung entwickeln:** In welchen Datenschemata sind die Daten organisiert, die die Grundlage für konkrete Datenarbeit bilden? Welche Grundannahmen liegen den bei der Verarbeitung verwendeten Programmbibliotheken zugrunde? Sind diese mit den zentralen Begriffen des Theorierahmens kompatibel?

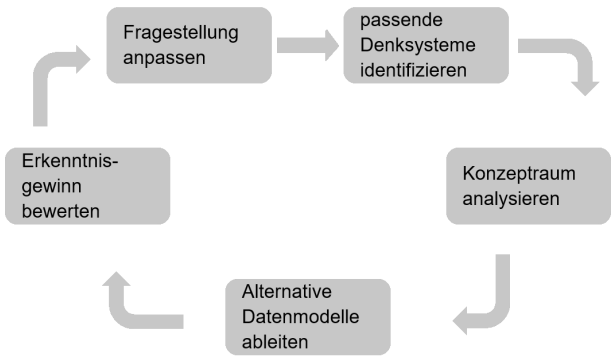


Abbildung 1: Workflow-Vorschlag zur Integration einer Konzeptraumanalyse als Übersetzungsschritt zwischen der Identifizierung passender Denksysteme und dem Ableiten geeigneter Datenmodelle. Bei jedem explorativen Durchlauf lässt sich der Mehrwert für die gestellte Fragestellung beurteilen.

Dem Konzeptraum kommt dabei wie beschrieben die entscheidende Übersetzungsleistung zwischen dem Theoriebezug und den in den Daten manifestierten Datenschemata zu. Wie in obigen Beispielen beschrieben (z.B. Kartendarstellungen für räumliche Datenanalysen), ist zu reflektieren, ob im jeweiligen Fall überhaupt Sekundärdaten herangezogen werden dürfen oder ob Daten explizit neu erhoben werden müssen, um die Forschungsfrage zu beantworten. Gibt es trotz der Verwendung von Metadatenstandards strukturelle Lücken in den Daten (z.B. bzgl. der Nachvollziehbarkeit der Forschungspraxis, durch die eine bestimmte Klassifikation erfasst wurde)? Kann eine zugrunde liegende Repräsentation (z.B. wie die oben genannte Karte) überhaupt als Grundlage für die Beantwortung einer bestimmten Forschungsfrage (Unsicherheit im Stadtraum) dienen?

Diesen Workflow werden wir nun an den zwei eingangs vorgestellten Beispielen näher erläutern.

Im Kontext von Krisendiskursen lohnen korpuslinguistische Betrachtungen zu Sprachgebrauchsmustern, um zu verstehen, wie Krisen sprachlich gebunden werden (Bubenhofer, 2009; Kremer und Walker, 2023). Dies betrifft akute Krisen wie Krankheiten (Semino et al., 2004) oder Pandemien (Kremer und Felgenhauer, 2022), aber durchaus auch alltägliches Erleben von Unsicherheit (Moura de Souza et al., 2022). Metaphern des Kampfs gegen die Krise (Semino, 2021) bedingen dabei die Suche nach der räumlichen Verortung der Krise (Brinks und Ibert, 2020), um Auswirkungen und Ursachen identifizieren, kontrollieren und bekämpfen zu können (Chapman und Miller, 2020). Die “Dashboard-Pandemic” (Walker, 2023) dominierte als digitales räumliches Artefakt dabei das Monitoring der COVID-19-Pandemie genauso wie die Kartierung von Gewaltverbrechen den Blick der räumlichen Kriminologie leitet (Eck und Weisburd, 1995). Adressiert der räumlich-kartographische Zugriff aber über eine plausible Metapher der Kontrolle hinaus den zentralen Bedarf, wenn Prävention und Schutz der Bevölkerung im Mittelpunkt stehen? Im Gefolge des Spatial Turns in den Kultur- und Sozialwissenschaften (Günzel, 2018) sind Ansätze wie diese innerhalb der Digital Humanities übrigens eher die Regel als die Ausnahme, wenn z.B. in der Digitalen Geschichte nach der raumzeitlichen Diffusion von Ideen gefragt wird (Koller, 2016).

Bei der Frage nach der erlebten Unsicherheit im Alltag ist aus der konzeptionellen gesellschaftswissenschaftlichen Diskussion bekannt, dass es sich um einen diffusen sprachlichen Zugriff auf fluide Räume handelt (Redepenning et al., 2010), die weder als unsicher empfundene Mobilitätsformen noch eine sich ständig verändernde Lage vor Ort widerspiegeln (Moura de Souza et al., 2022). Selbst als während der COVID-19-Pandemie eine Durchdringung der Bevölkerung mit dem Virus schon gegeben war, spielte die Frage nach einer räumlichen Herkunft, den “Hotspots” noch immer eine entscheidende Rolle in der Suche nach Ursachen (Kremer und Felgenhauer, 2022), obwohl die Abschätzung eines individuellen Risikos situatives Wissen in einer wesentlich höheren Dynamik erfordert, als es durch den rein räumlichen Zugriff auf einer vergleichsweise hohen administrativen Aggregationsebene (z.B. Landkreise) abgebildet werden kann. Interessanterweise haben sich im Alltag Medienrepertoires ausgebildet, die diesen Bedarf in einer Mischung aus Social Media und klassischen unidirektionalen TV-Sendungen adressieren (Moura de Souza et al., 2022) und somit ggf. eine besser geeignete Datengrundlage für den sprachlichen Aushandlungsprozess von Krisen darstellen als deren Visualisierung anhand kartengebundener Daten. Figure 2 zeigt die resultierende Konzeptraumanalyse.

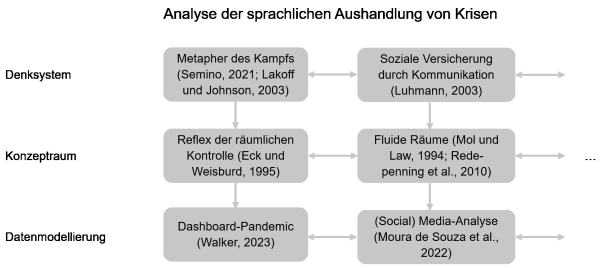


Abbildung 2: Beispiel für das Ergebnis einer vergleichenden Konzeptraumanalyse. Zwei unterschiedliche Theorieansätze führen zu unterschiedlichen konzeptionellen Zugängen (hier: Raumkonzepte), die sich in die konkrete Datenarbeit einschreiben. In der Zusammenschau ergeben sich so tiefere Antworten auf die gestellte Fragestellung. Zusätzliche Quellen: Lakoff und Johnson, 2003; Luhmann, 2003; Mol und Law, 1994.

Ein weiteres Beispiel: Im Kontext der Provenienzforschung zur NS-Raubkunst sind Lücken in den Besitzangaben zwischen 1933 und 1945 äußerst problematisch, da sie auf ungeklärte Unrechtskontexte hinweisen können. Sie müssen deshalb explizit markiert und erläutert werden (Lang, 2023). Im Moment gibt es noch kein Datenmodell, das eine semantische Modellierung von Lücken in Provenienzangaben erlaubt. Ein entsprechendes Projekt ist in Vorbereitung. Damit eine aussagekräftige Modellierung von Lücken stattfindet, muss in diesem Fall ein konzeptionelles Verständnis für die Entstehungsgründe und Arten von Lücken und zusätzliches Wissen über existierende Methodologie zur Datenmodellierung vorhanden sein. Folgende Anforderungen werden an das Modell gestellt: (1) Unterscheidung verschiedener Arten, Gründe und Kontexte, in welchen Lücken entstehen; (2) Abhängigkeiten verschiedener Lücken oder Verhältnis von Lücke, Zeit und Ort. Eine Modellierung von und Auseinandersetzung mit Lücken muss (3) soziale, räumliche und zeitliche Faktoren berücksichtigen und zudem fragen, (4) ob Lücken durch analoge Prozesse oder digitale Methoden generiert werden und wer entscheidet, was zur Lücke wird. (vgl. Daten als Ausdruck von Macht, vgl. Selwyn, 2021).

Ausblick

Gerade im Vergleich zu den Data Sciences stellt die theoriebezogene Artikulation von Forschungsfragen vor dem Hintergrund widerstreitender gesellschaftlicher Wirklichkeiten eine Stärke der Digital Humanities dar. Dies erfordert allerdings Übersetzungsvorgänge in doppelter Hinsicht: (1) Die Übersetzung von theoriegeleiteten Konzepten und Forschungsansätzen aus Humanities und Sozialwissenschaft in passende Datenschemata und Analysewerkzeuge zum Zweck einer digital gestützten Aufarbeitung und (2) die Erweiterung differenziert angelegter Metadatenschemata zur Erfassung, Pflege und Analyse von Daten zu Kunst- und Kulturgütern um solche Attribute, die die gesellschaftlichen Produktionsbedingungen und die zuschreibenden Akteure*innen hinter den Metadaten sichtbar, hinterfragbar und ebenfalls analysierbar machen. So entstehen nicht nur besser nutzbare Daten – sie eröffnen zudem spannende interdisziplinäre Forschungsperspektiven und Gelegenheit zur Methodeninnovation zwischen digitalen Kultur- und Sozialwissenschaften.

Bibliographie

Ahmad, Kashif, et al. 2022. “Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges.“ *Computer Science Review* 43: 100452.

Bekiari, Chryssoula et al. September 2021. ”Definition of the CIDOC Conceptual Reference Model.” Erarbeitet von der CIDOC CRM Special Interest Group, Version 7.2. https://cidoc-crm.org/Version/version-7.2.

Boyd, Danah, Kate Crawford. 2012. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.” In *Information, Communication & Society*, 15(5): 662-679. Doi: *10.1080/1369118X.2012.678878*.

Brinks, Verena, Oliver Ibert. 2020. “From corona virus to corona crisis: the value of an analytical and geographical understanding of crisis.“ In *Tijdschrift voor economische en sociale geografie* 111(3): 275–287. Doi: 10.1111/tesg.12428.

Bubenhofer, Noah. 2009. *Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse.* Berlin, New York: De Gruyter.

Chapman, Connor M., DeMond Shondell Miller. 2020. “From metaphor to militarized response: the social implications of “we are at war with COVID-19” – crisis, disasters, and pandemics yet to come.“ In *International Journal of Sociology and Social Policy* 40(9/10): 1107–1124.

Dalton, Craig, Jim Thatcher. 2014. ” What does a critical data studies look like, and why do we care?” In Society and Space (Magazine). https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care .

CIDOC CRM, https://www.cidoc-crm.org (abgerufen am 4. Dezember 2023).

Eck, John, David L. Weisburd. 1995. “Crime Places in Crime Theory.“ In *Crime and Place*, hg. von John Eck, David L. Weisburd, 1-33. Monsey, New York: Criminal Justice Press.

Glaser, Katja. 2017. *Street Art und neue Medien: Akteure – Praktiken – Ästhetiken.* Bielefeld: transcript Verlag. DOI: 10.14361/9783839435359.

Glasze, Georg. 2009. “Kritische Kartographie.“ *Geographische Zeitschrift* 97(4):181-191. https://www.jstor.org/stable/23031916 .

Günzel, Stephan. 2018. *Raum. Eine kulturwissenschaftliche Einführung.* Heidelberg, Berlin: JB Metzler.

Hackett, Edward J., Olga Amsterdamska, Michael Lynch, Judy Wajcman. 2008. *The handbook of*

science and technology studies. 3. Edition. Cambridge, Massachusetts: MIT Press.

Iliadis, Andrew, Federica Russo. 2016. "Critical data studies: An introduction." In *Big Data & Society* 3(2). https://doi.org/10.1177/2053951716674238.

Jannidis, Fotis, Hubertus Kohle, Malte Rehbein. 2017. *Digital Humanities. Eine Einführung.* Heidelberg, Berlin: JB Metzler.

Keller, Reiner. 2008. *Wissenssoziologische Diskursanalyse: Grundlegung eines Forschungsprogramms*. Berlin: Springer-Verlag.

Kitchin, Rob, Tracey P. Lauriault. 2014. "Towards critical data studies: Charting and unpacking data assemblages and their work." In The Programmable City Working Paper 2 . http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112 .

Koller, Guido. 2016. *Geschichte digital: historische Welten neu vermessen.* Stuttgart: Kohlhammer Verlag.

Kremer, Dominik, Blake Byron Walker. 2023. "Placing Wellbeing: Distant Reading Approaches for Exploratory Data Analysis of Geographies of Wellbeing." In *Geographical Research in the Digital Humanities. Spatial Concepts, Approaches and Methods. Reihe: Digital Humanities Research*, hg. von Finn Dammann & Dominik Kremer, 121-142. Bielefeld: transcript Verlag.

Kremer, Dominik, Blake Byron Walker. 2022. "Geodaten quantitativ, aber kritisch analysieren: die Methode der explorativen räumlichen Datenanalyse am Beispiel von COVID -19 in Brasilien." In *Handbuch Kritisches Kartieren*, hg. von Finn Dammann, Boris Michel, 307-324. Bielefeld: transcript Verlag.

Kremer Dominik, Tilo Felgenhauer. 2022. "Reasoning COVID-19: the use of spatial metaphor in times of a crisis." In *Humanities and Social Sciences Communications* 265 (9): 1-15. https://doi.org/10.1057/s41599-02.

Lakoff, George, Mark Johnson. 2003. *Metaphors we live by.* Nachdruck der Ausgabe von 1980. Chicago, Illionois: University of Chicago Press.

Lang, Sabine. 2023. ""Mind the Gap": Von Lücken in der Provenienzforschung und ihrer Präsenz im digitalen Raum." In *Book of Abstracts, Digital Humanities im deutschsprachigen Raum (DHd)*, 212-217. https://doi.org/10.5281/zenodo.7715420.

Locher, Hubert (Hg.). 2007. *Kunstgeschichte im 20. Jahrhundert. Eine kommentierte Anthologie.* Darmstadt: WBG.

Luhmann, Niklas. 2003. *Soziologie des Risikos.* Nachdruck der Ausgabe von 1991. Berlin, New York: De Gruyter.

Maganga, Matthew. 22.01.2023. "The AI Image Generator: The Limits of the Algorithm and Human Biases." In: *ArchDaily*. https://www.archdaily.com/992012/the-ai-image-generator-the-limits-of-the-algorithm-and-human-biases(zugegriffen: 6. Juli 2023).

Mayring, Philipp. 2016. *Einführung in die qualitative Sozialforschung.* Weinheim: Beltz.

Mol, Annemarie, John Law. 1994. "Regions, networks and fluids: anaemia and social topology." In *Social Studies of Science* 24(4): 641-671.

Moretti, Franco. 2013. *Distant Reading.* London, New York: Verso Books.

Moura de Souza, Cléssio, Dominik Kremer, Blake Byron Walker. 2022. "Placial-Discursive Topologies of Violence: Volunteered Geographic Information and the Reproduction of Violent Places in Recife, Brazil." In *ISPRS International Journal of Geo-Information* 11(10): 1-21.

Müller, Rebecca. 2023. *Die Vivarini: Bildproduktion in Venedig 1440 bis 1505.* Regensburg: Schnell & Steiner.

Noll, Thomas. 2011. "Ikonographie/Ikonologie" In *Metzler Lexikon Kunstwissenschaft,* hg. von Ulrich Pfisterer, 194-198. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-00331-7_76.

Redepenning, Marc, Henriette Neef, Edvânia Torres Aguiar Gomes. 2010. "Verflüssigende (Un-) Sicherheiten: Über Räumlichkeiten des Strassenhandels am Beispiel Brasiliens." In *Geographica Helvetica* 65(3): 207–216.

Rose, Gillian. 2001. *Visual methodologies: An introduction to the interpretation of visual materials.* Thousand Oaks, Kalifornien: Sage Publications.

Ross, Jeffrey Ian, Jeff Ferrell. 2016. *Routledge Handbook of Graffiti and Street Art*. London, New York: Routledge.

Schatzki, Theodore R. 2002. *The site of the social: A philosophical account of the constitution of social life and change*. University Park, Pennsylvania: Penn State University Press.

Selwyn, Neil. 2021. "What is Critical Data Studies?" In *Data Smart Schools.* https://data-smart-schools.net/2021/05/21/what-is-critical-data-studies/ (aufgerufen am 2.12.2023)

Semino, Elena. 2021. ""Not soldiers but fire-fighters"– metaphors and Covid-19." In *Health Communication* 36(1): 50–58. Doi: 10.1080/10410236.2020.1844989.

Semino Elena, John Heywood, Mick Short. 2004. "Methodological problems in the analysis of metaphors in a corpus of conversations about cancer." In *Journal of Pragmatics* 36(7):1271 1294. Doi: 10.1016/j.pragma.2003.10.013.

Street Art & Urban Creativity: Scientific Journal. Erscheint seit 2015. https://journals.ap2.pt/index.php/sauc/index(zugegriffen: 18. Juli 2023).

Taylor, Charles. 2004. *Modern social imaginaries.* Durham, North Carolina: Duke University Press.

Timpf, Sabine. 2002. "Ontologies of Wayfiding: a Traveler's Perspective." *Networks and Spatial Economics* 2: 9-33. https://doi.org/10.1023/A:1014563113112 .

Viehöver, Willy. 2001. "Diskurse als Narrationen." In *Handbuch Sozialwissenschaftliche Diskursanalyse: Band I: Theorien und Methoden*, hg. von Reiner Keller, Andreas Hirseland, Werner Schneider, Willy Viehöver, 177-206.

Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-99906-1.

Walker, Blake Byron. 2023. "Petrichor and Positionality: Occasion for a Situated Spatial Epidemiology in the Digital Humanities." In *Geographical Research in the Digital Humanities. Spatial Concepts, Approaches and Methods. Reihe: Digital Humanities Research,* hg. von Finn Dammann & Dominik Kremer, 40-51. Bielefeld: transcript Verlag.

Westerholt, René, Franz-Benjamin Mocnik, Alexis Comber. 2020. ”A place for place: Modelling and analysing platial representations.” *Transactions in GIS* 24(4): 811–818. DOI: 10.1111/tgis.12647.

Zhou, Kai-Qing, Hatem Nabus. 2023. “The Ethical Implications of DALL-E: Opportunities and Challenges.“ In *Mesopotamian Journal of Computer Science,* 17-23.

Lautstärke und Konflikt in Realismus und Naturalismus

Häußler, Julian
julian.haeussler@tu-darmstadt.de
Technische Universität Darmstadt, fortext lab, Deutschland
ORCID: 0000-0001-7490-8570

Guhr, Svenja
svenja.guhr@tu-darmstadt.de
Technische Universität Darmstadt, fortext lab, Deutschland
ORCID: 0000-0002-7686-3609

Gius, Evelyn
evelyn.gius@tu-darmstadt.de
Technische Universität Darmstadt, fortext lab, Deutschland
ORCID: 0000-0001-8888-8419

Realismus und Naturalismus im Vergleich

Der Naturalismus gilt als „radikale Form des Realismus“ (Fricke et al., 2000, S. 684). Die Autor:innen des Naturalismus widmeten sich den aus ihrer Sicht drängenden sozialen Fragen der Zeit. Mit den dadurch veränderten thematischen Schwerpunkten ging auch der Anspruch einher, wirklichkeitsnahe Literatur zu schreiben. Dies steht im Kontrast zu den Themen und poetologischen Prinzipien der parallel weiter existierenden Strömung des Realismus. Auf der Ebene der Textgestaltung lässt sich der Gegensatz der beiden Strömungen u. a. daran festmachen, dass Expositionen, also Anfänge von literarischen Texten, im Realismus typischerweise ereignisarm und gleichzeitig ausführlich gestaltet sind, während im Naturalismus das Prinzip des zeitdeckenden Erzählens (i. e., die Übereinstimmung von erzählter Zeit und Erzählzeit) wichtig wurde. Dieser sogenannte „Sekundenstil“ sollte als Erzähltechnik eine gewisse Unmittelbarkeit erzeugen. Neben der thematischen Neuausrichtung bzw. Zuspitzung des Naturalismus gibt es also zwischen Realismus und Naturalismus auch damit zusammenhängende Unterschiede in der sprachlichen Darstellung.

In unserem Beitrag untersuchen wir deutschsprachige Texte des Realismus und des Naturalismus auf Lautstärke und Konflikthaftigkeit. Wir nehmen an, dass die eben erläuterten poetologischen Prinzipien auch Auswirkungen auf die textliche Gestaltung der so genannten *soundscapes* (*sound* + *landscape*) (Schafer 1994) sowie von Konflikten in der Fiktion haben. Für die Analyse von Lautstärke und Konflikthaftigkeit haben wir in unterschiedlichen Projekten bereits Zugänge entwickelt (vgl. Guhr und Algee-Hewitt, 2023a und b und Häußler und Gius, 2023a und b), die an der sprachlichen Realisierung der Phänomene ansetzen, um sie zu identifizieren und zu kategorisieren.

Die Erkennung von Geräuschen des fiktionalen *soundscape* umfasst Umgebungs- sowie Figurengeräusche, die auf der Wortebene als in der Fiktion realisierte Geräusche annotiert werden. Die Untersuchung von Konflikten wiederum basiert konzeptuell auf interpersonellen Konflikten nach Glasl (2011) und erfasst Konflikthaftigkeit anhand von Schlüsselwörtern für Konflikt.

Mit diesen Methoden gehen wir im folgenden Beitrag auf die folgenden Hypothesen ein: (1) Aufgrund der grundsätzlich ablehnenden Haltung des Naturalismus gegenüber dem bürgerlichen Realismus erwarten wir zum einen, dass naturalistische Texte sowohl lauter als auch konflikthafter sind als realistische. (2) Zum anderen vermuten wir, dass die poetologischen Unterschiede der beiden Epochen sich auch auf eventuelle Zusammenhänge der beiden untersuchten Phänomene auswirken. Lautstärke und Konflikthaftigkeit im Realismus sollten eher negativ korrelieren, da dort Konflikte vermutlich eher abgeschwächt dargestellt werden, während im Naturalismus Konflikte auch lautstark ausgetragen werden und eine positive Korrelation zwischen Lautstärke und Konflikthaftigkeit bestehen dürfte.

Korpus

Zur Untersuchung dieser Hypothesen verwenden wir 192 Prosatexte aus dem Realismus und 69 Prosatexte aus dem Naturalismus. Es sind fiktionale Texte unterschiedlicher Länge, die in einschlägiger Sekundärliteratur dem Realismus und Naturalismus des deutschen Sprachraums zugeordnet werden (Böttcher und Geerdts, 1983; Brenner, 2011, Killy, 2016). Die Texte wurden dem deutschsprachigen

Prosakorpus *d-Prose 1870-1920* (Gius et al., 2020) entnommen und um Texte ergänzt, deren Urheberrecht nach 2020 erloschen ist oder die vor 1870 veröffentlicht wurden. Es sind Erzähltexte aus sogenannter Trivial- und Hochliteratur mit einem Mindestumfang von 2.000 Wörtern. Sie liegen als Reintextdateien vor und sind mit Metadaten zu Autor:innenname, Autor:innengender, Titel, Publikationsjahr, Dateiname, Epoche und Textlänge in Wörtern angereichert.

Tab. 1: Korpusbeschreibung.

	Realismus	Naturalismus
Summe der Korpustexte	192	69
Durchschnittliche Textlänge in Wörtern	39.116	33.470
längster Text	169.510	159.161
kürzester Text	2.773	2.713
Tokensumme	7.510.306	2.309.440
manuell annotiertes Testset in Wörtern	8.805	10.358

Erkennung von Geräuschen und ihrer Lautstärke

Die systematische Untersuchung von Geräuschen und ihrer Lautstärke ist ein Ansatz aus dem literaturwissenschaftlichen Anwendungsbereich der *sound studies*, die sich mit den Analysemöglichkeiten von Geräuschen in literarischen Texten beschäftigen. In unserer Studie analysieren wir Geräusche der fiktionalen Welt in literarischen Texten systematisch hinsichtlich ihrer Beschreibungen in der Narration (vgl. Schafer, 1994; Picker, 2003; Snaith, 2020). Die Erkennung und Analyse von Umgebungs- und Figurengeräuschen einer fiktionalen Welt eröffnet dabei eine neue Perspektive auf *scene setting* und Figurencharakterisierung. In einem heuristischen Verfahren weisen wir den erkannten Geräuschen auch ein Lautstärkelevel zu, indem wir auf der Grundlage von ermittelten Dezibelwerten faktualer Geräusche Geräusche in der Fiktion in Relation zueinander setzen und in Lautstärke-Levels gruppieren (Guhr und Algee-Hewitt, 2023b).

Für die Operationalisierung von Umgebungsgeräuschen und ihrer Lautstärke verwenden wir die in Guhr und Algee-Hewitt (2023b) fürs Englische entwickelte Geräuschdefinition mit ihrer Unterscheidung zwischen expliziten und interpretationsbedürftigen impliziten Geräuschbeschreibungen sowie die Methoden zur manuellen und lexikonbasierten Annotation von Geräuschen und übertragen diese auf deutschsprachige Prosa. Zum Beispiel wird im Satz „Der Zug fährt ratternd in den Bahnhof ein“ das ratternde Geräusch des einfahrenden Zuges explizit auf der Wortebene des literarischen Textes angegeben, sodass das Geräusch einer lexikalischen Einheit – hier dem Adverb „ratternd“ – zugeordnet werden kann. In der Geräuschannotation wird dieses Wort als Annotationseinheit betrachtet und lexikonbasiert um die Angabe eines Lautstärkelevels (1-5) von leisen bis sehr lauten Geräuschen erweitert.

Als ersten Schritt in Richtung einer Automatisierung der Geräuschannotation verwendeten wir einen lexikonbasierten Ansatz, in dem tokenisierte und lemmatisierte Korpustexte mit einem deutschsprachigen Geräuschwortlexikon abgeglichen wurden. Die Lexikoneinträge sind Schlüssel-Wertpaare von Geräuschwörtern, die einem nach Sachgruppen sortierten Wörterbuch (Dornseiff und Wiegand, 2004) entnommen (u. a. die Sachgruppen: „Geräusch“, „lautlos“, „Stimme“) und hinsichtlich der Dezibel-Heuristik mit Lautstärkelevels ausgezeichnet wurden (Geräuschlexikon = {Lemma : Geräuschlautstärkelevel}). Mithilfe eines String-Matching-Algorithmus und Lexikonabgleichs werden die Korpustexte automatisiert mit dem Tag „sound“ sowie einem Lautstärkelevel annotiert. Anschließend wird pro Korpustext und auch pro Korpus (Realismus- und Naturalismuskorpus) ein durchschnittlicher Lautstärkewert berechnet, der die annotierten Lautstärkelevels in Relation zur Anzahl der Geräuschwörter und der absoluten Anzahl an Wörtern pro Text abbildet.

Für die Evaluation wurden die Annotationen aus dem Lexikonansatz mit manuell erstellten Annotationen verglichen. Dafür wurden die bewährten Evaluationsmetriken *accuracy*, *precision*, *recall* und *F1-score* angewendet (s. Tab. 2). Die niedrigen Evaluationswerte deuten auf die Komplexität der Geräuschannotationsaufgabe hin. Außerdem ist zu vermuten, dass ein lexikonbasierter Ansatz ohne Einbezug des Kontexts die *soundscapes* der Fiktion nur bedingt erfassen kann. Für unser weiteres Vorgehen können diese Annotationen dennoch als Grundlage verwendet werden, weil die Geräuschwörter Hinweise auf in der Fiktion realisierte oder behandelte Geräusche geben. Im weiteren Vorgehen werden die Annotationstreffer des Lexikonansatzes manuell überprüft und um die *false positives* bereinigt, um so die *precision* der Ergebnisse zu steigern.

Tab. 2: Evaluation des Lexikonansatzes.

accuracy :	0.987
precision :	0.2
recall :	0.28
F1-score:	0.23

Messung von Konflikt

Die Analyse von Konflikthaftigkeit basiert auf der Anwendung und Adaption der im Projekt *SentiArt* entwickelten Sentimentanalyse (vgl. u. a. Jacobs, 2019). Dabei werden Sentimentwerte eines Wortes anhand eines *Word Embedding*-Modells berechnet, einem Verfahren der quantitativen Semantik, bei dem die Wortbedeutung mittels eines Ähnlichkeitsmaßes vergleichbar wird. Aus der Emotionsforschung entnommene Schlüsselwörter repräsentieren hierbei Pole für Sentimentdimensionen (z. B. ‚Angst‘ für den negativen Pol der emotionalen Valenz). Die anhand der Kosinusähnlichkeit gemessene Position eines Wortes zwischen diesen Polen entspricht dann einem bestimmten Sentimentwert. Dieser Ansatz der Sentimentanalyse wird hier adaptiert durch die Ersetzung der Sentiment-Schlüsselwörter durch Konflikt-Schlüsselwörter. Zur Bestimmung dieser Schlüsselwörter wurde aus einem nach Sachgrup-

pen sortierten Wörterbuch (Dornseiff und Wiegand, 2004) eine Auswahl an Wörtern jeweils zu den Sachgruppen ‚Konflikt' und ‚Harmonie' entnommen, um so gegensätzliche Pole zu bilden. Der positive Pol entspricht also dem Konzept ‚Konflikt', Schlüsselwörter dafür sind z.B. ‚töten', ‚Unglück' und ‚Gefahr'. Der negative Pol entspricht folgerichtig dem Konzept ‚Harmonie', repräsentiert z.B. durch die Schlüsselwörter ‚Glück', ‚leicht', ‚Lust' (für eine vollständige Übersicht der Schlüsselwörter s. Häußler und Gius, 2023a).

Diese Methodik wurde bisher genutzt, um Sentiment- und Konfliktwerte in Romantik-, Realismus-, und Naturalismus-Korpora zu erheben, mit dem Ziel konflikthafte Textstellen zu ermitteln und die Korpora im Hinblick auf ihre Konflikthaftigkeit zu betrachten. Darüber hinaus wurden Sentimentwerte erhoben, in der Erwartung, dass Sentiment ein Signal für Konflikt sein kann (vgl. Häußler und Gius, 2023a und b).

Für unseren Anwendungsfall wurden für das Realismus- und Naturalismuskorpus je ein *Word2Vec*-Modell erstellt und für jedes Wort ein Konfliktwert berechnet. Dabei wurde für jedes Wort die Kosinusähnlichkeit zu jedem der Schlüsselwörter berechnet und die durchschnittliche Ähnlichkeit zu den negativen Schlüsselwörtern von der zu den positiven Schlüsselwörtern subtrahiert. Mit den resultierenden Datensätzen aus den Wörtern und den dazugehörigen Konfliktwerten eines Textes können nun u. a. Verläufe der Werte innerhalb eines Textes betrachtet werden, etwa um Passagen mit extremen Werten auszumachen. Damit können wir nun auch das Auftreten von Geräuschwörtern in diesen Passagen betrachten, um zu ermitteln, ob es bei diesen Passagen auch zu einer Häufung von lauten Geräuschwörtern kommt (positive Korrelation von Lautstärke und Konflikt) oder ob konflikthafte Passagen auch ohne hohe Lautstärke auskommen (negative Korrelation). Des Weiteren können die Texte im Vergleich danach sortiert werden, wie stark sie welchen Wertebereich am ehesten vertreten (hier durch die Berechnung des durchschnittlichen Konfliktwertes eines Textes). Damit können Konflikt- und Lautstärkewerte auch auf Korpusebene verglichen werden.

Lautstärke und Konflikt im Vergleich

Um unsere Hypothesen zu überprüfen, dass (1) der Naturalismus durchschnittlich lauter und konflikthafter als der Realismus ist, und, dass (2) Konflikte im Naturalismus eher in laute bzw. im Realismus in leise *soundscapes* eingebettet sind, bringen wir im nächsten Schritt die beiden Verfahren zur Analyse von Konflikthaftigkeit und von Geräuschen inklusive Lautstärke zusammen. Dazu verglichen wir zunächst die Lautstärkewerte in den beiden Korpora, wobei wir das Realismus- und Naturalismuskorpus jeweils in lange und kurze Texte aufgeteilt betrachteten, um einen ggf. vorhandenen Textlängenbias abzuschwächen.

Mit Blick auf die dadurch erhaltenen Ergebnisse zeigt sich im Vergleich der Lautstärkewerte in den beiden Subkorpora zunächst keine auffälliger Unterschied zwischen Realismus und Naturalismus (vgl. Abb. 1). Betrachtet man jedoch die durchschnittlichen Konfliktwerte der einzelnen Texte, so fallen die Texte aus dem Naturalismus als konflikthafter auf (vgl. Abb. 2).

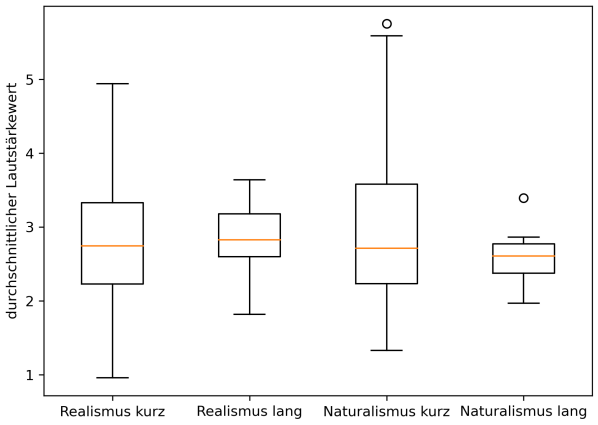


Abb. 1: Durchschnittlicher Lautstärkewert für alle kurzen (# 100.000 Wörter) und langen (>100.000 Wörter) Texte.

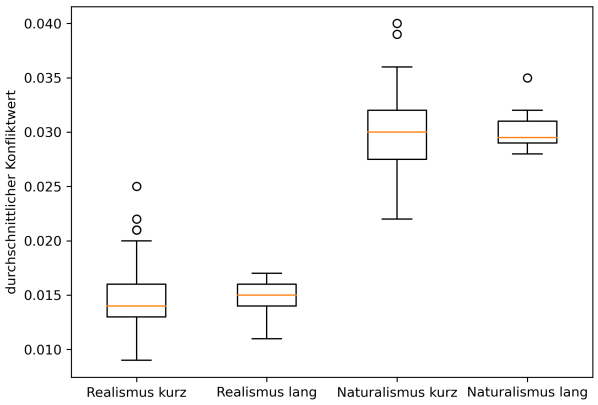


Abb. 2: Durchschnittlicher Konfliktwert für alle kurzen (# 100.000 Wörter) und langen (>100.000 Wörter) Texte.

In der Betrachtung der einzelnen Texte hinsichtlich ihrer Durchschnittswerte zeigt sich bei den lautesten bzw. leisesten Texte eine Gruppierung nach Autor:innen. So sind die drei durchschnittlich lautesten Texte des gesamten Korpus Texte der Naturalistin Clara Viebig (vgl. Tab. 4) und die drei konflikthaftesten Texte vom Naturalisten Ludwig Thoma. Der Realist Theodor Fontane fällt zudem als wenig „konflikthafter" Autor auf (vg. Tab. 4).

Tab. 3: Fünf lauteste Texte und leiseste Texte im Gesamtkorpus mit durchschnittlichem Lautstärkewert und Epochenkontext in Klammern.

Rang	Lauteste Texte	Leiseste Texte
1	Viebig: *Die Cigarrenarbeiterin* (5.75, Nat.)	von Saar: *Außer Dienst* (0.96, Real.)
2	Viebig: *Der Osterquell* (5.59, Nat.)	Thoma: *Onkel Peppi. Anfänge* (1,33, Nat.)
3	Viebig: *Am Totenmaar* (5.02, Nat.)	Fontane: *Meine Kinderjahre* (1.38, Real.)
4	Riehl: *Musiker-Geschichten. Demophoon von Vogel* (4.94, Real.)	von Polenz: *Luginsland* (1.39, Nat.)
5	Viebig: *Die Schuldige* (4.78, Nat.)	von Saar: *Die Pfündner* (1.54, Real.)

Tab. 4: Fünf konflikthafteste und am wenigsten konflikthafte Texte im Gesamtkorpus, durchschnittlicher Konfliktwert und Epochenzuweisung in Klammern.

Rang	Konflikthafteste Texte	Am wenigsten konflikthafte Texte
1	Thoma: *Nachbarsleute. Das alte Recht* (0.040, Nat.)	von Saar: *Der Hellene* (0.009, Real.)
2	Thoma: *Andreas Voest* (0.039, Nat.)	Fontane: *Mathilde Möhring* (0.009, Real.)
3	Thoma: *Die Fahnenweihe* (0.036, Nat.)	Fontane: *Die Poggenpuhls* (0.009, Real.)
4	von Suttner: *Die Waffen nieder* (0.035, Nat.)	Fontane: *Frau Jenny Treibel* (0.010, Real.)
5	Sudermann: *Der Katzensteg* (0.034, Nat.)	Fontane: *Irrungen* (0.010, Real.)

Um die Korrelation von Lautstärke und Konflikt zu prüfen, betrachten wir als nächstes, welchen Texten sowohl extreme Lautstärke-, als auch extreme Konfliktwerte zugewiesen wurden. Wir definieren hier als extrem die obersten bzw. untersten 10% in der Rangordnung der lautesten bzw. konflikthaftesten Texte und heben jene Texte hervor, die in überschneidenden extremen Gruppen auftreten (vgl. Tab. 5). Zwar bestätigen die Ergebnisse die Vermutung, dass die eher konflikthaften auch die eher lauten Texte sind (Thoma) und, dass die eher weniger konflikthaften auch die eher leisen Texte sind (Fontane), doch zeigt sich für den Naturalisten Thoma, dass auch eine starke negative Korrelation (hohe Konflikthaftigkeit, geringe Lautstärke) im Naturalismus vertreten ist.

Tab. 5.: Texte der Extremgruppen (alphabetische Sortierung innerhalb der Gruppen)

laute und konfliktarme Texte • Riehl: *Musiker-Geschichten. Demophoon von Vogel* (Real.) • Riehl: *Musiker-Geschichten. Gradus ad Parnassum* (Real.)	laute und konflikthafte Texte • Thoma: *Der vornehme Knabe* (Nat.) • Thoma: *Franz und Cora* (Nat.) • Thoma: *Nachbarsleute. Bismarck* (Nat.) • Viebig: *Das Weiberdorf* (Nat.)
leise und konfliktarme Texte • Fontane: *Die Poggenpuhls* (Real.) • Fontane: *Effi Briest* (Real.) • Fontane: *Irrungen, Wirrungen* (Real.) • Fontane: *Mathilde Möhring* (Real.) • Fontane: *Meine Kinderjahre* (Real.) • Raabe: *Kloster Lugau* (Real.) • von Saar: *Die Parzen* (Real.)	leise und konflikthafte • Thoma: *Onkel Peppi. Die Eigentumsfanatiker* (Nat.) • von Polenz: *Luginsland. Das Glück der Riegels von Petersgrün* (Nat.) • von Polenz: *Luginsland. Ein wilder Schoessling* (Nat.)

Um Konflikt und Lautstärke sowie deren Korrelation auch qualitativ zu betrachten, analysieren wir im Folgenden den Text *Der vornehme Knabe* von Ludwig Thoma (vgl. Abb. 3). Die Kurzgeschichte handelt vom gemeinsamen Spiel zweier Knaben, wobei der Protagonist das Modellschiff des anderen Knaben auf einem Weiher zur Explosion bringt. Der Protagonist entkommt, während der andere Knabe vom Besitzer des Weihers (Rafenauer) verprügelt wird. Der Konflikt zwischen den Knaben wird nicht gelöst, sie gehen im Streit auseinander.

Hinsichtlich der Lautstärke- und Konfliktwerte fallen höhere Konfliktwerte und mehr Lautstärkewörter in der zweiten Hälfte des Textes auf. Zwar entspricht die Eskalation von verbaler zu physischer Gewalt dem Höhepunkt der Konfliktwerte, doch treten z.B. Beleidigungen auch mit unterschiedlicher Konflikthaftigkeit auf. Eine Eskalation des Konfliktes korreliert hier mit einer Häufung der Lautstärkewörter, doch sind die Lautstärkewörter mit dem höchsten Wert nicht gleichbedeutend mit dem Höhepunkt der Eskalation. Die Knaben sprechen vor der eigentlichen Explosion darüber, ob Modellschiffe auch Munition verschießen können (‚schießt', ‚schießen' bzw. ‚knallen'). Arthur fragt z.B. den Protagonisten, ob der Verschuss der Munition „recht knallen wird". Die Eskalation kündigt sich dann mit der Erwartung der Explosion an (‚knallt'). Die Explosion selbst wird mit Worten beschrieben, die nicht den Lautstärkewert 5 besitzen bzw. nicht im Geräuschwortlexikon enthalten sind. Erst in den Momenten physischer Gewalt, den Ohrfeigen des Weiherbesitzers, knallt es wieder. Dieser unterstellt ihnen zudem, sein Haus sprengen zu wollen.

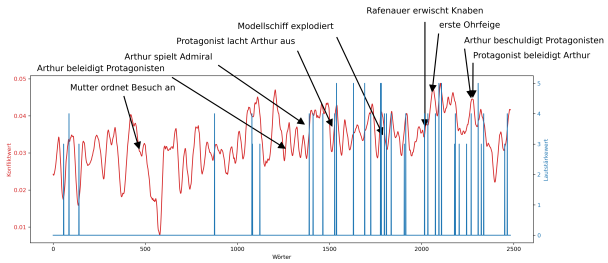


Abb. 3: Konfliktwerte (Kosinusglättung, Fenstergröße 40) und Geräuschwörter+ Lautstärkewerte in Thomas *Der vornehme Knabe.*

Fazit und Ausblick

Durch unsere Untersuchungen konnten wir zeigen, dass mit diesem ersten Ansatz zur Analyse von Konflikthaftigkeit und Geräuschlautstärke in fiktionalen Texten bereits einige vergleichende Rückschlüsse auf Texte des Realismus und Naturalismus getroffen werden können. Wir konnten herausstellen, dass naturalistische Texte auffällig konflikthafter sind (vgl. Tab. 4), während sie sich hinsichtlich ihrer Lautstärke kaum von Texten des Realismus unterscheiden (vgl. Tab. 3).

Interessant ist, dass sich mit Blick auf die Texte mit den höchsten/niedrigsten durchschnittlichen Lautstärkewerten und den höchsten/niedrigsten Konfliktwerten Autor:innencluster herausstellen lassen. Dabei entspricht die Beobachtung der Hypothese 1), da die lautesten Texte (Viebig) und konflikthaftesten Texte (Thoma) dem Naturalismuskorpus zugehören, während die am wenigsten konflikthaften Texte realistisch sind (Fontane). Die Verschränkung von Lautstärke und Konflikthaftigkeit (vgl. Hypothese 2) scheint zwar einerseits zu bestätigen, dass die konflikthaften Texte von Thoma (Naturalismus) auch laut sind (positive Korrela-

tion), doch trifft diese Beobachtung nicht auf alle Texte des Autoren (hier: *Onkel Peppi. Die Eigentumsfanatiker*) sowie die allgemeine Vermutung zu, da die auffällig konfliktharften und leisen Texte allesamt naturalistisch sind. Dadurch stellt sich für weiterführende Analysen einerseits die Frage, inwiefern Geräuschbeschreibungen und Konfliktdarstellung in literarischen Texten stilistische Unterschiede zwischen Autor:innen vorweisen und andererseits, ob und wie die Vermutung der positiven bzw. negativen Korrelation in Naturalismus bzw. Realismus auf Korpuseben untersucht werden kann.

Bei der Analyse des Beispieltextes (*Der vornehme Knabe*) war auf den ersten Blick eine leichte positive Korrelation zwischen hoher Lautstärke und Konflikthaftigkeit ersichtlich. Bei der genauen Lektüre des Textes, zeigt sich, dass hier tatsächlich ein Konflikt in einer lauten *soundscape* eingebettet ist (wie für den Naturalismus erwartet wurde).

Bibliographie

Böttcher, Kurt und Hans Jürgen Geerdts. 1983. *Kurze Geschichte der deutschen Literatur*. Berlin: Volk und Wissen. Volkseigener Verlag.

Brenner, Peter J. 2011. *Neue deutsche Literaturgeschichte: vom „Ackermann“ zu Günter Grass*. 3., überarb. und erw. Aufl. Berlin; New York: De Gruyter.

Dornseiff, Franz und Ernst Wiegand. 2004. *Der Deutsche Wortschatz Nach Sachgruppen*. 8., völlig neu bearbeitete und mit einem vollständigen alphabetischen Zugriffsregister versehenen Aufl. Berlin; New York: De Gruyter.

Fricke, Harald, Klaus Weimar, Klaus Grubmüller und Jan-Dirk Müller, Hrsg. 2000. *Reallexikon der deutschen Literaturwissenschaft*. Berlin: de Gruyter.

Gius, Evelyn, Svenja Guhr und Benedikt Adelmann. 2020. „d-Prose 1870-1920“. In *Zenodo* 10.5281/zenodo.4315209.

Glasl, Friedrich. 2011. *Konfliktmanagement: Ein Handbuch für Führungskräfte*, Beraterinnen und Berater. Bern: Haupt.

Guhr, Svenja und Mark Algee-Hewitt. 2023a. “On the Relation of Sound and Suspense in Literary Fiction.” In *Book of Abstracts. DH2023 Graz*, 166-167.

Guhr, Svenja und Mark Algee-Hewitt. 2023b. “What’s that Scary Sound? Ambient Sound in Gothic Fiction.” In *Conference Reader. CCLS 2023 Würzburg*, 256-280.

Häußler, Julian und Evelyn Gius. 2023a. “Operationalizing and Measuring Conflict in German Novels.” In *CHR 2023: Computational Humanities Research Conference 2023*, 426-440.

Häußler, Julian und Evelyn Gius. 2023b. “Towards a Conflict Heuristic. Detecting Conflict in Literary Texts by Adapting Word Embedding Based Sentiment Analysis.” In *Book of Abstracts. DH 2023 Graz*, 213-214.

Jacobs, Arthur M. 2019. “Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics.” *Frontiers in Robotics and AI* 6 53. 10.3389/frobt.2019.00053.

Killy, Walther. 2016. *Killy Literaturlexikon: Autoren und Werke des deutschsprachigen Kulturraums*. Darmstadt: WBG.

Picker, John M. 2003. *Victorian Soundscapes*. New York: Oxford University Press.

Schafer, R. Murray. 1994. *The soundscape: our sonic environment and the tuning of the world*. Rochester: Destiny Books.

Snaith, Anna, Hrsg. 2020. *Sound and literature*. Cambridge, United Kingdom; New York, NY: Cambridge University Press.

»LLMs for everything?« Potentiale und Probleme der Anwendung von In-Context-Learning für die Computational Literary Studies

Pichler, Axel

axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID: 0000-0002-9177-7645

Reiter, Nils

nils.reiter@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0003-3193-6170

Große Sprachmodelle, sogenannte Large Language Models (LLMs), haben das Natural Language Processing (NLP) seit dem Aufkommen der Transformer-Architektur in den letzten Jahren revolutioniert. Spätestens seit der Veröffentlichung von ChatGPT ist das Potential dieser Modelle auch der nicht akademischen Öffentlichkeit bekannt. Ein noch nicht vollständig erklärtes Merkmal dieser Modelle ist, dass sie mit zunehmender Größe – als Schwellenwert werden hier um die 10 Milliarden Parameter genannt, – auch Problemlösungskompetenzen entwickeln, für die sie nicht trainiert wurden (Wei et. al. 2022). Zu diesen sogenannten ›Emergent Abilities‹ zählt auch eine Trainingsmethode, bei der es sich im strengen Sinne gar nicht um eine ›klassische‹ Form des Fine-Tunings handelt, da dabei keine Anpassungen der Gewichte durchgeführt werden: das In-Context-Learning (ICL, Dong et al. 2023).

Darunter versteht man die Praxis, einem LLM durch die Eingabe von natürlichsprachlich verfassten Beispielen, das in diesen Beispielen inkorporierte und implizierte ›Wissen‹

zu vermitteln.[1] Diese Beispiele werden dann mit einer Aufgabe bzw. Frage zu einem sogenannten ›Prompt‹ zusammengeführt und dem Modell zur Vorhersage eingegeben.

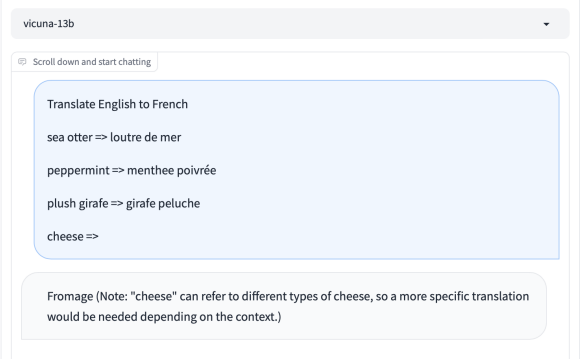


Abbildung 1: Das Few-Shot-Beispiel aus Brown et al. 2020 als Prompt in Vicuna-13b auf https://chat.lmsys.org

Wie bereits Brown et. al. (2020) für GPT-3 zeigten, können LLMs eine Vielzahl komplexer Aufgaben mithilfe von ICL lösen. Im Detail noch nicht geklärt sind die Gründe, warum sie das tun. Jüngere Untersuchungen lassen vermuten, dass dabei die Tatsache, dass die verwendeten Beispiele plausibel bzw. wahr für die Aufgabe sind, weniger wichtig ist, als andere Faktoren wie zum Beispiel die zugrundeliegende Verteilung der Beispiele bzw. deren Format (Min et al. 2022) oder die über die Trainingsdaten implizit vermittelten semantischen Relationen von Begriffen (Xie et al. 2021). Clavíe et. al. 2023 zeigen zum Beispiel, dass bei der binären Klassifikation der Qualifikationsvoraussetzungen für eine Stellenausschreibung große LLMs wie OpenAIs text-davinci-003-Model klassische ML-Ansätze wie SVM aber auch kleinere ›foundational models‹ wie DeBERTaV3 klar übertreffen.

Für die Digital Humanities im Allgemeinen und die Computational Literary Studies (CLS) im Besonderen ist das ICL auf den ersten Blick sehr attraktiv, da es *erstens* mit natürlichsprachlich verfassten Prompts arbeitet und daher weder profunde Programmier- noch Detailkenntnisse über die Modellierungspraxis von LLMs voraussetzt. Die Tatsache, dass mit natürlichsprachlich verfassten Eingaben gearbeitet wird, legt *zweitens* nahe, dass beim ICL – ohne größeren Operationalisierungs-Aufwand – die traditionellen Begriffsumgangs- und Definitionspraktiken der Geisteswissenschaften in das NLP bzw. die DH importiert werden können. *Drittens* ist das ICL nicht mit dem Zeitaufwand, den bekanntermaßen das (manuelle) Erstellen von Trainingsdaten mit sich bringt, verbunden.

Wir wollen im Folgenden das Potential von ICL an einem konkreten Beispiel aus den CLS überprüfen. Dabei handelt es sich um den Versuch, die Resultate der Operationalisierung und Modellierung von generischen Aussagen aus Andrew Pipers Cambridge Element *Can We Be Wrong?* aus dem Jahr 2020 zu reproduzieren bzw. zu übertreffen. Ziel von Pipers Operationalisierung und Modellierung ist es ›textempirisch‹ zu überprüfen, welche Rolle Generalisierungen in den *Literary Studies* spielen. Zur Beantwortung dieser Frage entwickelt er einen Workflow, der in groben Zügen mit jener Arbeitsablaufpraxis übereinstimmt, die wir zeitgleich im Rahmen von CRETA entwickelt haben und als ›reflektierte Textanalyse‹ bezeichnen (Reiter/Pichler 2020): Ausgehend von der besagten Frage erarbeiteten Piper und sein Team eine Operationalisierung des Konzepts von ›generalisierenden Aussagen‹ und von deren Subkonzepten, erstellten im Zuge dessen ein Annotationsschema, verbesserten dieses iterativ, um abschließend darauf unterschiedliche ML-Modelle zu trainieren (Piper 2020, 17–21). Im Rahmen des Operationalisierungsprozesses gelangten Piper und sein Team zu einer Bestimmung von *generalization*, die das Konzept als ein externalistisch zu validierendes linguistisches Phänomen begreift: »[G]eneralization is something that is at once linguistically legible at the statement level – it has certain criteria or qualities – and is also ambiguous. […] It depends on real-world knowledge of the generality of the terms being used with respect to some context and the definitional certainty that links subject and predicate.« (Piper 2020, 27). Der Annotations- und Modellierungsprozess gestaltete sich laut Piper wie folgt (Piper 2020, 32–34): In einem ersten Annotationslauf wurden 116 Sätze aus Artikeln, die nicht im späteren Trainingsdatensatz enthalten waren, von Piper und drei anderen Mitarbeitern des Teams annotiert und im Anschluss daran gemessen, wie Pipers Annotationen mit dem Mehrheitsvotum der anderen drei Kommentatoren übereinstimmten. Im Anschluss wurden dann von Piper an die 3500 weitere Sätze auf Basis der Annotationsrichtlinen annotiert. Diese Annotationen wurden abschließend dazu verwendet, um unterschiedliche Machine-Learning-Modelle zu trainieren. Dabei wurden die manuell noch separat annotierten Kategorien *generalization* und *exemplification* zu einer einzigen Kategorie zusammengefasst, da von letzteren bei der manuellen Annotation nur wenige Instanzen gefunden wurden. Es handelte sich also letztendlich um eine binäre Klassifikationsaufgabe. Wir fokussieren uns im Folgenden auf die Resultate dieser Modelle und vernachlässigen aus Platzgründen die ebenfalls höchst relevante Diskussion der Konsequenzen, die Piper aus der Anwendung dieser Modelle gezogen hat.

Die von Piper und seinem Team auf den annähernd ausgeglichenen Daten trainierten Modelle erzielten F1-Scores zwischen 0.591 und 0.769 sowie Accuracy-Werte zwischen 0.638 und 0.745, wobei es sich bei dem am besten performenden Modell um ein CNN mit ELMo-Embeddings handelt, bei dem der Recall die Precision deutlich übersteigt (Piper 2020, 34). Für unsere Experimente haben wir mit OpenAIs kostenpflichtigem[2] *text-davinci-003*-Model gearbeitet, das von OpenAI zum Zeitpunkt der Durchführung unsrer Experimente für diese Zwecke empfohlen wurde, und haben für dieses, nach einer kurzen Explorationsphase, 11 unterschiedliche ICL-Templates entwickelt.[3] Diese umfassten sowohl sogenannte Zero-shot-prompts, d.i. die Eingabe der Aufgabenbeschreibung ohne Beispiele, als auch unterschiedliche Arten von Few-shot-prompts, also Eingaben, die Beispiele mit oder ohne weitere kontextuelle Infor-

mationen umfassen. Unsere mit diesen Templates erzielten Resultate bewegen sich im Mittelfeld der Resultate von Piper und seinem Team, mit dem höchsten F1-Score von 0.69 und einer Accuracy von ebenfalls 0.69 mit einem Template, dass vier Beispiele mit der Instruktion »Determine the class of the incoming sentence as 'generalization' or 'neutral' on the base of the following examples« verbindet.

Tabelle 1: Klassifikationsergebnisse in Piper (2020) und Experimente mit einem LLM

	Modell	F1-Score	Accuracy
Piper (2020)	cnn + ELMo	0.769	0.745
	bilstm + ELMo	0.757	0.741
	stacked bilstm + ELMo	0.750	0.729
	lstm + ELMo	0.742	0.724
	bert	0.736	0.703
	cnn (GloVe)	0.696	0.696
	stacked bilstm (GloVe)	0.665	0.669
	lstm (GloVe)	0.595	0.641
	bilstm (GloVe)	0.591	0.638
Diese Arbeit	Eingabe des zu klassifizierenden Satzes + Aufforderung, ihn einer der beiden Kategorien zuzuordnen (zero_shot)	0.620	0.644
	Wie zero_shot, aber mit dem Zusatz »think step-by-step« (zero_shot_reason)	0.621	0.644
	Eingabe von vier Beispielen und ihrer Klassifikation + zu klassifizierender Satz (few_shot)	0.574	0.481
	Struktur wie few_shot, aber einschließlich der Erläuterung der Aufgabe (few_shot_inst)	0.678	0.678
	Struktur wie few_shot_inst, aber einschließlich der Erläuterung der Rolle des Models (few_shot_inst_role)	0.684	0.686
	Struktur wie few_shot_inst_role, aber einschließlich einer Beschreibung der beiden Klassen (few_shot_inst_role_exp)	0.659	0.659
	Struktur wie few_shot, aber andere Beispiele (few_shot_2)	0.649	0.596
	Struktur wie few_shot_inst, aber andere Beispiele (few_shot_inst_2)	0.691	0.696
	Struktur wie few_shot_inst_role, aber andere Beispiele (few_shot_inst_role_2)	0.669	0.683
	Struktur wie few_shot_inst_role_exp, aber andere Beispiele (few_shot_inst_role_exp_2)	0.598	0.642

Das beste ICL-Verfahren erzielt somit eine um 5-7 Prozentpunkte niedrigere Performance als das beste von Piper beschriebene Modell. Im Gegensatz zu Beispielen aus anderen Feldern zeigt sich also hier keine wesentlich bessere Performance als bei der Arbeit mit kleineren ›foundational models‹ wie z.B. BERT. In diesem konkreten Fall erachten wir unter anderem folgende Möglichkeiten als plausible Ursachen dafür: Erstens ist die Explikation der Unterscheidung zwischen ›generalization‹ und ›neutral‹ im allgemeinen Sprachgebrauch nicht üblich – man spricht zwar von generalisierenden Aussagen, bezeichnet aber gemeinhin nicht sämtliche Aussagen, die nicht unter diese Klasse fallen als ›neutral‹. Pipers theoretisch durchweg gerechtfertigtes Klassifikationsschema wird somit vom Sprachgebrauch nicht gestützt.[4] Folgt man Min et al. in ihrer – durchweg spekulativen – Hypothese, dass ICL umso besser funktioniert, je mehr es auf in den Trainingsdaten des LLMs bereits gegebene kategorische Differenzierungen aufbauen kann, könnte deren potentielles Fehlen in Letzteren die verhältnismäßig niedrigen Scores erklären. Zweitens besteht bei der Beispielauswahl durchweg noch Spielraum. Wir haben uns bei den Experimenten auf jene Beispiele konzentriert, die Piper selbst im Text seiner Monographie verwendet und die wir daher als exemplarisch erachteten. Wir haben jedoch weder auf bekannte Sampletechniken bei der Beispielauswahl zurückgegriffen noch die von Piper und seinem Team annotierten Daten im Detail auf ihre Repräsentativität manuell überprüft. Ob, und wenn ja welche, Performance-Gewinne derartig möglich sind, wäre zu klären.

Ergänzend zu diesen konkreten Fragen zur verhältnismäßig schwachen Performance von ICL in Hinblick auf Pipers Daten wollen wir auch noch auf weitere potentielle Problemfelder und offene Fragen in Hinblick auf den Einsatz von In-Context-Learning in den CLS hinweisen. Dazu zählt, erstens, die prinzipielle Gefahr, dass das ICL durch seinen Fokus auf Beispiele dazu einlädt, Begriffe undefiniert und unreflektiert zu verwenden. Wenn, wie in unserem Fall, die besten Resultate mit jenem Prompt erzielt werden, der keine Definition der verwendeten Begriffe beinhaltet, lädt dies dazu ein, auf die Bestimmung dieser Begriffe von Anfang an zu verzichten. Die problematischen Konsequenzen eines solchen Vorgehens liegen auf der Hand: Ohne die Begriffe definiert zu haben, läuft ein re-import der Resultate in den fachspezifischen Diskurs Gefahr, deren Umfang zu verunklaren, da die bloße Nennung von Beispielen unterschiedliche Interpretationen von der Extension dieser Begriffe zulassen. Eine ähnliche Gefahr besteht jedoch, zweitens, auch wenn der Begriff vor und für das ICL definiert wird, da die Mechanismen hinter selbigen noch nicht geklärt sind. Bei einem Prompt, der sich aus Definition, Instruktion und Beispiel zusammensetzt, wissen die Nutzenden nicht, welche der drei Komponenten für die Klassifikation letztendlich ausschlaggebend ist. Ob es tatsächlich die dabei verwendete Definition ist, bleibt unklar. Dies führt, drittens, zu einem weiteren prinzipiellen Problem beim Einsatz von kommerziellen LLMs, das hinlänglich bekannt ist: Kommerzielle Anbieter wie OpenAI stellen ihre Modelle nicht öffentlich zur Verfügung. Die per se bereits breit diskutierte vermeintliche Opazität von LLMs wird so noch zusätzlich verstärkt. Viertens sind LLMs wie das hier verwendete text-davinci-003-Model von OpenAI nicht deterministisch. Die Resultate sind dementsprechend nicht stabil.

In den CLS wird die Pflicht, dass man sich im Zuge des Operationalisierungs- bzw. Annotationsprozesses festlegt (welche Kategorien man wann vergibt, was diese bedeuten, wo Annotationen anfangen und aufhören, etc.) oft als Vorteil von computergestützten Verfahren gegenüber der ›traditionellen‹ Literaturwissenschaft genannt (z.B. Meister 1995), da deren Begriffe »in der Regel zu vage oder zu abstrakt [seien], als dass man sie eindeutig formalisieren könnte« (Meister 2012, 294). Die insbesondere von Harald Fricke seit mehreren Jahrzehnten propagierte Auffassung, dass literaturwissenschaftliche Begriffe ausgehend

vom standardsprachlichen Gebrauch zu präzisieren seien, um durch die solcherart hergestellte Exaktheit Vagheiten und Mehrdeutigkeiten aus dem literaturwissenschaftlichen Sprachgebrauch zu tilgen (Fricke 1989), bildet zwar mittlerweile das sprachtheoretische Fundament des *Reallexikons der deutschsprachigen Literaturwissenschaft*, scheint aber – wie Meisters Zitat nahelegt – die Praxis im Fach immer noch nicht zu dominieren. Mit der Verwendung von LLMs entfällt auch in den CLS diese Pflicht wieder, dementsprechend ungenau und unscharf werden potentiell abermals die Begrifflichkeiten.

Für den Einsatz von ICL in den CLS bedeutet das unseres Erachtens Folgendes: Erstens sollte man, unabhängig davon auf welches Sprachmodell man bei der Textanalyse zurückgreift, die für die Analyse zentralen Begriffe definieren und – idealerweise – manuell einen Referenzdatensatz erstellen. Dies erlaubt es, auch opake Modelle auf eine Art und Weise empirisch zu verankern, die den Nachvollzug sowie die Überprüfung der Validität der Analysen erleichtert bzw. in manchen Fällen überhaupt erst ermöglicht. Zweitens sollte man, falls man sich für den Einsatz von ICL entscheidet, zuerst mit kleineren Samples arbeiten, um zu überprüfen, ob das ICL überhaupt traditionelle Verfahren übertrifft: Bei Begriffen, deren Definitionen sich vom Alltagsgebrauch unterscheiden, ist das Fine-Tuning eines *pretrained Language Models* (PLMs) wie BERT ggf. zielführender. Drittens sollte man die fragestellungspezifische Leistung von kommerziellen LLMs der großen Techgiganten mit derjenigen von OpenSource-Modellen vergleichen.[5] Dies spart nicht nur Geld und Ressourcen, sondern liegt aus wissenschaftsethischen Gründen nahe, legen doch Modelle wie zurzeit zum Beispiel Stanford Alpaca[6] oder Vicuna[7] bei einer den kommerziellen Anbietern nahen Performance sowohl ihren Quellcode als auch Ihre Trainingsdaten offen.

Fußnoten

1. Durch die Verwendung der Anführungszeichen wollen wir hier darauf verweisen, dass es sich bei besagtem ›Wissen‹ unseres Erachtens nicht primär um propositionales Wissen im Sinne der Erkenntnistheorie handelt und dass LLMs nicht als rationale und damit epistemisches Wissen besitzende Agenten verstanden werden, sondern dass sie Implizites Wissen – häufig in Bezug auf den habituellen Sprachgebrauch – stochastisch reproduzieren. Zur laufendenden Debatte, wie und was LLMs ›verstehen‹, und deren primär inferentielle Semantik siehe Søgaard 2022.
2. Die Gesamtkosten beliefen sich auf ca. 260 US-Dollar.
3. Die Daten, das Python-Skript sowie ein jupyter notebook zur eigenen Exploration finden sich auf: https://github.com/nilsreiter/dhd2024-few-shot
4. So lautet das erste Exemplum eines neutralen Satzes in den annotierten Daten von Piper und seinem Team: »To this end, one of the main merits of Merleau-Ponty's framing of cinema as art is that it is not wedded to celluloid film and the oft-discussed reality effect its highly indexical-iconic images.« Die annotierten Sätze findet man hier: https://doi.org/10.6084/m9.figshare.12669329.v1
5. Als Ausgangspunkt eines solchen Vergleiches bietet sich die Gegenüberstellung der Score auf dem LLM Leaderboard von HuggingFace – https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard – mit den publizierten Scores der kostenpflichtigen Modelle an. Die dort verwendeten Vergleichsmetriken sind jedoch selbst in ihrer projektspezifischen Relevanz in Hinblick auf die jeweils gegebene Fragestellung zu evaluieren.
6. https://crfm.stanford.edu/2023/03/13/alpaca.html
7. https://lmsys.org/blog/2023-03-30-vicuna/

Bibliographie

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, u. a. 2020. „Language Models are Few-Shot Learners“. .

Clavié, Benjamin, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, und Thomas Brightwell. 2023. „Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification“. .

Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, und Zhifang Sui. 2023. „A Survey on In-context Learning“. .

Fricke, Harald. 1989. „Einführung“. In *Zur Terminologie der Literaturwissenschaft*, herausgegeben von Christian Wagenknecht, 1–9. Metzler.

Meister, Jan Christoph. 1995. Consensus ex Machina? Consensus qua Machina! Literary and Linguistic Computing, 10(4):263–270.

Meister, Jan Christoph. 2012. „Computerphilologie vs. ‚Digital Text Studies‘“. In *Literatur und Digitalisierung*, herausgegeben von Christine Grond-Rigler und Wolfgang Straub, 267–96. De Gruyter. https://doi.org/10.1515/9783110237887.267 .

Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi und Luke Zettlemoyer. 2022. „What makes In-Context-Learning work?“.

Pichler, Axel, und Nils Reiter. 2020. „Reflektierte Textanalyse“. In *Reflektierte algorithmische Textanalyse*, herausgegeben von Nils Reiter, Axel Pichler, und Jonas Kuhn, 43–60. De Gruyter. .

Piper, Andrew. 2020. *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*. 1. Aufl. Cambridge University Press. .

Søgaard, Anders. 2022. „Understanding Models Understanding Language“. *Synthese* 200 (6): 443. https://doi.org/10.1007/s11229-022-03931-4.

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, und Quoc V. Le. 2021. „Finetuned Language Models Are Zero-Shot Learners“. *CoRR* abs/2109.01652. .

Xie, Sang Michael, Aditi Raghunathan, Percy Liang, und Tengyu Ma. 2021. „An Explanation of In-context Learning as Implicit Bayesian Inference“. https://doi.org/10.48550/ARXIV.2111.02080

Modellierung von Gattungsunterschieden. Emotionen in Lyrik, Prosa und Drama

Kröncke, Merten
merten.kroencke@uni-goettingen.de
Universität Würzburg, Deutschland

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Universität Göttingen, Deutschland

Winko, Simone
simone.winko@phil.uni-goettingen.de
Universität Würzburg, Deutschland

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Göttingen, Deutschland
ORCID: 0000-0001-6944-6113

Einleitung

Literaturwissenschaftliche Untersuchungen zielen häufig darauf ab, verschiedene Textgruppen (zum Beispiel Gedichte, Romane und Dramen) hinsichtlich verschiedener Texteigenschaften (zum Beispiel Themen oder Emotionen) miteinander zu vergleichen.[1] Werden zu diesem Zweck Verfahren der Computational Literary Studies eingesetzt, bedeutet das oft, mit manuellen Annotationen trainierte Modelle auf verschiedene Textgruppen (Domänen) anzuwenden. In vielen Fällen stehen dabei aus Ressourcengründen nicht so viele Annotationen zur Verfügung, dass Modelle für jede Domäne auf ausschließlich domänenspezifischen Annotationen trainiert werden könnten. Stattdessen sind oft nur für *eine* Domäne hinreichend viele Annotationen verfügbar. Eine naheliegende Lösung besteht darin, für die weiteren, nicht umfassend annotierten Domänen dasselbe Modell wie für die ausführlich annotierte Domäne zu verwenden, oder zumindest Modelle, die maßgeblich auf Annotationen dieser Domäne basieren. Bei derartigen Prozessen kommt es bekanntermaßen zu Performanceeinbußen. Dies sollte allerdings nicht (nur) Anlass zu Enttäuschung geben, vielmehr lassen sich die zunächst suboptimal scheinenden Ergebnisse (auch) produktiv machen. Eben das zu demonstrieren, ist ein wichtiges Ziel dieses Beitrags. Vor allem soll anhand von Performanzunterschieden etwas über die Charakteristika der Domänen, und das heißt in diesem Fall: der literarischen Gattungen, ausgesagt werden.

In der Computerlinguistik (oder allgemeiner im gesamten Feld Machine Learning) wird die erläuterte Problemstellung unter dem Schlagwort Domain Adaptation intensiv beforscht (z.B. Ramponi und Plank 2020). Machine Learning-Probleme lassen sich als Versuch beschreiben, eine automatische Zuweisung von Datenpunkten x zu Labeln y zu lernen. Dabei wird unterstellt, dass alle Punkte x aus der gleichen Verteilung stammen, die zu lernende Zuweisung x#y also für jeden Datenpunkt ähnlich funktioniert. Diese Annahme ist in angewandter Forschung (darunter Computational Literary Studies) jedoch selten zu halten. Die Gründe für einen Domain Shift, also die Veränderung von x, während y stabil bleibt, können vielfältig sein (z.B. historischer Sprachwandel, Übersetzungen). Die Bereitstellung von Datensätzen, die dezidiert mehrere Domänen enthalten, ist also sowohl für Machine Learning Forschung als auch für die CLS hoch relevant.

Als Untersuchungsbeispiel dient die Gestaltung von Emotionen in deutschsprachigen Lyrik-, Prosa- und Dramentexten der zweiten Hälfte des 19. und des beginnenden 20. Jahrhunderts. In früheren Studien hat sich unsere Forschungsgruppe auf Lyrik konzentriert und für diese Gattung umfangreiche manuelle Annotationen erstellt.[2] Nun sollen ergänzend Prosa- und Dramentexte einbezogen werden; in beiden Fällen liegen deutlich weniger Annotationen vor. Da wir von der Emotionsontologie unserer Lyrikannotation ausgehen, können die Daten anderer Projekte, die sich mit der Analyse von Emotionen in deutscher Literatur befassen (z. B. Haider et al. 2020, Dennerlein et al. 2022), (noch) nicht einbezogen werden. Ob es in Hinsicht auf die Emotionsgestaltung Unterschiede zwischen Prosa, Drama und Lyrik gibt, ist nicht zuletzt gattungstypologisch interessant. Im Untersuchungszeitraum war die Auffassung verbreitet, dass Subjektivität und auch Emotionalität charakteristische, gegebenenfalls sogar definitionsrelevante Merkmale der Gattung Lyrik seien (vgl. Lamping 2000: 56f.). Zwar zählt die Gestaltung textueller Emotionen nicht zu den „gattungsbildende[n] Zentralmerkmal[en]“ (Zymner 2007: 36) in der Bestimmung von Prosa, Drama und Lyrik; jedoch wurde und wird – in Verallgemeinerung des Modells ‚Erlebnislyrik‘ bzw. epochenspezifischer Lyrikvarianten – Gedichten noch immer ein höherer Anteil an Subjektivität und Emotionalität zugeschrieben, wenn heute auch nur als fakultative Eigenschaft. Wenn sich die Gattungen im Untersuchungszeitraum abweichend verhalten, könnte dies aufschlussreich für die gattungstypologische Relevanz des Merkmals ‚Emotionalität‘ sein. Insgesamt soll nun also danach gefragt werden, welche Schwierigkeiten sich dabei ergeben, auf Lyrikannotationen trainierte Modelle zur Emotionserkennung auf Prosa- und Dramentexte anzuwenden,

und was sich so über etwaige Gattungsunterschiede lernen lässt.

Ressourcen

Die Studie verwendet drei Korpora: ein vergleichsweise großes Lyrikkorpus, für das umfangreiche manuelle Annotationen vorliegen, und zwei deutlich kleinere, zu Testzwecken zusammengestellte Korpora mit einerseits Prosa- und andererseits Dramentexten. Das Lyrikkorpus besteht aus Texten in Anthologien aus dem Untersuchungszeitraum, die sich auf Gedichte von Zeitgenoss:innen konzentrieren. Die Anthologien stammen aus der Zeit von 1859 bis 1919 und enthalten mehr als 6000 Gedichte, von denen 1412 (270k Token) annotiert wurden.[3] Die Prosa- und Dramenkorpora bestehen aus jeweils 5 vollständig annotierten Texten aus der Zeit um 1900 (Prosakorpus: 17k Token, Dramenkorpus: 34k Token).[4]

Die Emotionsannotation zielt darauf ab, die im Text gestalteten Emotionen (und nicht die Emotionen der Leser:innen) 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.[5] Da für viele einzelne Emotionen nur eine sehr geringe Zahl von Annotationen vorliegen, 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). Die Lyrik-, Prosa- und Dramentexte wurden alle auf dieselbe Weise annotiert. Das Inter-Annotator-Agreement (Mahet 2015) für die Emotionsgruppen beträgt 0.71 # (Lyrik), 0.61 # (Prosa) und 0.59 # (Drama).[6] Für die Agreement-Differenzen ist möglicherweise mitverantwortlich, dass die Annotator:innen bislang deutlich weniger Prosa- und Dramen- als Lyriktexte annotiert haben und insofern mit den Gattungen unterschiedlich vertraut sind.

Tabelle 1: Annotierte Segmente nach Gattung und Emotion.

	Agitation	Fear	Anger	Sadness	Joy	Love
Lyrik	591	532	870	3955	4233	4159
Drama	134	178	161	238	148	144
Prosa	66	42	45	41	50	23

Klassifikation

Tabelle 2 zeigt die Qualität der Emotionsklassifikation in den drei Gattungen. Verwendet wird ein Modell, welches lediglich mit den Annotationen für Lyrik trainiert ist. Basis ist das deutsche Bert-Modell gbert-large (Chan et al. 2020). Dieses wird zusätzlich auf Lyrik angepasst[7] und anschließend trainiert[8] (Konle et al. 2022).

Tabelle 2: Evaluation nach Gattung und Emotion (F1 Macro).

	Agitation	Fear	Anger	Sadness	Joy	Love	MEAN
Lyrik (CV9)	.62	.79	.71	.74	.73	.77	.73
Drama	.51	.61	.54	.43	.58	.58	.54
Prosa	.52	.53	.57	.57	.64	.56	.56

Mögliche Erklärungen für Unterschiede in der Klassifikationsperformanz

Die folgenden Abschnitte beschäftigen sich mit der Suche nach möglichen Erklärungen für die großen Qualitätsunterschiede (siehe Tab. 2). Es werden sowohl Eigenschaften des Modells als auch die Verteilung der annotierten Emotionen, die Zusammensetzung einzelner Emotionen und das zugrundeliegende Textmaterial untersucht.

Modellanalyse

Zunächst lässt sich danach fragen, wie sicher sich das Modell bei den Klassifikationen ist (Abb. 1). Blickt man auf die Vorhersage von (vorhandenen) Emotionen, ist die Sicherheit erwartungsgemäß bei Lyrik am größten, gefolgt von Prosa und danach Drama. Bei der Vorhersage ‘keine Emotion’ ist sich das Modell hingegen im Fall von Dramentexten besonders sicher (sogar noch sicherer als im Fall von Gedichten) und im Fall von Prosa besonders unsicher. Zum einen scheint die Klassifikationsperformance also mit der Sicherheit des Modells zusammenzuhängen; zum anderen weisen die Differenzen zwischen Prosa- und Dramentexten in puncto ‘Emotion’/’Keine Emotion’ auf klassifikationsrelevante Gattungsunterschiede hin.

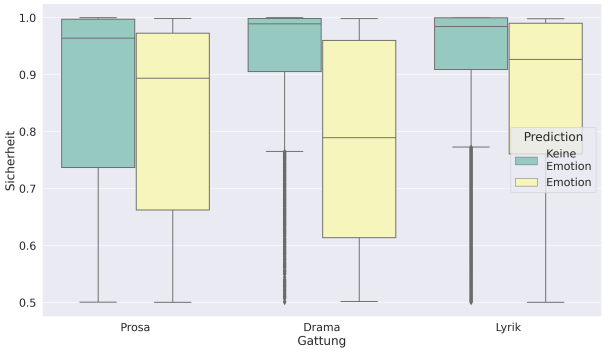


Abbildung 1: Sicherheit der Klassifikation nach Gattung und Label. Sicherheit wird hier mit der Wahrscheinlichkeit (nach Softmax Transformation) für die ausgegebene Klasse gleichgesetzt.

Emotionsverteilung

Um einen Eindruck von der Emotionsverteilung innerhalb der Gattungen zu erhalten, werden jeweils 50 Segmente (Verse bzw. Sätze) zu einer Einheit zusammengefasst, die als Vektor über die Anzahl der enthaltenen Emotionen repräsentiert wird. Um diese Vektoren zu visualisieren, werden sie in den 2-dimensionalen Raum projiziert (siehe McInnes 2018). Das Resultat (Abb. 2) zeigt, dass die annotierten Gedichte stärker streuen als die annotierten Texte der übrigen Gattungen, also vielfältigere Mischungen an Emotionen enthalten. Auffällig ist zusätzlich die Häufung von Dramen und Prosa im oberen rechten Bereich der Grafik. Die Emotionsverteilung innerhalb der beiden Gattungen ähnelt sich nach diesem Befund und weicht zugleich von der Verteilung in den meisten lyrischen Texten ab.

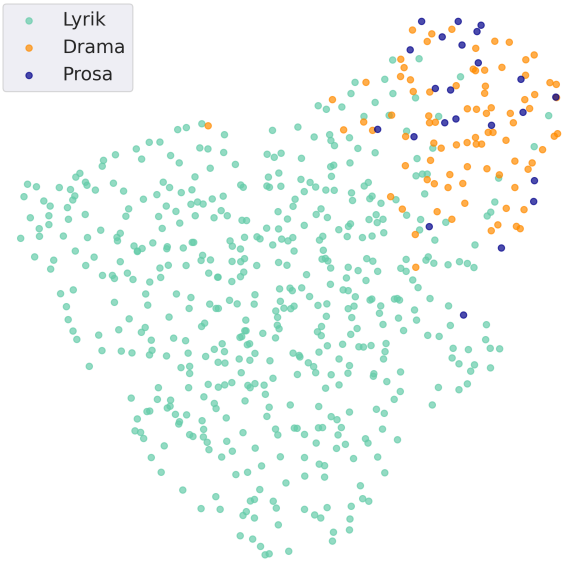


Abbildung 2: Annotierte Emotionen über je 50 Segmenten aggregiert und in den 2D Raum projiziert.

Abbildung 3 macht deutlich, dass die Gattungen Emotionen in stark unterschiedlicher Häufigkeit gestalten. Lyrik enthält mit Abstand die meisten Emotionen, beinahe das Dreifache im Vergleich zu Dramen. Diese enthalten wiederum das Doppelte an Emotionen, gemessen an Prosa.

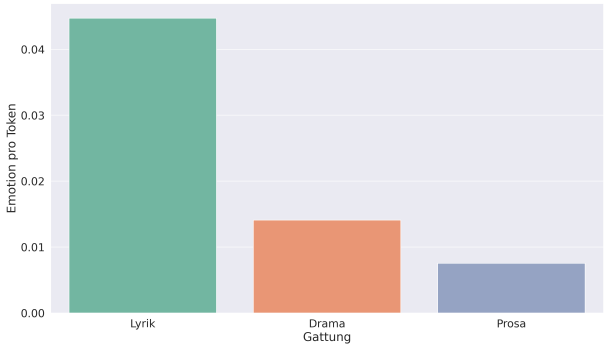


Abbildung 3: Emotionen pro Token in Gattungen.

Abbildung 4 ermöglicht einen Einblick in die einzelnen Emotionsgruppen nach Shaver (1987). Während Lyrik eine deutliche Ungleichverteilung zugunsten von Trauer, Freude und Liebe zeigt, sind die Gruppen in Dramen und Prosa nahe an einer Gleichverteilung.

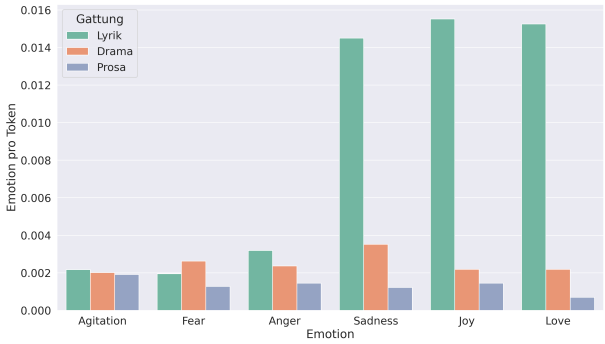


Abbildung 4: Emotionsgruppen pro Token in Gattungen.

Zusammensetzung der Emotionsgruppen

Abbildung 5 zeigt beispielhaft für die Emotionsgruppe Erregung/Überraschung, wie sich die Gruppe je nach Gattung anteilig zusammensetzt. Es zeigen sich erhebliche Unterschiede: Während in lyrischen Texten die Kategorie ‘Emotionalität’ dominiert, die vor allem für unspezifische Emotionen eingesetzt wird (‘Er war ein grundsätzlich emotionaler Mensch’ usw.), kommt in den annotierten Dramen ‘Aufregung’ am häufigsten vor; in den annotierten Prosatexten ist wiederum die Einzelemotion ‘Spannung’, verglichen mit den anderen Gattungen, besonders verbreitet. Diese Unterschiede erzeugen ein großes Fehlerpotential, da sich mit der Zusammensetzung auch die Repräsentation der Gruppe im Modell ändert. Während in Lyrik bereits gute Ergebnisse erzielt werden können, wenn lediglich die Ein-

zelemotion ‘Emotionalität’ erkannt wird, ist diese für Prosatexte nutzlos. Umgekehrtes gilt für Spannung.

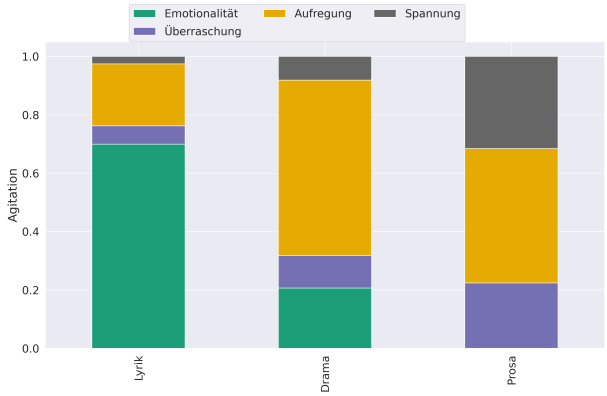


Abbildung 5: Zusammensetzung der Gruppe Erregung/Überraschung nach Gattungen.

Texteigenschaften

Nachdem die Verteilungsunterschiede in den Emotionen dargestellt sind, werden Differenzen in der sprachlichen Gestaltung der annotierten Texte untersucht.

Tabelle 3: Emotionswörter pro Emotion.

Lyrik	Drama	Prosa
0.22	0.27	0.23

Neben der bislang betrachteten Emotionsannotation wurde separat festgehalten, welche Wörter im Text über ihre lexikalische Bedeutung markieren, dass eine Emotion gestaltet wird, zum Beispiel Ausdrücke wie ‚Angst‘, ‚lachen‘ oder ‘jauchzen’. Eine Emotionsannotation wird meist, aber nicht immer, von der Annotation entsprechender Emotionswörter begleitet; umgekehrt kommen Emotionswörter nie ohne zugehörige Emotionsannotation vor. Tabelle 3 zeigt, wie viele Emotionswörter pro Emotionsannotation je nach Gattung vorkommen. In Dramen werden etwas mehr Emotionswörter verwendet als in Prosa und Lyrik, eine explizitere Nennung von Emotionen in Lyrik als mögliche Fehlerquelle in der anschließenden Klassifikation der anderen Gattungen kann damit ausgeschlossen werden.

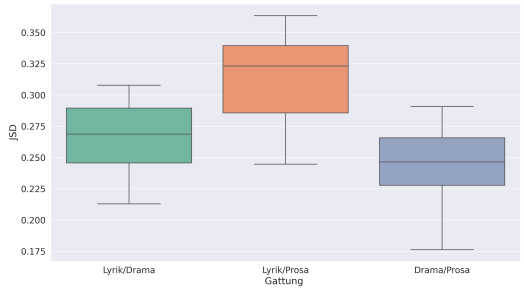


Abbildung 6: Jensen-Shannon-Divergenz zwischen Gattungen repräsentiert durch die jeweils 8000 häufigsten Wörter der einzelnen Gattungen.

Abschließend lässt sich danach fragen, wie groß der Abstand zwischen den Gattungen hinsichtlich des Textmaterials ist und ob die Unterschiede in der Klassifikationsperformance dazu ‘passen’. Abb. 6 zeigt, dass der Abstand zwischen Lyrik und Prosa am größten, der Abstand zwischen Lyrik und Drama bereits deutlich geringer und der Abstand zwischen Drama und Prosa am geringsten ausfällt. Dass Lyrik also, was das Textmaterial angeht, eher den einbezogenen Dramen- als den Prosatexten ähnelt, schlägt sich jedoch nicht unmittelbar in der Klassifikationsperformance nieder, die nämlich im Fall von Dramen nicht besser als im Fall von Prosa ist.

Diskussion

Die Studie ist von der Frage ausgegangen, wie sich etwaige Unterschiede zwischen den literarischen Großgattungen Prosa, Drama und Lyrik in puncto Emotionsgestaltung mit Fragen der Domain Adaptation verknüpfen und aus dieser Perspektive modellieren lassen. Der (zu erwartende) Befund, dass ausschließlich auf Lyrikannotationen trainierte Modelle deutlich schlechter performen, wenn sie auf Prosa- und Dramentexte angewendet werden, kann mit einer ganzen Reihe von Faktoren zusammenhängen, von denen einige näher untersucht wurden. Neben pragmatischen Gesichtspunkten, zum Beispiel den etwas niedrigeren Inter-Annotator-Agreement-Werten, scheinen vor allem Spezifika der Domänen eine Rolle zu spielen. Erhebliche Unterschiede zwischen den Gattungen zeigen sich unter anderem, wenn man die Häufigkeit und Verteilung der gestalteten Emotionen betrachtet und wenn man danach fragt, aus welchen Einzelemotionen sich die Emotionsgruppen zusammensetzen. Demgegenüber deuten weitere Ergebnisse darauf hin, dass sich manche Gattungsunterschiede *nicht* – oder zumindest nicht unvermittelt – wie erwartet in den Klassifikationsergebnissen niederschlagen. Zumindest hat sich gezeigt, dass Dramen sowohl relativ viele Emotionswörter enthalten als auch lyrischen Texten auf sprachlicher Ebene vergleichsweise stark ähneln und dass für sie trotzdem keine besseren Klassifikationsergebnisse erzielt wurden als für Prosa.

Die Ergebnisse deuten indizienhaft an, dass Gedichte, verglichen mit Dramen und Prosatexten, besonders häufig Emotionen gestalten. Dieser Befund passt zu den in der Einleitung erwähnten gattungstypologischen Vermutungen, wenngleich berücksichtigt werden muss, dass an dieser Stelle nur sehr wenige Prosa- und Dramentexte einbezogen werden konnten. Um noch besser abgesicherte Schlüsse über die Gattungen zu ermöglichen, werden wir weitere Texte annotieren und verschiedene Verfahren der Domain Adaptation testen, um letztlich auch für Prosa und Dramen zuverlässige Klassifikatoren trainieren zu können.

Fußnoten

1. Wir danken Melanie Andresen für das sehr detaillierte Feedback und genaue Lektüre unseres Beitrags.
2. Vgl. z. B. Konle u. a. 2022a, Kröncke u. a. 2023.
3. Korpus: Winko u. a. 2022 a; Korpusbeschreibung: Winko u. a. 2022 b.
4. Prosatexte: Detlev von Liliencron, *Unter flatternden Fahnen* (1888); Bjarne P. Holmsen (Arno Holz/Johannes Schlaf), *Ein Tod* (1889); Peter Altenberg, *Fünfundzwanzig* (1896); Hugo von Hofmannsthal, *Reitergeschichte* (1899); Maria Janitschek, *Darüber kommt kein Weib hinweg …* (1902). Dramen: Hugo von Hofmannsthal, *Der Tod des Tizian* (1892); Arthur Schnitzler, *Lebendige Stunden* (1902); Hennie Raché, *Belsazar* (1904), Paul Scheerbart, *Lachende Gespenster* (1904); August Stramm, *Erwachen* (1914).
5. Annotationsrichtlinien: Kröncke u. a. 2022.
6. Verwendet wurden die Default-Einstellungen. Für weitere Ausführungen siehe Reiter und Konle 2022.
7. Hyperparameter: 500 steps, batchsize 30, learning rate 2e-5 (see Konle and Jannidis 2020, Gururangan et al. 2020)
8. Hyperparameter: 20 Epochen, batchsize 20, learning rate 1e-4
9. Cross Validation

Bibliographie

Chan, Brandon; Schweter, Stefan and Möller, Timo. 2020. German's next language model In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), pp. 6788–6796.

Dennerlein, Katrin; Schmidt, Thomas; Wolff, Christian. 2023. Computational emotion classification for genre corpora of German tragedies and comedies from 17th to early 19th century, Digital Scholarship in the Humanities, Volume 38, Issue 4, 2023, Pages 1466–1481, https://doi.org/10.1093/llc/fqad046

Ekman, Paul. 1992. An Argument for Basic Emotions. Cognition and Emotion, vol. 6, no. 3-4, pp. 169–200.

Ekman, Paul. 1999. Basic Emotions. Handbook of Cognition and Emotion, edited by John Tim Dagleish and Mich J. Power. Wiley, pp. 45-60.

Dennerlein, Katrin; Schmidt Thomas; Wolff Christian. 2023. Computational emotion classification for genre corpora of German tragedies and comedies from 17th to early 19th century, Digital Scholarship in the Humanities, Volume 38, 4, pp. 1466–1481, https://doi.org/10.1093/llc/fqad046

Gururangan, Suchin; Marasović, Ana; Swayamdipta, Swabha; Lo, Kyle; Beltagy, Iz Downey, Doug and Smith, Noah A. 2020. Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Haider, Thomas; Eger, Steffen; Kim, Evgeny; Klinger, Roman and Menninghaus, Winfried. 2020. PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1652–1663, Marseille, France. European Language Resources Association.

Mathet, Yann; Widlöcher, Antoine and Métivier, Jean-Philippe. 2015. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. Computational Linguistics, 41(3), pp. 437–479.

Konle, Leonard; Kröncke, Merten; Jannidis, Fotis; Winko, Simone. 2022. Emotions and Literary Periods. DH2022. Tokyo.

Kröncke, Merten; Jannidis, Fotis; Konle, Leonard; Winko, Simone. 2022. Annotationsrichtlinien Emotionsmarker und Emotionen, https://doi.org/10.5281/zenodo.6021152.

Krönke, Merten; Konle, Leonard; Winko, Simone; Jannidis, Fotis. 2023. Gattungen und Emotionen in der Lyrik des Realismus und der frühen Moderne, in: DHd2023: Open Humanities Open Culture. Konferenzabstracts, Belval/Trier, 13.–17. März 2023, DOI: 10.5281/zenodo.7715402.

Lamping, Dieter. 2000. Das lyrische Gedicht. Definitionen zu Theorie und Geschichte der Gattung. 3. Aufl. Göttingen.

McInnes, Lelland; Healy, John; Saul, Nathaniel and Großberger, Lukas. 2018. UMAP: Uniform Manifold Approximation and Projection. The Journal of Open Source Software, 3(29).

Plutchik, Robert. 1980a. Emotion: A Psychoevolutionary Synthesis. Harper & Row.

Plutchik, Robert. 1980b. "A general psychoevolutionary theory of emotion." Emotion: Theory, Research and Experience. Theories of Emotion, edited by Robert Plutchik and Henry Kellerman. Academic Press, vol. 1, pp. 3–33.

Plutchik, Robert. 2001. "The Nature of Emotions." American Scientist, vol. 89, no. 4, pp. 344–350.

Ramponi, Alan and Plank, Barbara. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Reiter, Nils; Konle, Leonard. 2022. Messverfahren zum Inter-annotator-agreement (IAA). DARIAH-DE Working Papers No. 44. Göttingen: DARIAH-DE, 2022. DOI: 10.47952/gro-publ-103.

Winko, Simone; Konle, Leonard; Kröncke, Merten; Jannidis, Fotis. 2022a. Lyrik-Anthologien 1850-1910, https://doi.org/10.5281/zenodo.6053952.

Winko, Simone; Konle, Leonard; Kröncke, Merten; Jannidis, Fotis. 2022b. Korpusbeschreibung der Lyrik-Anthologien 1850-1910, https://doi.org/10.5281/zenodo.6204787.

Zymner, Rüdiger. 2007. Texttypen und Schreibweisen. In: Thomas Anz (Hg.): Handbuch Literaturwissenschaft. Bd. 1: Gegenstände und Grundbegriffe. Stuttgart, Weimar, pp. 25–80.

Musikhistorische Daten, Netzwerkanalyse und Migration

Stadler, Peter

pstadler@mail.uni-paderborn.de
Universität Paderborn, Deutschland
ORCID: 0000-0002-2544-1481

Grund, Vera

vgrund@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Einführung

Bei der ersten DHd-Tagung 2014 in Passau wurde ein Thesenpapier zur Standortbestimmung der Digital Humanities in Hinblick auf ihr Selbstverständnis, ihre Aufgabenstellung und Entwicklungsperspektiven erarbeitet. Besonders betont wurde dabei die Bedeutung von Interdisziplinarität für die Digital Humanities sowie ihr besonderes methodisches Potential, Erkenntnisgewinn durch empirische Praxis zu erzielen. Als wissenschaftliche Handlungsräume wurden die "Annotation, Analyse, Aggregation und Rekombination von geisteswissenschaftlichen Objekt- und Metadaten in Kombination von algorithmischen und hermeneutischen Prozeduren"[1] benannt.

Um einen Material- bzw. Datenkorpus, der in diesem Sinne als Grundlage für ein Digital Humanities Forschungsprojekt dienen kann, handelt es sich bei der bereits in den 1980er Jahren durch das „U.S. RISM Libretto Project" digital erschlossene Albert Schatz-Librettokollektion (McClymonds/Parr Walker 1986; Needham Costonis 1991). Es handelt sich dabei um eine der größten Sammlungen von Opernlibretto-Drucken. Diese wurden in der Regel in Zusammenhang mit Aufführungen erstellt und enthielten daher neben dem Text der Oper außerdem Informationen zu den aufgeführten Werken und den Aufführungen selbst. Diese Informationen wurden bei der Erschließung durch das RISM Libretto Projekt feingranular aufgenommen, so dass nicht nur basale Metadaten wie Titel, Komponist:in, Librettist:in, Druckort und -jahr erfasst sind, sondern auch die in den Libretti enthaltenen Angaben zu den Ausführenden der jeweiligen Aufführung. Aufgrund der Qualität und Menge der dabei generierten Daten können diese für die quantitative Netzwerkforschung weiter genutzt werden, oder wie es in dem Thesenpapier heißt, "zur Formulierung neuer qualitativer geisteswissenschaftlicher Fragestellungen" sowie für die "Operationalisierung qualitativer geisteswissenschaftlicher Fragestellungen auf der Basis formaler, logisch-mathematischer Verfahren der Informatik" die Basis bilden.[2] Das Projekt möchte durch die Nachnutzung von Daten zwischen frühen Digitalisierungsprojekten und den Digital Humanities vermitteln, indem es die Potentiale der Daten ebenso wie die Schwachstellen aufzeigt und so Kriterien für die Nachhaltigkeit von Daten erkennen lässt. In diesem Sinne kann es prototypisch für die in den Thesen formulierten Kriterien von Digital Humanities stehen und als Fallbeispiel in Hinblick auf die Methodik bei der Erstellung von Datenkorpora in Digitalisierungsprojekten und deren Nachnutzbarkeit für die Digital Humanities stehen. Im Rahmen des Papers sollen erste Ergebnisse zur Erforschung von Künstler:innen-Netzwerken auf der Basis der Daten aus der Schatz-Sammlung im Verhältnis zu den Thesen zu den Digital Humanities präsentiert sowie ihre Potentiale für weiterführende Forschung aufgezeigt werden. Weiter sollen auf der Basis des Projekts Ergänzungen der Thesen insbesondere in Hinblick auf in den Daten enthaltene Differenzkategorien und den bewussten Umgang damit dargestellt werden.

Die Sammlung Schatz

Die Library of Congress kaufte im Jahr 1908 die Libretto-Kollektion des Rostocker Sammlers Albert Schatz als Vorlass, bevor dieser im Jahr 1910 verstarb. Schatz beschrieb die Sammlung als „Ergebnis eigener Sammeltätigkeit von über 42 Jahren" (Sonneck 1914), die er aufgrund von Interesse für die Oper und ihre Geschichte begann. Diese forcierte er zusätzlich, als er im Herbst 1873 Besitzer der Rostocker Musikalienhandlung Ludwig Trutschel wurde. Da laut Schatz „die bisherigen Darstellungen der Operngeschichte [...] völlig unzureichend und unzuverlässig waren, da sie in der Hauptsache nicht unmittelbar auf den Quellen aufgebaut, sondern ohne [ihre] Kenntnis" (Sonneck 1914) verfasst wurden, entschloss er sich auf der Basis seiner Librettosammlung eine Operngeschichte zu schreiben. Zwar blieb diese ein Fragment, jedoch dokumentierte er die Erkenntnisse aus seiner Sammeltätigkeit in Form von circa 80.000 Zetteln, mit denen Schatz laut eigener Aussage 30.000 Opernaufführungen nachweisen konnte.[3]

Die von der Library of Congress angekaufte Libretto-Sammlung enthielt 12.240 Textbücher, bei denen es sich abgesehen von wenigen hundert Oratorien und Kantaten, die vermutlich per Zufall in die Sammlung kamen, um Programme zu Opernaufführungen handelt. Seltene Librettodrucke wurden auch als handschriftliche Kopien in die Sammlung aufgenommen. Der geografische Schwerpunkt lag auf deutschen und italienischen Texten, wohingegen französische Opern eher in deutschen Übersetzungen Eingang in die Sammlung fanden (Sonneck 1914). Bei der Erschließung der Albert-Schatz-Collection durch das „U.S. RISM Libretto Project“ wurden neben den Grunddaten zu den mit den Libretti in Verbindung stehenden Werken auch Figurennamen sowie die in den Textbüchern enthaltenen Informationen zu Aufführungen wie Theater und beteiligte Personen erfasst.

Datenaufbereitung

Von besonderem Interesse für die Netzwerkforschung ist dabei die systematische und normierte Erfassung bzw. Zuschreibung von „relator terms“ zu den beteiligten Personen nach dem kontrollierten Vokabular MARC Code List for Relators[4]. So sind beispielsweise in dem Metadatensatz zum Libretto Le villanelle astute zur Aufführung im venezianischen Teatro San Samuele im Jahr 1786 die auf der Titelseite genannten Personen Giovanni Maria Foppa, Francesco Bianchi, Catarina Casalis und Eusebio Luzzi als „librettist“, „composer“, „vocalist“ und „dancer“ ausgewiesen (Foppa 1786).

Die Daten wurden durch das „U.S. RISM Libretto Project“ nicht nur gut und strukturiert erfasst, sie werden auch von der Library of Congress über eine offene Programmierschnittstelle verfügbar gemacht, so dass die gesamten Angaben zu den Libretti, die über den Online-Katalog zugänglich sind, auch maschinell abgerufen werden können.[5] Aus diesen Datensätzen haben wir die folgenden Informationen extrahiert:

Personennamen und Funktionen

Diese beiden Angaben finden sich in innerhalb der Beschreibung des „item“ als „contributor_names“. Jede Person ist einzeln aufgenommen, allerdings ist der „relator term“ in Klammern angefügt, sodass dieses Feld entsprechend analysiert werden muss, um die Namen von den Funktionen zu separieren. Aus dem Eintrag „Luzzi, Eusebio. (dancer)“ (siehe obiges Beispiel) wird dann entsprechend der Name „Luzzi, Eusebio“ mit der Funktion „dancer“ abgeleitet. Im Gegensatz zu den relator terms sind die Eigennamen nicht normalisiert, sondern tauchen oft in verschiedenen orthographischen Varianten oder Abkürzungen (von z.B. Vornamen) auf. Insgesamt ergeben sich so in unserem Korpus über 43.000 verschiedene Namensformen, die wir mithilfe von OpenRefine (vorsichtig) geclustert haben, um diese Namensvarianten zusammenzuführen.

Aufführungsort

Den Aufführungsort gewinnen wir aus den „location“-Feldern. Hierbei gehen wir davon aus, dass der Druckort dem Aufführungsort entspricht. Für unsere Untersuchung ziehen wir nur die Libretti heran, die explizit die beteiligten Künstler:innen wie „vocalist“ erwähnen, sodass bei diesen Libretti davon auszugehen ist, dass sie auch am Aufführungsort gedruckt worden sind.

Die Orte sind ähnlich wie die Personennamen nicht normalisiert (Venedig erscheint beispielsweise als „Venetia“, „Venezia“ oder auch „In Venezia“ etc.) und mussten analog der Personennamen geclustert werden. Der Umfang war allerdings wesentlich geringer: Insgesamt existierten 748 verschiedene Ortsangaben im Gesamtkorpus, die wir auf 427 normalisierte Angaben abbilden konnten.

Aufführungsjahr

Die Datumsangaben wurden bei der Katalogisierung ebenfalls von den Titelblättern der Drucke übernommen und liegen in zwei Varianten vor. Einerseits als Freitextvariante (z.B. “18??”) und andererseits eine offenbar davon abgeleitete normalisierte Form mit zwei Einträgen (“1800-01-01T00:00:00Z” und “1899-01-01T00:00:00Z”). In der Katalogansicht der Library of Congress wird für die Facettierung dann diese normalisierte Form genutzt, was zu verzerrten Angaben führt, wenn für die Jahre 1800 und 1899 mehrere Hundert Treffer ausgegeben werden, für 1801 aber z.B. nur 80. Wir haben daher für die Bestimmung des Aufführungsjahres diese Unschärfen der ursprünglichen Datenerfassung aufgefangen und ggf. Datumsbereiche oder unsichere Datumsangaben markiert.

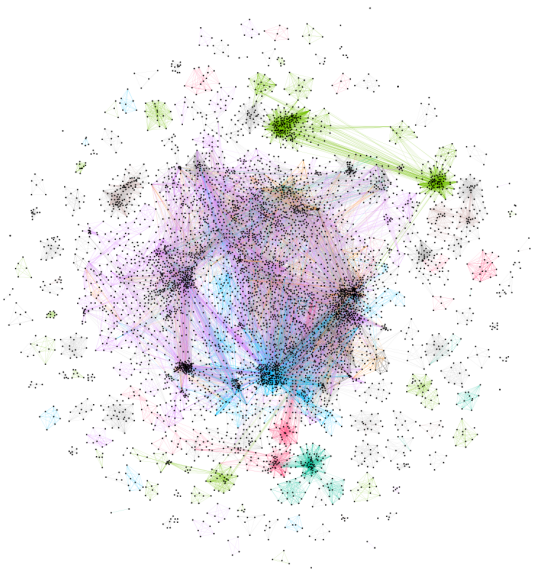
Netzwerke

Aus den so normalisierten Daten zu Personen, Funktionen, Orten und Jahr haben wir anschließend eine Einschränkung auf Sänger:innen (vocalists) im 18. Jahrhundert vorgenommen, um entsprechende Thesen zu Mobilität und Migration aus der musikwissenschaftlichen Forschung untersuchen zu können (Strohm 2001; Zur Nieden 2015).

Personennetzwerk

Zunächst haben wir zwecks eines initialen, explorativen Vorgehens ein einfaches ungerichtetes Personennetzwerk erstellt, bestehend aus den Personen als Knoten und den Aufführungen – bei denen zwei Personen gemeinsam künstlerisch gewirkt haben – als Kanten. Dieses Netzwerk der Sänger:innen-Beziehungen des 18. Jahrhunderts besteht aus insgesamt 6203 unterschiedlichen Personen (Knoten), die über 50.450 Relationen miteinander verbunden sind. Die Abbildung 1 gibt einen ersten Eindruck von der Größe des Netzwerks, sichtbar wird aber durch die Einfär-

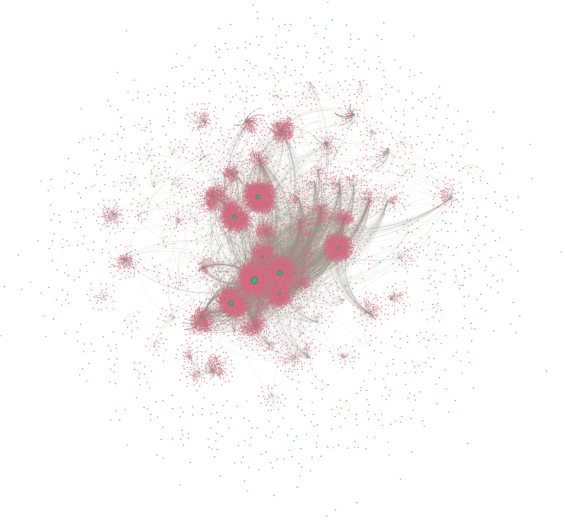
bung der Kanten nach Orten auch direkt die Dominanz Venedigs (in Lila) und die räumliche Trennung von z.B. Paris (Grün) und Berlin (Rot).



Bemerkenswert ist aber ein Blick auf die Zentralitätsmaße: Hier sticht Nicola Grimaldi mit dem höchsten gewichteten Grad von 415 und einem Eigenvektor Zentralitäts-Score von 1 hervor.[6] Dies verwundert nicht, ist Grimaldi doch in der Musikgeschichte als erfolgreicher und bekannter Sänger überliefert.

Bipartites Personen-Orts-Netzwerk

Um das Verhältnis bzw. den Einfluss von den Orten auf die Personen genauer zu untersuchen, haben wir im Anschluss ein (ungerichtetes) bipartites Personen-Orts-Netzwerk erstellt. Hierbei wurden Orte (grün eingefärbt) und Personen (rot eingefärbt) als Knoten aufgenommen und als Kanten die durch Aufführungen dokumentierten Beziehungen zwischen Personen und Orten (vgl. Abb.2). Die nicht verbundenen Knoten zeigen Orte und Personen an, die lediglich erwähnt sind, aber keine Beziehung zu einem Knoten des anderen Typs herstellbar ist. Im Falle von Orten sind dies Aufführungen ohne Beteiligung von Sänger:innen und bei Personen sind dies Aufführungen mit unbekanntem Aufführungsort.

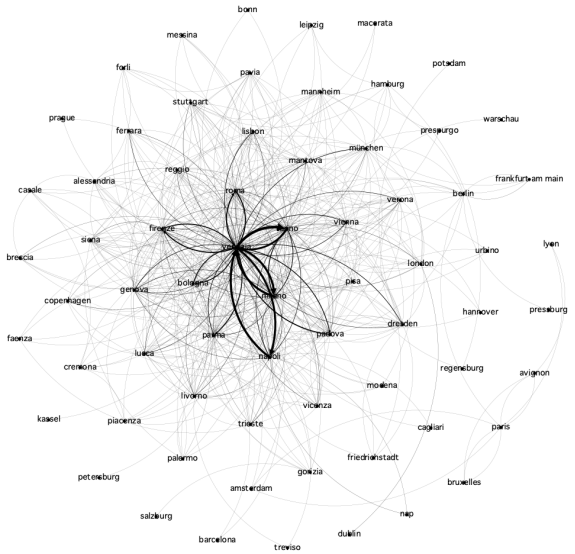


Wie bereits durch die Einfärbung der Kanten im Personennetzwerk angedeutet, manifestiert sich hier eine deutliche Dominanz der italienischen Opernzentren, insbesondere Venedigs. Unser Hauptaugenmerk liegt aber auf der Mobilität der Sänger:innen, die hier ebenfalls sichtbar wird. So finden sich zahlreiche Personen, die mit 5 oder mehr Orten verknüpft sind, d.h. eine hohe Mobilität aufweisen. Spitzenreiter mit 11 verschiedenen Orten ist Girolamo Crescentini, ein berühmter italienischer Kastrat, der um 1800 wirkte. Dabei fehlen bei unserer Beschränkung auf das 18. Jh. sogar noch Stationen, denn Crescentini war im 19. Jahrhundert noch in Wien und Paris tätig.

Die überwiegende Mehrzahl der Personen ist aber nur mit einem Ort in Verbindung zu bringen. Ob das nun an einer größeren Seßhaftigkeit oder aber der Datenlage oder der Ausschnittsbildung liegt, lässt sich aktuell noch nicht beantworten.

Ortsnetzwerk

Zuletzt haben wir noch einen gezielten Blick auf die Migrationsbewegungen zwischen den Orten gelegt und dafür ein gerichtetes Ortsnetzwerk erstellt. Diesem liegen die chronologisch sortierten Auftritte jeder Person im Korpus zugrunde, sodass sich aus der Aufeinanderfolge von zwei Auftritten einer Person an verschiedenen Orten eine Bewegung inferieren lässt.



In Hinblick darauf ist das 18. Jahrhundert als Zeitraum von besonderem Interesse. Denn wie Reinhard Strohm feststellte (Strohm 2001), wandelte sich aufgrund von zunehmender Mobilität das Verhältnis in Bezug auf die von ihm konstatierte „italian musical diaspora“. Wie Strohm darstellte, hatte diese zuvor als mediale Wanderbewegung von Notendrucken stattgefunden. Da große Teile der Kultur des 18. Jahrhunderts durch „Italianità“ als Modetrend bestimmt war, ging damit ein starkes Interesse an italienischen Künstlerinnen und Künstlern im europäischen Raum einher, weswegen laut Strohm Wanderbewegungen zunahmen. Er ging davon aus, dass sich die Karriereverläufe von Bühnenkünstler:innen, insbesondere von Sängerinnen und Sänger, daher stark änderten. In Hinblick auf italienische Kastraten-Sänger behauptete Strohm sogar, dass es sich um “Exportprodukte“ handelte. Unsere Daten belegen einen regen Austausch zwischen den verschiedenen europäischen Musikzentren, allerdings kann von einem reinen Export nicht die Rede sein.

Fazit und Ausblick

An dieser Stelle kann nur ein Vorgeschmack auf den „Daten-Schatz“ gegeben werden, der große Mengen an Informationen enthält u.a. zu tausenden von Künstler:innenbiographien, die mit wenigen Mausklicks – zwar nicht vollständig, aber als belastbare Basis – zur Verfügung stehen. Besonders von Interesse ist dies auch in Hinblick auf Tänzer:innen-Karrieren, über die nur in den seltensten Fällen Forschung existiert. Zwar liegt mit der Datenbank operabuffa.uni-bayreuth.de Forschung zu Akteur:innen von Opere buffe zwischen 1740 und 1765 vor. Diese reichen allerdings nicht über das komische Genre und den genannten Zeitraum hinaus. Die Daten aus der Schatz-Collection bieten hingegen eine Gesamtschau von Oper im 18. und 19. Jahrhundert, womit statt heuristisch vom Einzelfall ausgehend auf das Allgemeine zu schließen, umgekehrt vorgegangen werden kann.

Neben den bisher auf Personen beschränkten Wanderbewegungen, ließe sich damit auch die Mobilität von bestimmten Werken oder Sujets untersuchen, Fragen in Bezug auf die Bedeutung von bestimmten Orten als Ausbildungsstätten könnten daran gestellt werden, anhand der Betrachtung von Widmungsträger:innen ließen sich Mäzenatentum bzw. Kultursponsoring untersuchen. Weitere für die musikhistorische Forschung zu hebende „Schätze“ liegen beispielsweise in der Untersuchung zu Dauern von Karrieren oder auch zu Geschlechterunterschieden bei Karriereverläufen.

Um diese Fragen adäquat adressieren zu können, bedürfen die Daten aber noch einiger Nachbearbeitung und im Idealfall der Anreicherung mit Normdaten zur eindeutigen Identifikation und Verlinkung mit externen Ressourcen.

In einem weiteren Schritt ließen sich die digitalisierten Librettotexte außerdem als Datenkorpus für Distant Reading Projekte in Hinblick auf Genrefragen untersuchen, wie es z.B. Giovannini und Daniil (2023) an einem kleinen Libretto-Bestand in der Dramendatenbank dracor.org bereits versucht haben oder wie es im Projekt „Emotionen im Drama“ als „Sentiment Analysis“ auf einen Bestand von Opernlibretti der Hamburger Gänsemarktoper von 1678 bis 1730 angewandt wurde (Dennerlein et al. 2019). Aufgrund der Größe des zur Verfügung stehenden Korpus, den die digitalisierte Schatz-Collection bildet, bietet es sich insbesondere an, computergestützte Verfahren auszuprobieren oder weiterzuentwickeln. Zwar mit erheblichem Aufwand verbunden, aber auch besonders reizvoll wäre es, eine Verbindung von überlieferten Musikpartituren zu den Libretti herzustellen und die Distant Reading Methoden auch auf Musik zu übertragen. Ein groß angelegtes Projekt könnte in diesem Fall sowohl methodisch wie inhaltlich die Digital Humanities um die musikwissenschaftliche Perspektive ergänzen. In Hinblick auf die 2014 formulierten Thesen des DHd könnte ein derartiges Projekt aufgrund der Interdisziplinarität ebenso entsprechen, wie es das Potential digitaler Methoden voll ausschöpfen könnte, in Hinblick auf „Annotation, Analyse, Aggregation und Rekombination von geisteswissenschaftlichen Objekt- und Metadaten in Kombination von algorithmischen und hermeneutischen Prozeduren”[7] wie es im Thesenpaper gefordert wurde.

Fußnoten

1. Digital Humanities 2020, https://dig-hum.de/thesen-digital-humanities-2020 (zugegriffen: 12.7.23).
2. Digital Humanities 2020.
3. Zettelkatalog von Schatz ebenfalls in der Library of Congress.
4. https://www.loc.gov/marc/relators/relaterm.html (zugegriffen: 19.7.2023).
5. Vgl. die API-Beschreibung unter https://www.loc.gov/apis/json-and-yaml/ (zugegriffen: 19.7.2023).

6. Berechnet mit Gephi 0.10.1 202301172018
7. Digital Humanities 2020, https://dig-hum.de/thesen-digital-humanities-2020 (zugegriffen: 12.7.23).

Bibliographie

Dennerlein, Karin, Thomas Schmidt und Christian Wolff. 2019. „Emotionen im kulturellen Gedächtnis bewahren“, https://zenodo.org/records/6327957 (zugegriffen: 3.12.23)

Foppa, Giuseppe Maria. Le villanelle astute. 1786. Venedig: Fenzo. https://lccn.loc.gov/2010663602 (zugegriffen: 24.5.2023).

Giovannini, Luca, und Skorinkin, Daniil. 2023. “Computational approaches to opera libretti". In Journal of Computational Literary Studies, 2 (im Druck). Preprint: https://github.com/DanilSko/opera/blob/main/giovannini_skorinkin_libretti_2023pp.pdf (zugegriffen: 3.12.23).

McClymonds, Marita P. und Parr Walker, Diane. 1986. “U.S. Rism Libretto Project: With Guidelines for Cataloguing in the MARC Format”. In *Notes* 43/1: 19–35

Needham Costonis, Maureen. 1991. “The U.S. Rism Libretto Project: Accessible Online Database for Researchers”. In *Dance Research Journal* 23/2: 55–57

Sonneck, Oscar George Theodore. 1914. Preface. In *Library of Congress, Catalogue of Opera Librettos printed before 1800*, hg. von Oscar Sonneck. Washington: 1–19. https://archive.org/details/catalogueoperal00sonngoog/ (zugegriffen: 12.7.23).

Strohm, Reinhard. 2001. „Italian Operisti North of the Alps, c. 1700 - c. 1750“. In *The Eighteenth-Century Diaspora of Italian Music and Musicians*, hg. von Reinhard Strohm. Turnhout

Zur Nieden, Gesa. 2015. „Frühneuzeitliche Musikermigration nach Italien: Fragen, Verflechtungen und Forschungsgebiete einer europäischen Kulturgeschichtsschreibung der Musik“. In *Musicisti europei a Venezia, Roma e Napoli / Europäische Musiker in Venedig, Rom und Neapel / Les musiciens européens à Venise, à Rome et à Naples (1650–1750)*, hg. von Anne-Madeleine Goulet und Gesa zur Nieden. Kassel: Bärenreiter (= Analecta Musicologica 52): 9–30

My Body is a Cage: Human Pose Estimation und Retrieval in kunsthistorischen Inventaren

Schneider, Stefanie
stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-4915-6949

Theoretisch fundiert ist die Gestik als Ausdruck nonverbaler Kommunikation seit dem 17. Jahrhundert (Knowlson, 1965). Ihre Relevanz für die bildende Kunst wurde jedoch nur vereinzelt betont (Barasch, 1987), etwa als die Antike rezipierende *Pathosformel* (Warburg, 1998). Diese Punktualität mag nicht nur auf die große Menge an traditionell manuell zu verarbeitenden Daten zurückzuführen sein, sondern ebenso auf das Fehlen eines die Gestik ‚kodifizierenden‘ Vokabulars. Zwar hat sich 1912 der finnische Kunsthistoriker Johan Jakob Tikkanen an einer Typologie kunsthistorischer (Bein-)Stellungsmotive versucht, die er mit potenziellen Entwicklungsketten versah (Tikkanen, 1912). Streng ökonomisch motiviert sind darüber hinaus jedoch allenfalls Studien zur Handgestik, die einige wenige Stereotypen differenzieren (u. a. Bulwer, 1644; Demisch, 1984). Um diesem Desiderat zu begegnen, konzentrieren wir uns auf die quantitativ-fundierte Exploration von Gesten- und Positurtypen in der bildenden Kunst. Die Einreichung knüpft an eigene Vorarbeiten an und erweitert diese (Springstein et al., 2022; Schneider und Vollmer, 2023). Unter Positur (engl. *posture*) wird gewöhnlich eine statische, „bewusst eingenommene *Stellung*“ des Körpers verstanden,[1] im Gegensatz zur Geste (engl. *gesture*), die als dynamische, „bewusst eingesetzte *Bewegung*“ des Körpers definiert wird;[2] auf diese Unterscheidung greifen auch wir, soweit möglich, im Folgenden zurück. Wir verweisen zudem auf Mulder (1996).

Unser Ansatz fußt auf zwei Modulen: Zunächst werden ‚Gelenkpunkte‘ von menschlichen Figuren detektiert (*Human Pose Estimation; HPE*). Diese werden in ‚Deskriptoren‘ überführt und aufgrund ihrer Nähe zueinander typisiert (*Human Pose Retrieval; HPR*). Ein webbasiertes *Retrieval* im zweidimensionalen Raum rundet die Pipeline ab. Es gibt zwar Ansätze zur *HPE* in kunsthistorischen Inventaren, diese fokussieren jedoch auf restriktive Datenkorpora: Impett und Süsstrunk (2016) analysieren Tafeln aus Warburgs Bilderatlas *Mnemosyne*; Madhu et al. (2023) beziehen sich auf griechische Vasenmalerei.

Semi-überwachte Human Pose Estimation

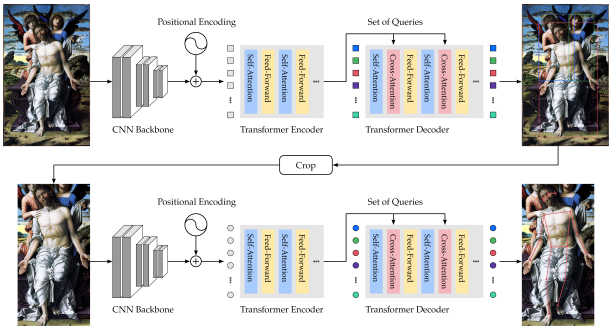


Abb. 1: Unser Ansatz zur *HPE* ist zweistufig: Zunächst werden menschliche Figuren durch *Bounding Boxes* lokalisiert und diese Begrenzungsrahmen dann auf *Keypoints* analysiert (Springstein et al., 2022).

Der hier vorgeschlagene Ansatz zur *HPE* gründet auf der bewährten *Top-Down*-Strategie (Li et al., 2021; Wang, Sun et al., 2021): In einem Bild werden zunächst menschliche Figuren durch *Bounding Boxes* detektiert. Diese Begrenzungsrahmen werden dann auf 17 *Keypoints* untersucht, die Gelenkpunkte des menschlichen Körpers approximieren. Auf diese Weise soll eine maschinell effizient handhabbare Abstraktion des menschlichen Skeletts, und damit der figuralen Gestik oder Positur, erzeugt werden. Abb. 1 zeigt die Gesamtarchitektur des Ansatzes.

Methodik

Die erste Phase basiert auf dem *Detection-Transformer*-Framework (*DETR*; Carion et al., 2020). Ein *Convolutional-Neural-Network*-Backbone berechnet Merkmalsdeskriptoren, die durch ein *Positional Encoding* angereichert werden. Dieser Input wird umgewandelt in eine Sequenz visueller Merkmale und in einen *Transformer*-Encoder gespeist; der Output des Encoders wird in den *Cross-Attention*-Modulen des *Transformer*-Decoders verwendet. Nach der Verarbeitung durch den Decoder wird die Ausgabe in zwei *Multilayer-Perceptron*-Köpfe geleitet: Der erste Kopf fungiert als Klassifikator, der zwischen Figur und Bildhintergrund unterscheidet; der zweite führt eine Regression auf die Koordinaten der jeweiligen *Bounding Box* durch. Das Vorgehen in der zweiten Phase ist äquivalent, nur dass hier der Kopf für jede zuvor identifizierte *Bounding Box* die Koordinaten der 17 *Keypoints* vorhersagt.

Um das jeweils verfügbare Trainingsmaterial in beiden Phasen zu erweitern, integrieren wir einen Ansatz des *semi-überwachten Lernens (Semi-supervised Learning; SSL)*, der auf dem von Xu et al. (2021) motivierten Lehrer-Schüler-Paradigma aufsetzt. In diesem übernimmt der Lehrer, dessen Gewichte aus dem *Exponential Moving Average* des Schülers abgeleitet werden (Tarvainen et al., 2017), die Rolle eines *Pseudo-Label*-Generators: Er generiert *Bounding-Box*- und *Keypoint*-Annotationen für unbeschriftete Daten.

Daten

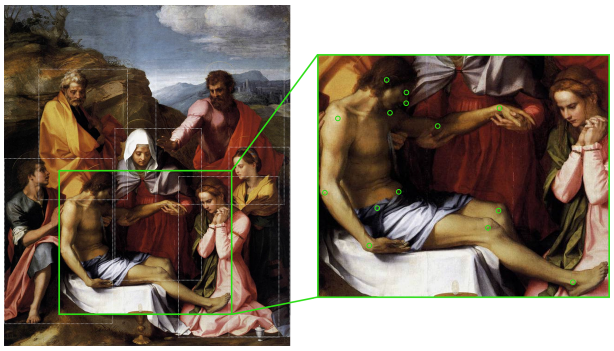


Abb. 2: Es lassen sich *Bounding-Box*- und *Keypoint*-Annotationen unterscheiden. Wie in Andrea del Sartos *Pietà mit Heiligen* (1523–24) gezeigt, werden menschliche Figuren zunächst von *Bounding Boxes* umschlossen. Dann werden bis zu 17 *Keypoints* zugewiesen, die in der Detailansicht durch grüne Kreise gekennzeichnet sind.

Für das Training der Modelle werden fünf Datensätze verwendet; vier sind beschriftet, drei mit *Keypoint*-Annotationen: Als realweltliche Datengrundlage dient *COCO 2017* (123.287 Bilder; Lin et al., 2014). Um die Effizienz von *Style-Transfer-(ST)*-Ansätzen zu evaluieren, generieren wir zusätzlich eine stilisierte Version, die dem jeweiligen Modell anteilig zugeführt wird (Chen et al., 2021). Kunsthistorisches Material fließt zum einen über den *People-Art*-Datensatz ein, der *Bounding Boxes* von menschlichen Figuren annotiert (4.851 Bilder; Westlake et al., 2016). Zum anderen wird der von uns in Schneider und Vollmer (2023) eingeführte *PoPArt*-Datensatz integriert, der ebenfalls *Keypoints* auf 2.454 Bildern enthält. Alle Datensätze folgen dem Microsoft *COCO*-Format, in dem bis zu 17 *Keypoints* pro Figur zusätzlich zu *Bounding Boxes* gespeichert werden (Lin et al., 2014). Es gibt fünf *Keypoints* für den Kopf, die Nase, Augen und Ohren repräsentieren; sechs für den Oberkörper, die Handgelenke, Ellbogen und Schultern repräsentieren; und sechs für den Unterkörper, die Knöchel, Knie und Hüften repräsentieren (Abb. 2). Unbeschriftete Daten stammen aus *ART500K* (318.869 Bilder; Mao et al., 2017).

Evaluation

Test Set	Training Set(s)	ST	SSL	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR
People-Art	COCO 2017	0 %		0.312	0.511	0.318	0.008	0.212	0.329	0.673
	COCO 2017	0 %	✓	0.370	0.597	0.389	0.001	0.212	0.395	**0.735**
	COCO 2017	50 %		0.369	0.611	0.387	0.005	0.239	0.394	0.726
	COCO 2017	50 %	✓	0.374	0.628	0.379	0.002	0.219	0.401	0.730
	COCO 2017	100 %		0.373	0.626	0.392	**0.024**	**0.241**	0.398	0.717
	COCO 2017	100 %	✓	**0.385**	**0.633**	**0.405**	0.012	0.231	**0.411**	0.722
	People-Art	0 %		0.428	0.728	0.435	**0.068**	0.212	0.464	0.704
	People-Art	0 %	✓	**0.443**	**0.738**	**0.459**	0.051	**0.241**	**0.477**	**0.729**

Tab. 1: Ergebnisse der ersten Phase der *HPE*, in der *Bounding Boxes* menschlicher Figuren detektiert werden.

Test Set	Training Set(s)	ST	SSL	AP	AP_{50}	AP_{75}	AP_5	AP_M	AP_L	AR
PoPArt	COCO 2017	0 %		0.468	0.574	0.523		**0.336**	0.471	0.523
	COCO 2017	0 %	✓	**0.610**	**0.771**	**0.671**		0.287	**0.619**	**0.697**
	COCO 2017	50 %		0.566	0.723	0.625		0.306	0.575	0.663
	COCO 2017	50 %	✓	0.592	0.752	0.649		0.311	0.601	0.683
	COCO 2017	100 %		0.573	0.741	0.633		0.320	0.581	0.663
	COCO 2017	100 %	✓	0.592	0.747	0.663		0.282	0.602	0.673
	People-Art, PoPArt	0 %		0.680	0.865	0.753		**0.354**	0.687	0.784
	People-Art, PoPArt	0 %	✓	**0.740**	**0.884**	**0.799**		0.266	**0.754**	**0.813**

Tab. 2: Ergebnisse der zweiten Phase der *HPE*, in der für jede identifizierte *Bounding Box* die Koordinaten von 17 *Keypoints* vorhergesagt werden.

Für die verwendeten Modellparameter wird auf Springstein et al. (2022) verwiesen. Wie aus Tab. 1 ersichtlich, verbessert *SSL* die *Bounding-Box*-Erkennung wesentlich sowohl in Bezug auf *Average Precision (AP)* als auch *Average Recall (AR)*. Mit AP_{50} = 0,738 ist die Leistung unseres Ansatzes für *People-Art* zudem deutlich höher als die von Kadish et al. (2021) mit AP_{50} = 0,68 und als die von Gonthier et al. (2022) mit AP_{50} = 0,583. Noch ausgeprägter ist der Unterschied in der *Keypoint*-Schätzung (Tab. 2).[3] Ebenso zeigt sich, dass es zwar nicht notwendig ist, große Mengen an domänenspezifischem Material zu annotieren, aber kleinere Mengen in den Trainingsprozess einbezogen werden sollten, anstatt sich – wie in Madhu et al. (2023) – auf synthetisch generierte Bilder zu stützen.

Blickwinkel-invariantes *Human Pose Retrieval*

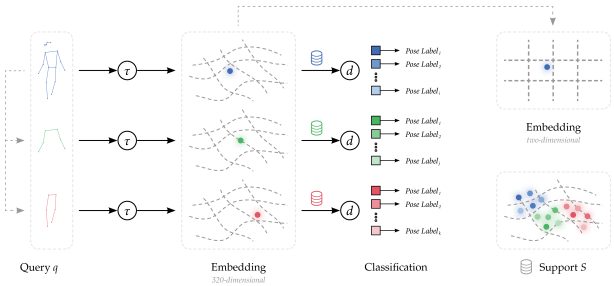


Abb. 3: Der *HPR*-Ansatz besteht aus drei Schritten: Zunächst wird eine *Query* gefiltert und in ein 320-dimensionales *Embedding* überführt (Sun et al., 2020). Dieses *Embedding* wird dann mit Hilfe einer *Support*-Menge klassifiziert.

Unser dreistufiger *HPR*-Ansatz baut direkt auf der *HPE* auf: Die *Keypoints* werden hier in semantisch-plausible Gestendeskriptoren übersetzt und mit Hilfe einer kleinen *Support*-Menge typisiert. Abb. 3 stellt die Gesamtarchitektur dar.

Methodik

Ausgehend von der *HPE* des Ganzkörperskeletts werden Ober- und Unterkörper zusätzlich getrennt abgelegt; diese ‚Konfigurationen' ergeben die *Query* q . In einem *Pre-Processing*-Schritt werden zunächst Konfigurationen mit hoher Unsicherheit entfernt, d. h. solche, die weniger als τ = 0,5 der jeweils möglichen *Keypoints* haben. Alle verbleibenden, hinreichend sicheren Konfigurationen werden in ein 320-dimensionales *Embedding* überführt, das als Gestendeskriptor fungiert. Bisherige Verfahren integrieren dazu entweder Informationen über die absolute Position der *Keypoints* (So und Baciu, 2005) oder über winkelbasierte Maße zwischen ihnen (Chen et al., 2011). In beiden Fällen sind die erzeugten *Embeddings* nicht Blickwinkel-invariant: Eine kauernde Figur würde, wenn sie einmal von vorne und einmal von hinten dargestellt wird, nicht auf den gleichen Punkt im *Embedding Space* abgebildet. Um diesen für das *HPR* gravierenden Mangel abzuschwächen, adaptieren wir die *Pr-VIPE*-Architektur von Sun et al. (2020), in der probabilistische *Embeddings* durch Blickwinkel-augmentierte *Keypoints* im zweidimensionalen Raum gelernt werden. Der Ansatz zielt darauf ab, dass das *Embedding* von zweidimensionalen Gesten in einem hochdimensionalen Raum den Abstand zwischen dreidimensionalen Gesten im euklidischen Raum widerspiegelt. Mit anderen Worten: Wenn zwei dreidimensionale Gesten einander ähnlich sind, sollten ihre zweidimensionalen Pendants im *Embedding Space* nahe beieinander liegen. Die Ähnlichkeit dreidimensionaler Gesten beruht auf ihrer visuellen Ähnlichkeit unter Berücksichtigung der menschlichen Wahrnehmung; zwei Gesten können mathematisch unterschiedlich sein, aber je nach Betrachtungswinkel visuell ähnlich erscheinen.

Jedes *Embedding* wird anschließend klassifiziert. Da bislang kein Datensatz vorliegt, der kunsthistorisch bedeutsame Gesten benennt und illustriert, konstruieren wir eine Taxonomie auf Basis von *Iconclass* (van de Waal, 1973–1985). Sie besteht aus vier Notationsgruppen: „postures of the human figure" (31A23), „postures and gestures of arms and hands" (31A25), „postures of the legs" (31A26) und „movements of the human body" (31A27). Notationen unter 31A23 und 31A27 dienen der Klassifizierung des Ganzkörperskeletts, 31A25 des Ober- und 31A26 des Unterkörpers. Die Notationen für den Oberkörper (31A25) und den Unterkörper (31A26) sind mit 22 bzw. 19 annähernd gleich häufig. Bei den Ganzkörpernotationen gibt es jedoch eine Diskrepanz: Notation 31A23 hat 19 und Notation 31A27 nur 8 verwendbare Unternotationen, sodass sich insgesamt 27 Unternotationen ergeben. Unser Vorgehen orientiert sich am Prinzip des *One-shot-Lernens (OSL)*: Für die insgesamt 69 Subnotationen identifizieren wir jeweils ein repräsentatives Bildbeispiel einer Figur in *Wikidata*,[4] erstellen ihre *Ground-Truth*-Annotation und generieren ihr *Embedding*. D. h. anstatt wie in typischen *OSL*-Ansätzen einen *One-Shot*-Klassifikator separat zu trainieren (u. a. Jadon et al., 2020), nutzen wir die *Pr-VIPE-Embeddings* nach und berechnen die Abstände zwischen den *Embeddings*. Diese *Support*-Menge S wird verwendet zur Typisierung der Konfigurationen; die Kosinusdistanz d misst den Abstand zwischen dem jeweiligen *Query-Embedding* und den *Embeddings* der *Support*-Menge. Dies ermöglicht eine feingranulare Erschließung der Gestik oder Positur, auch wenn Teile einzelner Konfigurationen nur unzureichend geschätzt werden konnten. Gleichzeitig wird keine feste, semantisch zweifelhafte Kategorisierung in Gruppen vorgegeben, wie dies bei agglomerativen Clusterverfahren der Fall ist (Impett und Süsstrunk, 2016).

Daten

Wir extrahieren 644.155 kunsthistorische Objekte durch Abfragen des *Wikidata-SPARQL*-Endpunkts.[5] Um *Query Timeouts* zu vermeiden, gehen wir iterativ vor: Zuerst werden 171 ‚Klassenentitäten' extrahiert, die direkte Unterklassen der Knoten „visual artwork" (wdt: Q4502142) oder „artwork series" (wdt:Q15709879) sind. Für jede Klassenentität werden dann ‚Objektentitäten' abgefragt, denen eine zweidimensionale Reproduktion (wdt:P18) zugeordnet ist und die entweder Instanzen dieser Klassenentität oder Unterklassen davon sind. Zwar ist nicht auszuschließen, dass auch *Wikidata* mehrere Knoten für dasselbe Objekt führt und somit mehr als eine Reproduktion nach demselben Original zurückliefert. Unseres Erachtens ist der Anteil jedoch geringer als bei Aggregatdatenbanken wie Prometheus.[6]

Evaluation

Mangels eines annotierten Testdatensatzes ist die Evaluation des *HPR* im Gegensatz zur *HPE* rein qualitativer Natur. Um dennoch eine möglichst verlässliche Aussage über die Güte des verwendeten Ansatzes treffen zu können, untersuchen wir den erzeugten *Embedding Space* auf Aggregat- und Individualebene. Die 644.155 Objekte aus *Wikidata* durchlaufen die gesamte Pipeline von *HPE* und *HPR*; 385.481 werden mit 2.355.592 Figuren als potenziell relevant erkannt.

Aggregatebene

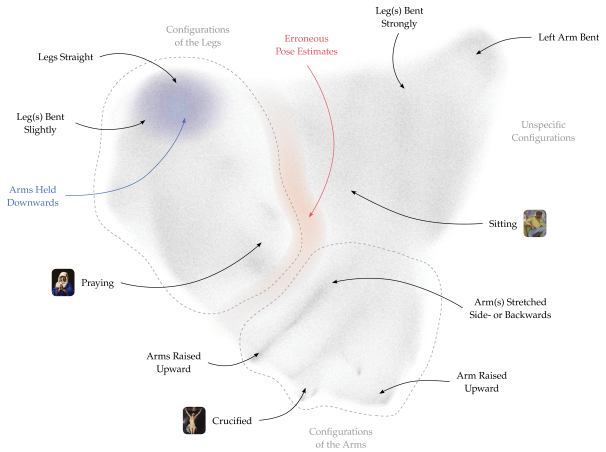


Abb. 4: Im dimensionsreduzierten *Embedding Space* fallen zwei marginal abgetrennte Gruppen auf, die insbesondere Konfigurationen des Ober- und Unterkörpers referenzieren.

Die Auswertung des *Embedding Space* erfolgt durch eine Reduktion der 320 Dimensionen des Ganzkörperskeletts auf zwei. Gängige Methoden zur Dimensionsreduktion wie *t-SNE* (van der Maaten und Hinton, 2008) oder *UMAP* (McInnes et al., 2018) fokussieren entweder auf die Erhaltung lokaler oder globaler Strukturen, so dass häufig falsche *Cluster* projiziert werden, die im hochdimensionalen Raum nicht existieren. Wir verwenden daher *Pairwise Controlled Manifold Approximation Projection* (*PaCMAP*; Wang, Huang et al., 2021). Abb. 4 zeigt den so reduzierten *Embedding Space*, den wir auf Basis von *PixPlot* auch interaktiv explorierbar machen.[7] Es sind zwei annähernd clusterartige Strukturen erkennbar, die vor allem Konfigurationen des Ober- und Unterkörpers entsprechen, und damit spezifischeren Arm- und Beinhaltungen, die mit Hilfe der *Iconclass*-annotierten *Support*-Menge typisiert werden konnten. Deutlich wird jedoch, dass es sich bei der Typisierung lediglich um eine Hilfskonstruktion handelt, die die Interaktion im *Embedding Space* erleichtern und mögliche Clusterbildungen schneller identifizieren soll. Insbesondere Haltungen mit stärker gebeugten Gliedmaßen – hockende, kauernde oder sitzende Figuren – bilden eine dritte Gruppe, die semantisch mehrdeutig zu erfassen ist. Falsche Schätzungen des Ganzkörperskeletts finden sich am häufigsten in den schwach besetzten Zwischenräumen, die zur Mitte konvergieren.

Individualebene

Abb. 5: *Retrieval*-Ergebnisse für die links abgebildete Figur aus James Tissots *Le Coup de Lance* (1886–1894) mit den jeweils geschätzten *Keypoints* in *grün*.

Für das *HPR* einzelner Gesten oder Posituren wird eine Indexstruktur erstellt, in die die 320-dimensionalen *Embeddings* der Figuren geladen werden. Wir verwenden mit *Hierarchical Navigable Small World* (*HNSW*; Malkov und Yashunin, 2020) einen *Approximate-k-Nearest-Neighbor*-Ansatz mit polylogarithmischer Komplexität, der andere graphbasierte Ansätze wie *Faiss* (Johnson et al., 2021) in *Precision* und *Recall* übertrifft (Aumueller et al., 2023). Als Beispiel- *Query* filtern wir eine Figur aus James Tissots *Le Coup de Lance* (1886–1894): den gekreuzigten Schächer zur Rechten Christi. In Abb. 5 ist eine Auswahl von *Retrieval*-Ergebnissen mit geringer Distanz zur *Query* dargestellt. Es dominieren naturgemäß vor allem Figuren aus Kreuzigungsgruppen, wenn auch meist mit Oberkörper-Konfigurationen in klassischer *T*- oder leichter *Y*-Form. Die angewinkelten Arme des Schächers werden in Pietà-Darstellungen aufgegriffen, z. B. in der Kopie nach Marcello Venusti (erste Zeile, fünftes Bild von links). Eine interessante Fehlschätzung findet sich in Jacques Louis Davids *Napoleon am Großen St. Bernhard* (1801; dritte Zeile,

drittes Bild von links): Der vordere Teil des aufgerichteten Pferdes wird in der *HPE* fälschlicherweise als Figur mit nach hinten gestreckten Armen und angewinkelten Beinen erkannt – eine Konfiguration, die im *HPR* große Ähnlichkeit mit der des Schächers hat.

Fazit und Ausblick

Obwohl das *HPR* nur qualitativ auf der Aggregatebene durchgeführt werden konnte – oder auf der Individualebene exemplarisch am Beispiel des gekreuzigten Schächers in Tissots *Le Coup* – wird deutlich, dass der Ansatz in Kombination mit einer semi-überwachten *HPE* eine vielversprechende Basis schafft für die quantitativ-fundierte Exploration von Gestentypen in der bildenden Kunst: Das menschliche Skelett wird durch ein Blickwinkel-invariantes 320-dimensionales *Embedding* ganzheitlich erfasst. Indem neben dem Ganzkörperskelett auch Ober- und Unterkörper separat abgelegt werden, lassen sich Gestik oder Positur feingranular erschließen und typisieren, auch wenn einzelne Konfigurationen nur unzureichend geschätzt werden.

Es ist geplant, die Pipeline anhand von zwei disparaten Anwendungsfällen kunsthistorisch näher zu evaluieren: der kompositionell restriktiven Ikonographie des Sündenfalls und der zeitlich dynamischer variierenden Kreuzabnahme Christi. Beide lassen sich auf dominante Gestentypen oder zeitabhängige Phänomene hin untersuchen, wie sie für den Manierismus durch die Überstreckung der Gliedmaßen charakteristisch sind. Intra- und interikonografisch wiederkehrende Motive, deren teils radikal veränderte Semantik befremdet, sind in diesem Zusammenhang zu diskutieren.

Danksagung

Diese Arbeit wurde teilweise von der Deutschen Forschungsgemeinschaft (DFG) unter der Projektnummer 415796915 gefördert.

Fußnoten

1. https://www.duden.de/node/113566/revision/1343578, wie alle URLs zugegriffen: 9. November 2023
2. https://www.duden.de/node/57136/revision/1454368.
3. Die Auswertung erfolgt auf dem vollständig mit *Bounding Boxes* annotierten *PoPArt*-Datensatz, im Gegensatz zu Springstein et al. (2022).
4. https://www.wikidata.org.
5. https://query.wikidata.org/bigdata/namespace/wdq/sparql.
6. https://prometheus-bildarchiv.de/de.
7. https://github.com/YaleDHLab/pix-plot.

Bibliographie

Aumueller, Martin, Erik Bernhardsson und Alec Faitfull. 2023. *ANN Benchmarks*. https://ann-benchmarks.com (zugegriffen: 19. Juli 2023).

Barasch, Moshe. 1987. *Giotto and the Language of Gesture*. Cambridge: Cambridge University Press.

Bulwer, John. 1644. *Chirologia. Or the Naturall Language of the Hand.* London: Thomas Harper.

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov und Sergey Zagoruyko. 2020. „End-to-end Object Detection with Transformers.“ In *Computer Vision – ECCV 2020. Lecture Notes in Computer Science* 12346: 213–229 10.1007/978-3-030-58452-8_13.

Chen, Cheng, Yueting Zhuang, Feiping Nie, Yi Yang, Fei Wu und Jun Xiao. 2011. „Learning a 3D Human Pose Distance Metric from Geometric Pose Descriptor.“ *IEEE Transactions on Visualization and Computer Graphics* 17.11: 1676–1689 10.1109/TVCG.2010.272.

Chen, Haibo, Lei Zhao, Zhizhong Wang, Zhang Hui Ming, Zhiwen Zuo, Ailin Li, Wei Xing und Dongming Lu. 2021. „Artistic Style Transfer with Internal-external Learning and Contrastive Learning.“ In *35th Conference on Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2021/file/df5354693177e83e8ba089e94b7b6b55-Paper.pdf (zugegriffen: 19. Juli 2023).

Demisch, Heinz. 1984. *Erhobene Hände. Geschichte einer Gebärde in der bildenden Kunst.* Stuttgart: Urachhaus.

Gonthier, Nicolas, Saïd Ladjal und Yann Gousseau. 2022. „Multiple Instance Learning on Deep Features for Weakly Supervised Object Detection with Extreme Domain Shifts.“ *Computer Vision and Image Understanding* 214 10.1016/j.cviu.2021.103299.

Impett, Leonardo und Sabine Süsstrunk. 2016. „Pose and Pathosformel in Aby Warburg’s Bilderatlas.“ In *Computer Vision – ECCV 2016 Workshops. Lecture Notes in Computer Science* 9913: 888–902 10.1007/978-3-319-46604-0_61.

Jadon, Shruti und Aryan Jadon. 2020. *An Overview of Deep Learning Architectures in Few-shot Learning Domain*. arXiv:1412.6980.

Johnson, Jeff, Matthijs Douze und Herve Jegou. 2021. „Billion-scale Similarity Search with GPUs.“ *IEEE Transactions of Big Data* 7: 535–547 10.1109/TBDATA.2019.2921572.

Kadish, David, Sebastian Risi und Anders Sundnes Løvlie. 2021. „Improving Object Detection in Art Images Using Only Style Transfer.“ In *International Joint Conference on Neural Networks. IJCNN 2021*, 1–8 10.1109/IJCNN52387.2021.9534264.

Knowlson, James R. 1965. „The Idea of Gesture as a Universal Language in the XVIIth and XVIIIth Centuries.“ *Journal of the History of Ideas* 26: 495–508.

Li, Ke, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu und Zhuowen Tu. 2021. „Pose Recognition with Cascade Transformers." In *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2021*, 1944–1953.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár und C. Lawrence Zitnick. 2014. „Microsoft COCO. Common Objects in Context." In *Computer Vision – ECCV 2014. Lecture Notes in Computer Science* 8693: 740–755 10.1007/978-3-319-10602-1_48.

van der Maaten, Laurens und Geoffrey Hinton. 2008. „Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (zugegriffen: 19. Juli 2023).

Madhu, Prathmesh, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas K. Maier und Vincent Christlein. 2023. „Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-grounded Style Transfer Learning." *ACM Journal on Computing and Cultural Heritage* 16.1: 1–17 10.1145/3569089.

Malkov, Yu A. und D. A. Yushunin. 2020. „Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.4: 824–836 10.1109/TPAMI.2018.2889473.

Mao, Hui, Ming Cheung und James She. 2017. „DeepArt. Learning Joint Representations of Visual Arts." In *MM '17. The 25th ACM International Conference on Multimedia*, 1183–1191 10.1145/3123266.3123405.

McInnes, Leland, John Healy, Nathaniel Saul und Lukas Großberger. 2018. „UMAP. Uniform Manifold Approximation and Projection." *Journal of Open Source Software* 3.29 10.21105/joss.00861.

Mulder, Axel. 1996. *Hand Gestures for HCI*. Vancouver: Simon Fraser University.

Schneider, Stefanie und Ricarda Vollmer. 2023. *Poses of People in Art. A Data Set for Human Pose Estimation in Digital Art History*. arXiv:2301.05124.

So, Clifford Kwok-Fung und George Baciu. 2005. „Entropy-based Motion Extraction for Motion Capture Animation." *Computer Animation and Virtual Worlds* 16.3–4: 225–235 10.1002/cav.107.

Springstein, Matthias, Stefanie Schneider, Christian Althaus und Ralph Ewerth. 2022. „Semi-supervised Human Pose Estimation in Art-historical Images." In *MM '22. The 30th ACM International Conference on Multimedia*, 1107–1116 10.1145/3503161.3548371.

Sun, Jennifer J., Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam und Ting Liu. 2020. „View-invariant Probabilistic Embedding for Human Pose." In *Computer Vision – ECCV 2020. Lecture Notes in Computer Science* 12350: 53–70 10.1007/978-3-030-58558-7_4.

Tarvainen, Antti und Harri Valpola. 2017. „Mean Teachers are Better Role Models. Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results." In *5th International Conference on Learning Representations. ICLR 2017*.

Tikkanen, Johan Jakob. 1912. *Die Beinstellungen in der Kunstgeschichte. Ein Beitrag zur Geschichte der künstlerischen Motive.* Helsingfors: Druckerei der finnischen Litteraturgesellschaft.

van de Waal, Henri. 1973–1985. *Iconclass. An Iconographic Classification System. Completed and Edited by L. D. Couprie with R. H. Fuchs*. Amsterdam: North-Holland Publishing Company.

Wang, Jingdong, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu und Bin Xiao. 2021. „Deep High-resolution Representation Learning for Visual Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10: 3349–3364 10.1109/TPAMI.2020.2983686.

Wang, Yingfan, Haiyang Huang, Cynthia Rudin und Yaron Shaposhnik. 2021. „Understanding How Dimension Reduction Tools Work. An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization." *Journal of Machine Learning Research* 22.201: 1–73. https://jmlr.org/papers/v22/20-1061.html (zugegriffen: 19. Juli 2023).

Warburg, Aby. 1998 [1905]. „Dürer und die italienische Antike." In *Die Erneuerung der heidnischen Antike. Kulturwissenschaftliche Beiträge zur Geschichte der europäischen Renaissance. Gesammelte Schriften*, hg. von Horst Bredekamp und Michael Diers, 443–449. Berlin: Akademie Verlag.

Westlake, Nicholas, Hongping Cai und Peter Hall. 2016. „Detecting People in Artwork with CNNs." In *Computer Vision – ECCV 2016 Workshops. Lecture Notes in Computer Science* 9913: 825–841 10.1007/978-3-319-46604-0_57.

Xu, Mengde, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai und Zicheng Liu. 2021. „End-to-end Semi-supervised Object Detection with Soft Teacher." In *IEEE/CVF International Conference on Computer Vision. ICCV 2021*, 3040–3049 10.1109/ICCV48922.2021.00305.

Normdaten Quo Vadis

Rettinghaus, Klaus

klaus.rettinghaus@enote.com
Enote GmbH, Deutschland
ORCID: 0000-0003-1898-2543

Normdaten bilden heute einen zentralen Pfeiler der DH; quasi kein Projekt kommt ohne sie aus. Aus bibliothekarischen Bedürfnissen erwachsen (hauptsächlich zum Zwecke der Disambiguierung), pflegen heute nahezu alle Nationalbibliotheken weltweit ihre eigenen Normdatenbanken.[1] Das Online Computer Library Center Inc. (OLCL) bietet

mit dem Virtual International Authority File (VIAF)einen Dienst, der die internationalen Identifikatoren miteinander verknüpft.[2] Gerade für die Digital Humanities im deutschsprachigen Raum ist die Gemeinsame Normdatei (GND) der Deutschen Nationalbibliothek die zentrale Anlaufstelle für Normdaten. Dennoch äußern sich auch heute noch gerade Fachwissenschaftler zumeist kritisch über die GND und bemängeln fehlerhafte oder fehlende Daten. Zeit, deren Geschichte, Datenqualität und Kollaborationen mit der Community erneut zu beleuchten.

Geschichte der GND

Die GND führte 2012 die bis dahin getrennt geführten Verzeichnisse für Personen, Organisationen, Schlagworte, etc. zusammen und ersetzte sie gleichfalls. (Scheven und Ohde, 2019) Den umfangreichsten Teil bildete dabei die Personennamendatei (PND), die zu diesem Zeitpunkt nahezu 2 Millionen Namen umfasste und die ab 1994 auf Empfehlung des Bibliotheksausschusses der Deutschen Forschungsgemeinschaft (DFG) an der Deutschen Nationalbibliothek entwickelt wurde.[3]

Seit 2018 läuft das erneut von der DFG geförderte Projekt „GND4C – GND für Kulturdaten“, dessen Ziel die spartenübergreifende Öffnung der GND für nicht-bibliothekarische Einrichtungen wie Museen, Archive, Universitätssammlungen, Denkmalbehörden, wissenschaftliche Institutionen und Mediatheken voranzubringen.[4]

Einzelne Projekte wie z.B. Bach digital oder musicon-n.performance haben ihren Blick auf die Ergänzung und Verbesserung von Werknormdaten zur Musik gerichtet.[5] (Bicher und Wiermann, 2018) Auch beim Répertoire International des Sources Musicales (RISM), dem Internationalen Quellenlexikon der Musik, wird seit einiger Zeit über Werknormdaten nachgedacht.[6]

Zukünftig soll die (möglichst automatisierte) Einspielung von mehr Informationen aus Forschungsprojekten durch die NFDI-Konsortien gesichert werden. (Kailus, 2022) Hierbei ist insbesondere auf die Datenqualität zu achten. Zwar ist davon auszugehen, dass die Forschenden das meiste Wissen zu Personen, Ereignissen und Geografika zusammengetragen haben und als die Experten anzusehen sind, doch muss möglichst verhindert werden, dass die GND durch weitere Dubletten Qualitätseinbußen zu beklagen hat.[7]

Die einfache Vernetzungsmöglichkeit zwischen Informationsanbietern per Normdaten-Identifikatoren ist dem Verein Wikimedia Deutschland zu verdanken, der die Entwicklung und Verbreitung des simplen BEACON-Formats vorangetrieben hat.[8] (Schindler, 2011) (Voss et al., 2011)

Daten finden

Für die Recherche in der GND stehen der Katalog der Deutschen Nationalbibliothek (DNB-OPAC) bereit, die OGND des Bibliotheksservice-Zentrums Baden-Württemberg (BSZ) sowie der Dienst lobid-gnd des Hochschulbibliothekszentrums des Landes NRW (hbz), wobei der letztere auch eine Linked-Open-Data-Schnittstelle (LOD-API) und einen Reconciliation-Service für OpenRefine (Steeg und Pohl, 2021) anbietet.

Keine Einsicht in den Gesamtbestand der GND bietet Wikidata, doch in der Regel sind hier erfasste Entitäten[9] mit einer GND-ID versehen. Wird man nicht mit einem der oben genannten Tools fündig, was z.B. an abweichenden Ansetzungsformen liegen mag, kann man bei Wikipedia nachsehen, ob zu der entsprechenden Person ein Eintrag existiert. Falls dies der Fall ist und zudem Identifikatoren anderer Normdatenbanken dort verzeichnet sind, jedoch keine der GND, ist dies ein recht sicheres Indiz, dass für diese Person noch kein Eintrag in der GND vorliegt.

Korrekturwünsche

Personendaten sind nach wie vor die wichtigsten Normdaten innerhalb der GND für die Geisteswissenschaften, deswegen sind viele DH-Projekte sehr an der Korrektheit dieser Daten interessiert. Es existieren verschiedene Möglichkeiten den GND-Redaktionen Fehler zu melden und die Datenqualität sukzessiv zu erhöhen.

Der einfachste und direkteste Weg ist sicherlich jener, direkt aus der Ansicht des jeweiligen Datensatzes im Katalog der Deutschen Nationalbibliothek heraus. Dort findet sich im „Aktionen“-Menü der Punkt „Korrekturanfrage“, über die man per Kontaktformular eine „Korrekturanfrage für den Personendatenbestand der Gemeinsamen Normdatei“ absetzen kann. Der Nachteil an dieser Methode ist jedoch, dass man keinerlei Rückmeldung über etwaige Korrekturen erhält und man die jeweiligen beanstandeten Datensätze selbst im Auge behalten muss. Zudem ist es sehr mühselig, falls man mehrere Korrekturwünsche anmelden möchte.

Etwas leichter ist es, sich (per Mail) direkt an eine der GND-Redaktionen zu wenden. Das erlaubt es seine Anfragen zu bündeln und mit zusätzlichen Informationen, wie z.B. detaillierten Quellenangaben zu versehen. Hierbei ist die jeweilige Zuständigkeit zu beobachten. Für Korrekturen an GND-Datensätzen von Personen mit einer Erstveröffentlichung bis einschließlich 1850 ist beispielsweise die Bayerische Staatsbibliothek (BSB) in München zuständig, für Personen die in Österreich ab 1800 veröffentlicht haben die Österreichische Nationalbibliothek (ÖNB) in Wien.[10] Hier ist die Wahrscheinlichkeit eine persönliche Rückmeldung zu erhalten zwar ungleich höher, eine Sicherheit dafür gibt es aber leider auch nicht.

Mittlerweile ist es auch möglich über das eingerichtete GND-Webformular Datensätze zu ändern oder neue hinzuzufügen. Die Zugangsvoraussetzungen hier könnten dennoch eine Hürde sein, denn selbstverständlich möchte die Deutsche Nationalbibliothek keinen generellen Zugang eröffnen.

Bereits seit 2005 werden über die deutschsprachige Wikipedia Dubletten und Fehler in den Personendaten an die Nationalbibliothek gemeldet.[11] Insbesondere bei der ein-

deutigen Identifizierung von nicht-individualisierten Datensätzen war die Wikipedia-Community behilflich.[12] Die kollaborative Verbesserung der Normdaten wird heute hauptsächlich über Wikipedia (und Wikidata) vorangetrieben. Monat für Monat wird von Wikipedianern eine Liste mit GND-Fehlermeldungen zusammengestellt und – je nach Zuständigkeit – an die GND-Redaktion in München und die GND-Zentrale in Frankfurt am Main gesandt. Diese Listen werden dann bearbeitet und für abgeschlossene Listen erfolgt eine Rückmeldung an die Wikipedianer.[13] Die zugehörigen Artikel in Wikipedia und Einträge in Wikidata werden dann aktualisiert und die ursprünglichen Anfragen mit einem der Rückmeldung entsprechenden Hinweis versehen, d.h. ob die gewünschte Korrektur auf Seiten der GND durchgeführt oder abgelehnt wurde. Falls eine positive Rückmeldung der GND erfolgt, die Änderungen aber noch nicht im DNB-Portal sichtbar sind oder anderweitig vom erwarteten Ergebnis abweichen, wird eine Extramarkierung gesetzt und in der Regel erfolgt dann eine erneute Meldung.

Die monatlichen Listen umfassen regelmäßig etwa zwischen 250 und 450 Korrekturwünsche. Die Meldungen der vergangenen Monate und Jahre werden stets archiviert und bleiben so für die Nachwelt erhalten. Zwar dauert die Abarbeitung der Meldungen eine Weile,[14] dafür bekommt man eine nachvollziehbare Rückmeldung (sowie kostenlose Betreuung durch andere Wikipedianer).

Wikipedianer, die an einer speziellen Schulung teilgenommen haben, können seit 2016 über das oben erwähnte GND-Webformular auch direkt Personendatensätze neu anlegen bzw. korrigieren.[15]

Mit Stichtag 15. Oktober 2023 sind in der deutschsprachigen Wikipedia 488.600 Artikel zu Personen mit GND-IDs versehen; das entspricht 53,36% aller Personenartikel.[16]

Fazit

Es ist bemerkenswert, dass der mögliche Mehrwert durch die Auszeichnung mit Normdaten gerade in der Wikipedia-Community erkannt wurde und massiv vorangetrieben wurde. Vor allem für Fachwissenschaftler, die innerhalb von Projekten im Laufe ihrer Forschungstätigkeit regelmäßig Fehler in der GND entdecken und diese korrigiert sehen möchten, empfiehlt sich hier die Arbeit in und mit der Wikipedia. Falls nur eine einmalige Meldung von einigen wenigen Datensätzen erfolgen soll, ist es vermutlich einfacher sich direkt an die jeweilige GND-Redaktion zu wenden.

Im Rahmen des Vortrages werden die GND-Fehlermeldungen der Wikipedia und Möglichkeiten zur Mitarbeit genauer vorgestellt und eine detaillierte statistische Auswertung der Meldungen an sich sowie der erfolgten Rückmeldungen der GND-Redaktionen präsentiert werden, die zur weiteren Beteiligung einladen sollen. Zudem werden verschiedene projektspezifische Vorschläge zur aktiven und konstruktiven Mitarbeit and der GND unterbreitet und diskutiert.

Fußnoten

1. Genannt seien hier als internationale Beispiele die Library of Congress Name Authority File (NAF) der US-amerikanischen Library of Congress (LOC) und der Identifiants et Référentiels (IdRef) des französischen Verbundkatalogs Système universitaire de documentation (SUDOC).
2. Von der Verwendung der Identifikatoren der VIAF innerhalb von wissenschaftlichen Projekten ist abzuraten, da die Datensätze und Verknüpfungen automatisiert erstellt werden und daher die URIs – gleichwohl als „Permalink“ bezeichnet – nicht persistent sind. Nur für manche Identifikatoren wird eine Weiterleitung eingerichtet, andere einfach entfernt, sodass eine VIAF ID im ungünstigen Fall eine HTTP 404 Meldung zurückgibt.
3. Bei Abschluss des ursprünglichen Projektes 1998 umfasste die PND 1.863.984 Datensätze. (Scheven und Ohde, 2019)
4. Projektseite im BSZ-Wiki: .
5. Die Sächsische Landes- und Universitätsbibliothek (SLUB) in Dresden zeichnete hier verantwortlich für die Neuansetzung und Bereinigung der Werkdatensätze in der GND.
6. So fand vom 9. bis 11. Mai 2019 in der Akademie der Wissenschaften und der Literatur in Mainz die Konferenz „Werke, Werktitel, Werknorm - Perspektiven der Einführung einer Werkebene bei RISM“ statt. (Köppl, 2019)
7. Auch ohne dies tauchen aus unterschiedlichen Ursachen immer wieder neue unerwünschte Dubletten in der GND auf.
8. Die Format-Spezifikationen finden sich unter .
9. Hauptsächlich Personen, aber auch Körperschaften, Bauwerke und Sachbegriffe/Schlagworte.
10. Vgl. GND-Redaktionsanleitung und die Hinweise zur Korrektur von Personendaten in der GND.
11. Siehe auch den Abschnitt Geschichte auf Wikipedia:Normdaten.
12. Reine Namensansetzungen, d.h. nicht individualisierte Personen (sogenannte Tn-Sätze), sind seit 2020 nicht mehr Bestandteil der GND (Hartmann, 2020). Nach der Bereinigung blieben weit über 5 Millionen individualisierte Personendatensätze in GND (Jahresbericht 2020).
13. Üblicherweise für jeden zugänglich, tabellarisch als PDF über Google Drive.
14. Aktuell beträgt die „voraussichtliche Bearbeitungsdauer“ etwa ein Jahr.
15. Eine Übersicht von neuangelegten und korrigierten GND-Einträgen zeigt die Wikipedia-Seite *Portal:Bibliothek, Information, Dokumentation/Normdaten/GND-Kooperation* .
16. Die Wikipedia-Vorlage NORMDATENCOUNT gibt eine Übersicht der in der deutschsprachigen Wikipedia verwendeten Normdaten.

Bibliographie

Bicher, Katrin und Barbara Wiermann. 2018. “Normdaten zu „Werken der Musik“ und ihr Potenzial für die digitale Musikwissenschaft” In *Bibliothek Forschung und Praxis* 42 (2018), 222-235. https://doi.org/10.1515/bfp-2018-0043 (zugegriffen: 1. Juli 2023).

Deutsche Nationalbibliothek. 2023. “ GND-Redaktionsanleitung . Version 2.5, Stand: 30. Januar 2023. ” https://wiki.dnb.de/download/attachments/90411323/Redaktionsanleitung.pdf(zugegriffen: 1. Juli 2023).

Deutsche Nationalbibliothek. 2020. “ GND-Webformular ” https://www.dnb.de/gndwebformular (zugegriffen: 1 . Juli 20 23).

Deutsche Nationalbibliothek. 2021. – Jahresbericht 2020: 49. URN: urn:nbn:de:101-2021051859 https://d-nb.info/1234429616/34(zugegriffen: 1. Juli 2023).

Hartmann, Sarah. 2020. “Abschaffung von Tn-Normdaten in der GND mit Auswirkung auf die Titeldaten der DNB” https://wiki.dnb.de/display/ILTIS/Abschaffung+von+Tn-Normdaten+in+der+GND+mit+Auswirkung+auf+die+Titeldaten+der+DNB (zugegriffen: 1. Juli 2023).

Kailus, Angela. 2022. “ GND and NFDI4Culture together for better data ” https://nfdi4culture.de/id/E3118 (zugegriffen: 1 . Juli 20 23).

Köppl, Chantal. 2019. “ Works, Work Titles, Work Authorities: Perspectives on Introducing a Work Level in RISM (Tagungsbericht) ” https://www.musikforschung.de/publikationen/berichte/tagungsberichte/2019/2209-2(zugegriffen: 1. Juli 2023).

Scheven, Esther und Maike Ohde. 2019. “Geschichte der GND”. URN: urn:nbn:de:101:1-201909030845147331858 2. https://prezi.com/p/i86nojr2q6rs/geschichte-der-gnd/ (zugegriffen: 1. Juli 2023).

Schindler, Mathias. 2011. “Der Datengarten - Kollaborative Pflege von Norm- und Metadaten” https://books.ub.uni-heidelberg.de/arthistoricum/reader/download/163/163-17-75515-1-10-20160919.pdf (zugegriffen: 1 . Juli 20 23).

Steeg, Fabian und Adrian Pohl. 2021. “ Abgleich & Anreicherung eigener Daten mit der GND. OpenRefine Reconciliation mit lobid-gnd ” http://slides.lobid.org/2021-kim-reconcile/ (zugegriffen: 1 . Juli 20 23).

Voss, Jakob, Mathias Schindler und Christian Thiele. 2011. “ Link server aggregation with BEACON ” In International Symposium for Information Science 2011 http://eprints.rclis.org/15407/ (zugegriffen: 1 . Juli 20 23).

PhiWiki: ein semantisches Wiki für die Digitalphilosophie

Bailly, Kolja
kolja.bailly@tib.eu
Technische Informationsbibliothek Hannover, Deutschland

Geiger, Jonathan D.
jonathan.geiger@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland
ORCID: 0000-0002-0452-7075

Podschwadek, Frodo
frodo.podschwadek@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland
ORCID: 0000-0003-1248-4228

Vater, Christian
christian.vater@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland
ORCID: 0000-0003-1367-8489

Digital Humanities und Philosophie

In der Gegenwart ist unbestritten, dass Digitalisierung als allumfassender Prozess gesehen werden muss, der unsere Gesellschaft, unsere Arbeitswelt und die Disziplinen der Wissenschaft mitunter disruptiv verändert. Hiervon sind auch unsere Techniken des „allmählichen Verfertigens von Büchern“ (Böhle et. al., 1997) betroffen, wie bereits seit den 1980er Jahren auch im deutschsprachigen Raum festgehalten und erprobt werden konnten (vgl. auch Matejovsky und Kittler, 1996). Eine zentrale Bildungs- und Forschungsressource der Philosophie sind hierbei Wörterbücher, deren letzte ‚Überarbeitungswelle‘ in einer erst rudimentär digitalen und größtenteils unvernetzten Welt stattfand. Inhaltlich zeigt dies der Lemmatabestand in Auswahl und Zuschnitt, technisch die Art der Publikation. In der Gegenwart gibt es also die Bedarfe, (a) Lemmata der Digitalität zu ergänzen und/oder bestehende Lemmatalisten zu erweitern und (b) die Möglichkeiten zeitgemäßer technischer Angebote hierbei zu nutzen. Dem Selbstverständnis der Philosophie als selbstreflexive, hermeneutische Disziplin ermöglicht (c) der Einsatz neuer Techniken hierbei auch eine neuerliche Reflexion über Technologie, insbesondere über das Wech-

selverhältnis von Denkprozess und technischer Vermittlung. Besondere Schwerpunkte liegen hierbei auf Fragen nach der „technologischen Bedingung“ der Philosophie, der Vielfältigkeit von alternativen Ontologien/Taxonomien als Struktur des Denkens und der Möglichkeit der Erzeugung des Neuen durch Verbindung und Rekombination des verdateten Bekannten. Hiermit ergeben sich auch interessante Perspektiven auf das Spannungsfeld automatisierbaren Rechnens und individuell zu bildender Urteilskraft. Einen kurzen Überblick über das Verhältnis von Philosophie und Digital Humanities liefert Heßbrüggen-Walter (2018) und zur Verortung der Philosophie als digitale Geisteswissenschaft im Spannungsfeld „digitale Philosophie“ und „Philosophie der Digitalität“ vgl. Gramelsberger (2023).

Mit der programmatischen Verwendung eines semantischen Wiki-Systems als digitaler Denk-, Diskurs- und Publikationsraum, der mit vernetzten Normdatensystemen verkoppelt ist, betritt die deutschsprachige Philosophie Neuland. Die technische Umsetzung und soziale Koordination eines solchen Unternehmens ist auch ein Realexperiment, das ohne die Öffnung und Einbindung einer breiten Community gar nicht denkbar wäre. Hypertextuelle oder fragmentarisch-organisierte Texträume sowie strukturierte Wissensordnungen sind der Philosophie in ihrer Geschichte nicht fremd (Beispiele sind die Universalcharakteristik Leibnizens oder die Arbeit der französischen Enzyklopädisten Diderot und d’Alembert), doch eine enzyklopädische Wissenssammlung nicht (nur) als Text, sondern (auch) als Forschungsdaten zu betrachten, und in eine Norm- und Forschungsdateninfrastruktur operational einzubinden ist durch seinen genuin digitalen und vernetzten Charakter ein Novum.

Kontext und Hintergrund des semantischen Wikis

Die Arbeitsgruppe „Philosophie der Digitalität / philosophische Digitalitätsforschung “ wurde 2020 in der Deutschen Gesellschaft für Philosophie (DGPhil) gegründet (siehe https://digitale-philosophie.de/) und 2022 um differenzierte thematische Fokusgruppen erweitert. Diese Arbeitsgruppe entstand aus dem Bedürfnis der Community heraus, die verschiedenen Akteur:innen, die sich mit digitalen Phänomenen aus dezidiert philosophischer Perspektive beschäftigen, zusammenzuführen und einen gemeinsamen Diskursraum zu schaffen. So prägen das Programm der AG Themen wie digitale Ethik, Künstliche Intelligenz, epistemologische Reflexionen der digitalen Geisteswissenschaften oder Virtualität. Doch auch praktische Aspekte des digitalen Philosophierens sowie der digitalen Philosophiedidaktik werden thematisiert. Durchaus wird dabei auch die Perspektive digitaler Infrastrukturen für die Philosophie eingenommen und diskutiert. Eines der Ziele der AG ist daher auch die Verfertigung eines Glossars zur digitalen Philosophie, das auch die Publikationen eines wissenschaftlichen Netzwerks (in Beantragung bei der DFG) unterstützen soll. Dieses Glossar (das „PhiWiki “) dient hierbei gleichzeitig auch als Labor zur Erprobung zeitgemäßer digitaler Arbeitstechniken mit Fokus auf gemeinsame Textproduktion und Einbindung und Vernetzung bestehender Datenbestände von hoher Qualität. Auf Grundlage der Bedarfe der Fokusgruppe „Begriffs- und Wissensgeschichte“ und in enger Absprache mit dem Team „Forschungsdatenmanagement & Tools“ (beide Teil der AG) wurde an der Akademie der Wissenschaften und der Literatur Mainz eine Wikibase-Instanz mit semantischer Erweiterung aufgesetzt (Semantic-Media-Wiki) und modular weiterentwickelt. Eine enge Kooperation im Hinblick auf Data-Science-Methoden besteht hierbei mit dem Open-Science-Labor der Technischen Informationsbibliothek (TIB) Hannover. Die AG kooperiert eng mit dem Fachinformationsdienst Philosophie an der Universität Köln. Das PhiWiki ist dabei auch als Publikationsorgan gedacht, das nicht nur der internen Wissenssammlung dienen soll, sondern nach außen hin einsehbar und zitierbar ist und einen niedrigschwelligen Zugang zur Mitarbeit anbietet. Aktuell besteht kein explizites Funding des Projekts, es handelt sich noch um ein pet project im early stage development – an der Antragstellung wird allerdings gearbeitet.

Technische Beschreibung des Wikis

Auf Grundlage von MediaWiki 1.35 und des Wikibase-Pakets dev-REL1_35 wurde eine Instanz an der Akademie der Wissenschaften und der Literatur Mainz aufgesetzt. Diese Instanz steht als Docker-Image im RLPGitLab bereit. Sie wurde erweitert um die Semantic-Wiki-Extension 3.2, mit der nicht nur semantische Datenoperationen, sondern auch fortgeschrittene Visualisierungsfunktionen zur Verfügung stehen; weitere Extensions zur Implementierung sind vorgesehen (SemanticDrilldown, SemanticResultFormats, Maps_formerly Semantic Maps). Der Wikibase-Stack wurde ergänzt um einen gesonderten Annotation Service, der an der TIB entwickelt wurde. Mit Hilfe dieses Annotation Service als operativem Layer können vorhandene Datenbestände eingebunden werden, insbesondere authority control (VIAF, GND), Taxonomien – z. B. die auf der Stanford Encyclopedia of Philosophy (SEP, siehe https://plato.stanford.edu/) basierende InPhO (siehe https://www.inphoproject.org/) – und Literaturangaben (z. B. von der DNB) (vgl. dazu auch Menzel et al., 2021). Die gemeinsame Textarbeit wird durch die Wikioberfläche ermöglicht, die nicht nur eine zuverlässige kommentierbare Versionierung, sondern auch einen leistungsfähigen Kategorienbaum bereitstellt, über den nicht nur Such- und Sortierfunktionen laufen, sondern der auch Inhalte und Normdaten/Vokabulare verbindet. Zudem sind – wie in jedem MediaWiki – leistungsstarke Redaktionswerkzeuge vorhanden. Um die Nachhaltigkeit der gemeinsamen Arbeit zu sichern, wird das Wiki als Forschungsdatensatz begriffen, dessen Archivierung der Verstetigung dient. Die generische SPARQL-Schnittstelle von Wikibase dient der Einbindung in die entstehende „Wiki-Förderation “ .

Als Tool für die hypertextuelle Begriffsarbeit richtet sich das PhiWiki also an die philosophische Fachcommunity in mehreren Hinsichten, die auch die konkreten Zielgruppen konstituieren. Für Autor:innen von (primär deutschen, andere Sprache sind wünschenswert) Artikeln ist die Schaffung von redaktionellen Workflows in einer gewohnten Wiki-Umgebung angedacht, sodass neue Artikel verfasst oder bestehende Artikel angepasst oder erweitert werden können. Durch die Versionskontrolle und Änderungshistorie werden unterschiedliche Autorschaftsanteile dokumentiert und damit transparent und nachvollziehbar. Durch Kommentierungsfunktionen wird Peer-Review möglich, um die wissenschaftliche Qualität der Artikel zu gewährleisten. Forscher:innen stehen nicht nur die Artikel zur Verfügung, sondern ebenfalls eine Bandbreite technischer Funktionalitäten (insb. hinsichtlich der Recherche). Abschließend adressiert das PhiWiki auch Data Scientists und -Manager der Community, da Normdaten und Ontologien sowie inference rules für das semantic reasoning stets erweitert und angepasst werden müssen, was sowohl technische als auch fachwissenschaftliche Kenntnisse erfordert.

Weiterentwicklung und Prospektion

Es ist naheliegend, sich auf die domänen-spezifische (oben erwähnte) InPhO zu stützen, um ein bereits vorhandenes Vokabular nachzunutzen. InPhO wurde an der Indiana University unter der Leitung von Colin Allen entwickelt und vom National Endowment of the Humanities Fund und der Deutschen Forschungsgemeinschaft gefördert (vgl. auch Buckner et al., 2011 sowie Eckert et al., 2010). Diese Ontologie ist ein aufbereitetes Datenaggregat aus der Stanford Encyclopedia of Philosophy, steht unter einer CC BY-NC-SA-3.0-Lizenz und kann per Datendump oder per REST API verknüpft werden. Allerdings ist die InPhO für den Einsatz von s emantic r easonern nicht tiefenstrukturiert genug und weist nur eine geringe Anzahl semantischer Relationen auf. Zudem sind die downloadbaren Datendumps seit 2020 leer. Aus einer Korrespondenz mit Allen von 2020 ging hervor, dass die InPhO nicht weiterentwickelt wird und sowohl Finanzierungsmöglichkeiten als auch Personal mit der für diese Aufgabe notwendigen Kombination technischen und fachwissenschaftlichen Know-hows fehlen. Für das PhiWiki sollen die Begriffssets der InPhO zwar berücksichtigt, aber nicht als semantische Trägerstruktur des Wikis insgesamt verwendet werden. Ein zweites Ziel des PhiWiki ist also die Entwicklung eines domänen-spezifischen Vokabulars mit semantischen Relationen (Ontologie). Dieses konzentriert sich in erster Linie auf Namen (Personen, Orte, Werke, Events etc.) sowie die Relationen zwischen Personen, Institutionen und Texten. Hierfür sollen so weit wie möglich bestehende Datenstrukturen nachgenutzt werden, um ein (auch automatisiertes) Matching zu vereinfachen: das CIDOC Conceptual Reference Model (CRM) für Text- und Zitationsbeschreibungen, ICONCLASS für bildwissenschaftliche Auszeichnungen, das SCHEMA für Personennetzwerke und der Wikidata-Bestand für offene Fälle. Führende ID für jedes Datenobjekt ist hierbei ihr Eintrag in der Gemeinsamen Normdatei (GND) der Deutschen Nationalbibliothek (DNB) (so vorhanden). Die begriffliche Struktur des Wikis spiegelt den Aufbau des Historischen Wörterbuchs der Philosophie in seiner digitalen Bearbeitung durch Margarita Kranz. Durch diese Auswahl wird die Grundfunktionalität des PhiWikis als kollaborativer, digitaler und vernetzter „Zettelkasten“ gesichert. Aufgrund von Bedarfserhebung in der Community wird das PhiWiki schrittweise erweitert werden. Hierzu gehören (1) ein „Laborbuch“ für Marginalien und Miscellanea, (2) eine Genealogie der Besetzung philosophischer Lehrstühle und von Promotionslinien und (3) eine Kartierungsfunktion zur Geovisualisierung, z. B. für Drittmittelprojekte oder Forschungsverbünde.

Weitere Bestrebungen zur Vertiefung stellen Vernetzungen mit anderen Akteur:innen dar. Ein Beispiel für eine moderne Arbeitsumgebung für lexikographische und begriffsorientierte Arbeit ist die ORGANON terminology toolbox (siehe https://gkorganon.userpage.fu-berlin.de/). ORGANON basiert auf dem Open Encyclopedia System (OES), welches ebenfalls ein differenziertes technisches Konzept zur Versionierung und Referenzierung von Text bereithält. Ein technischer wie auch theorieorientierter Austausch mit dem Team hinter ORGANON insbesondere mit Werner Kogge (HU Berlin) besteht bereits (vgl. auch Kogge, 2021). Weitere Vernetzungen sind wünschenswert.

Auch wissenschaftstheoretische und epistemologische Reflexionen sollen die Arbeit mit und am PhiWiki begleiten. Hierbei ist die philosophische und medientheoretische Community eingeladen, mitzudenken an der Deutung des Wikis als Denkraum, sowohl was die Hypertextualität generell (vgl. dazu Bolter, 2001 und Coy, 1997), als auch speziell in den Digital Humanities bzw. der Philosophie als digitale Geisteswissenschaft betrifft (vgl. dazu Reichert, 2017; Drucker, 2021 und Schöch, 2017). Eine besondere Chance des PhiWikis ist es zudem, nicht nur bekannte Inhalte in das neue Medium zu übersetzen und zu transkodieren und dabei zu erproben, wie stark die Bedingungen der Re-Mediation unser Denken am Übergang von der Gutenberg- in die Turinggalaxis technisch vorprägen, sondern auch, die verborgenen Aushandlungsprozesse und Entstehungsbedingungen sichtbar zu machen, die mit der Verfertigung einer „enzyklopädischen Struktur“ verbunden sind – als Explikation der technischen Bedingung, aber auch als Ausleuchtung „verborgener Lehren“. Eine besondere Herausforderung ist hierbei stets die besondere Eigenschaft der Philosophie als Disziplin, gleichermassen stets neue zeitbezogene Denksysteme und Bezugsstrukturen zu entwickeln wie diese in Frage zu stellen.

Prospektion und Herausforderungen

Eine der Herausforderungen der aktuellen Phase der Entwicklungsarbeit ist die Erzeugung eines geeigneten Kategorienbaumes (siehe oben). Hierbei kann die technische „Unbegrenztheit“ genutzt werden, um alternierende Kate-

goriensysteme als auszuwählende Alternative einzubinden (beispielsweise der redaktionell kuratierten Standardwerke „Historisches Wörterbuch der Philosophie“ (vgl. Ritter, 1971), „Enzyklopädie Philosophie und Wissenschaftstheorie“ oder der Stanford Encyclopedia of Philosophy sowie WikiData als ‚crowdsourced commons ‘). Diese Offenheit ist dabei allerdings gleichzeitig auch eine Herausforderung. Offenheit verstanden als Textoffenheit markiert den Charakter der Wissenssammlung als nicht abgeschlossen oder prinzipiell nicht abschließbar. Hierfür sind also Techniken und Strategien notwendig, um ein gewisses Level wissenschaftlicher Qualität zu gewährleisten und praktischen Anforderungen wie der Zitierbarkeit Rechnung zu tragen.

Technisch stellen sich Herausforderungen im Bereich der „Dockerisierung “ des gesamten Stacks inklusive der versionsabhängigen Interaktion der verwendeten Komponenten. Die Einbindung der Ergebnisse via SPARQL-Endpoint in die entstehende „Förderation der Wikis“ stellt sich vor allem als soziale bzw. kommunikative Aufgabe; ebenso wie die Frage nach der Langzeitverfügbarkeit (was auch Hosting- und Kuratierungsaufgaben umfasst) des PhiWikis (aktuell ist die Staging-Instanz auf den Servern der Akademie der Wissenschaften und der Literatur Mainz gehostet).

Ein weiterer Punkt ist die Konzeption des Redaktionsprozesses. Kollektive Wissenssammlungen benötigen Workflows, um Zugangs- und Partizipationsmöglichkeiten zu schaffen sowie Mechanismen des Qualitätsmanagements. Die Ausbalancierung von möglichst offenen Bearbeitungsrechten der Wikiseiten (Artikelseiten, Diskussionsseiten etc.) und der Gewährleistung eines wissenschaftlichen Qualitätsstandards ist eine wesentliche Herausforderung solcher Systeme. Notwendig wird die Organisation einer redaktionellen Community sein – hier ist zentral an die DGPhil AG „Philosophie der Digitalität / philosophische Digitalitätsforschung “ gedacht. Die AG umfasst aktuell ca. 200 Mitglieder, die sich nicht nur durch philosophische Expertise, sondern auch durch eine gewisse (digital)technische Affinität auszeichnen. Zudem stellt sie eine der wesentlichen Zielgruppen des PhiWikis dar. Selbstverständlich ist die redaktionelle Community aber offen für alle Interessierte auch außerhalb dieser AG, auch unabhängig von akademischen Titeln. Denkbar sind regelmäßige Meetings dieser Community, um organisatorische, technische und inhaltliche Fragen gemeinschaftlich zu klären sowie (virtuelle) Workshops für Einsteiger:innen oder zur Entwicklungszielsetzung. In der Gruppe erarbeitete Leitfäden und ein Rollenkonzept flankieren die Redaktion.

Vortragsformat und Diskussionspunkte

In dem Vortrag wird das semantische Wiki der Digitalphilosophie „PhiWiki “ vorgestellt. Dabei wird auf die Vorgeschichte, die Bedarfe der Community und den Kontext in der DGPhil Arbeitsgruppe „Philosophie der Digitalität / philosophische Digitalitätsforschung “ und der Akademie der Wissenschaften und der Literatur Mainz eingegangen (vgl. auch Kranz, 2012). Der aktuelle Stand des Wikis, ebenso wie die Konzeptualisierung der neu zu entwickelnden Ontologie (sowie Anschlussmöglichkeiten an bereits bestehende Vokabulare und Normdatensätze) werden präsentiert und die technischen Spezifikationen erläutert und (soweit möglich) demonstriert. Zudem werden künftige Herausforderungen skizziert und in der Diskussion das Feedback aus den relevanten Fachcommunities eingeholt (Philosophie, Wikis, Data Science).

Obwohl im Anschluss nur eine begrenzte Zeit für die Diskussion zur Verfügung steht, sollen dennoch drei wesentliche Aspekte bzw. Herausforderungen des PhiWikis im Plenum diskutiert werden.

1. Wie lassen sich derartige dezentrale, hypertextuelle Texträume für eine geisteswissenschaftliche Community gut organisieren? (Gibt es Erfahrungswerte hinsichtlich des Rollenmanagements, des Betriebs oder der Gewährleistung der wissenschaftlichen Qualität der Artikel?)
2. Welche Möglichkeiten bzw. Notwendigkeiten zur inhaltlichen, technischen und organisatorischen Verschaltung mit anderen Werken bzw. Projekten gibt es? (Wie lässt sich Interoperabilität zu anderen lexikographischen oder philosophischen Werken herstellen?)
3. Für den Einsatz welcher digitaler Methoden eignet sich das PhiWiki als Semantic-Media-Wiki? (Gibt es potenziell gewinnbringende Methoden aus dem Bereich der Data Science, des Informationsdesigns oder der Korpuslinguistik, die bereits bei der Konzeptionierung des Datenmodells mitgedacht werden sollen?)

Eine Verortung der Philosophie im Diorama der Digital Humanities wird hierbei ebenfalls versucht und technische, infrastrukturelle und organisatorische Anschlussmöglichkeiten als Ergebnisse einer ersten Heuristik aufgezeigt. Das PhiWiki als virtuelle Forschungsumgebung stellt dabei einen (möglichen) neuartigen Weg für die philosophische Begriffsarbeit einer in die Zukunft verlängerbaren Gegenwart dar.

Bibliographie

Böhle, Knud, Ulrich Riehm, Bernd Wingert. 1997. *Vom allmählichen Verfertigen elektronischer Bücher. Ein Erfahrungsbericht.* In *Veröffentlichungen des Instituts für Technikfolgenabschätzung und Systemanalyse 5*. Frankfurt am Main: Campus.

Bolter, Jay D. 2001. *Writing Space. Computers, Hypertext, and the Remediation of Print*. Hillsdale (NJ): Erlbaum. (=erw. Neuauflage von 1991)

Buckner, Cameron, Mathias Niepert, Colin Allen. 2011. „From encyclopedia to ontology: toward dynamic representation of the discipline of philosophy.“In *Synthese* 182: 2. 205–233.

Coy, Wolfgang. 1997. „turing@galaxis II“. In *HyperKult. Geschichte, Theorie und Kontext digitaler Medien*, hg. von Martin Warnke, Wolfgang Coy und Georg Christoph Tholen, 15–33. Basel u. Frankfurt am Main: Stroemfeld.

Drucker, Johanna. 2021. *The Digital Humanities coursebook. An Introducition to Digital Methods for Research and Scholarship*.New York: Routledge.

Eckert, Kai, Mathias Niepert, Christof Niemann, Cameron Buckner, Colin Allen, Heiner Stuckenschmidt. 2010. „Crowdsourcing the assembly of concept hierarchies.“In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10). Association for Computing Machinery*. New York. 139–148. DOI: https://doi.org/10.1145/1816123.1816143.

Gramelsberger, Gabriele . 2023. *Philosophie des Digitalen. Eine Einführung*. Hamburg: Junius.

Heßbrüggen-Walter, Stefan . 2018. „Philosophie als digitale Geisteswissenschaft.“ In *Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden, hg. von Martin Huber und Sybille Krämer* (=Sonderband der Zeitschrift für digitale Geisteswissenschaften, 3). DOI: https://doi.org/10.17175/sb003_006.

Kogge, Werner . 2021. „Begriffsforschung im Interdisziplinären Kontext. Neuansätze einer Methode. Erster Teil.“ In *Archiv für Begriffsgeschichte* 63: 1. 105–134.

Kranz, Margarita . 2012. „Begriffsgeschichte institutionell - Teil II. Die Kommission für Philosophie der Akademie der Wissenschaften und der Literatur Mainz unter den Vorsitzenden Erich Rothacker und Hans Blumenberg (1949–1974).“ In *Archiv für Begriffsgeschichte* 54, 119–194.

Matejovsky, Dirk, Friedrich Adolf Kittler (Hrsg.) **.** 1996. „Literatur im Informationszeitalter.“ In *Schriftenreihe des Wissenschaftszentrums Nordrhein-Westfalen 2*. Frankfurt am Main: Campus.

Menzel, Sina, Hannes Schnaitter, Josefine Zinck, Vivien Petras, Clemens Neudecker, Kai Labusch, Elena Leitner und Georg Rehm . 2021. „Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten.“ In *Qualität in der Inhaltserschließung, hg. von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl und Hans Schürmann*. Berlin, Boston: De Gruyter Saur. 229–258. DOI: https://doi.org/10.1515/9783110691597-012.

Reichert, Ramón . 2017. „Theorie Digitaler Medien.“ In *Digital Humanities. Eine Einführung, hg. von Fotis Jannidis, Hubert Kohle, Malte Rehbein*. Stuttgart: J.B. Metzler. 19–34.

Ritter, Joachim . 1971. „Vorwort“. In *Historisches W ö rterbuch der Philosophie, Band 1 (A – C)*. Basel: Schwabe.

Schöch, Christo f . 2017. „Digitale Wissensproduktion.“ In *Digital Humanities. Eine Einführung, hg. von Fotis Jannidis, Hubert Kohle, Malte Rehbein*. Stuttgart: J.B. Metzler. 206–212.

Project Overhaul und Refactoring der digitalen Edition der ‘Urfehdebücher der Stadt Basel’ mithilfe von GPT-4 und LLM

Pollin, Christopher
christopher.pollin@uni-graz.at
Institut Zentrum für Informationsmodellierung,
Universität Graz, Österreich
ORCID: 0000-0002-4879-129X

Scholger, Martina
martina.scholger@uni-graz.at
Institut Zentrum für Informationsmodellierung,
Universität Graz, Österreich
ORCID: 0000-0003-1438-3236

Steiner, Elisabeth
elisabeth.steiner@uni-graz.at
Institut Zentrum für Informationsmodellierung,
Universität Graz, Österreich
ORCID: 0000-0001-9116-0402

Lang, Sarah
sarah.lang@uni-graz.at
Institut Zentrum für Informationsmodellierung,
Universität Graz, Österreich

Galka, Selina
selina.galka@uni-graz.at
Institut Zentrum für Informationsmodellierung,
Universität Graz, Österreich
ORCID: 0000-0003-4476-315X

Schiller-Stoff, Sebastian
sebastian.stoff@uni-graz.at
Institut Zentrum für Informationsmodellierung,
Universität Graz, Österreich
ORCID: 0000-0001-6941-113X

Einleitung und Begriffsdefinitionen

Ziel dieses Beitrags ist die Erprobung des Einsatzes von Large Language Models (LLMs) wie GPT-4 *(OpenAI*

2023) für das Overhauling und Refactoring von Projekten in den Digital Humanities am Beispiel der digitalen Edition “Urfehdebücher der Stadt Basel” (UFBAS; https://gams.uni-graz.at/ufbas). Im Rahmen dieses Refactoring- und Overhauling-Prozesses, der sich auf die Überarbeitung von Daten, Datenmodell, Website und Code bezieht, werden LLMs für die verschiedenen Arbeitsschritte eingesetzt, getestet und evaluiert. Ziel ist es, zu evaluieren, inwieweit LLMs in der Lage sind, Arbeitsschritte KI-assistiert abzuarbeiten, und ob ihre Outputs den Standards der Digital Humanities entsprechen und in der Praxis anwendbar sind.

LLMs sind fortgeschrittene KI-Systeme, die auf die Erzeugung und das Verarbeiten menschlicher Sprache spezialisiert sind. Sie “lernen” aus großen Textmengen und sind in der Lage, kontextbezogene, qualitativ hochwertige Texte zu generieren. Ein prominentes Beispiel ist GPT-4 (Generative Pre-training Transformer 4), das aufgrund seiner Vielseitigkeit und Leistungsfähigkeit in verschiedenen Bereichen eingesetzt werden kann, darunter Datenmodellierung, -transformation, -visualisierung *(Chen et al. 2023)* und -programmierung *(Poldrack, Lu, und Beguš 2023; Tian et al. 2023)*. Insbesondere in Kombination mit dem ChatGPT Plugin Code Interpreter erweist sich GPT-4 als eines der leistungsfähigsten Werkzeuge in diesem Bereich *(AI Explained 2023a; AI Explained 2023b)*. Die Anwendungsmöglichkeiten von LLMs erstrecken sich auch auf Bereiche wie Datengenerierung, Prototyping und Ontology Engineering bzw. Datenmodellierung, wie *(Pollin 2023)* zeigen konnte. Das größte Potenzial kann entfaltet werden, wenn sehr gute Kenntnisse im Prompt Engineering mit hohem Domänenwissen kombiniert werden und das “Human in the Loop”-Paradigma verfolgt wird. Mit anderen Worten: Es findet eine kontinuierliche Bewertung des Outputs eines LLM durch Expert*innen aus der Domäne statt und die KI wird assistierend in den Arbeitsprozess integriert. Die Custom GPTs von OpenAI (https://openai.com/blog/introducing-gpts) ermöglichen es, solche Prozesse mit einer eigenen Wissensbasis und System Prompts (Instruction) zu optimieren.

In der Softwareentwicklung bezeichnet der Begriff **Overhaul** eine tiefgreifende Umstrukturierung von Systemen mit dem Ziel, deren Leistungsfähigkeit zu verbessern oder sie an die neuesten Standards und Anforderungen anzupassen. Dies beinhaltet häufig erhebliche Änderungen der Architektur, der Funktionalität und des Designs von Websites und Webschnittstellen. Ein Redesign wird dann notwendig, wenn ein Projekt so veraltet ist, dass es nicht mehr den aktuellen Anforderungen an Wartung, Sicherheit und Benutzerfreundlichkeit entspricht. Ein solcher Prozess kann im Rahmen eines Refactorings stattfinden. **Refactoring** ist ein Prozess in der Softwareentwicklung, bei dem der Quellcode überarbeitet wird, um seine Struktur zu verbessern, ohne das äußere Verhalten der Software zu verändern. Ziel ist es, den Code übersichtlicher, wartbarer und flexibler zu machen. Häufig wird ein Refactoring im Rahmen einer Neugestaltung einer Webpräsenz durchgeführt, insbesondere um die technischen Aspekte des Systems zu optimieren.

Ziel dieser Studie ist es, das Unterstützungspotential von LLMs wie GPT-4 im Bereich des Code Overhauling und Refactoring zu untersuchen. Dabei soll nicht nur der Prozess der Codeoptimierung beleuchtet werden, sondern auch die Möglichkeiten und Grenzen von LLMs in diesem Kontext aufgezeigt werden. Es soll analysiert werden, inwieweit solche Modelle und Methoden als Lösungen für aktuelle Herausforderungen in den Digital Humanities dienen können.

Fallstudie UFBAS

Das Projekt UFBAS *(Burghartz, Calvi und Vogeler 2017; Pollin und Vogeler 2017)* unter der Leitung von Susanna Burghartz umfasst die digitale Edition des Urfehdebuchs der Stadt Basel für die Jahre 1563 bis 1569, die trotz ihres inhaltlichen Werts in ihrer technischen Umsetzung Verbesserungspotential aufweist und dem heutigen Stand der Technik angepasst werden muss. UFBAS wurde 2017 im Rahmen eines einjährigen Projekts am Zentrum für Informationsmodellierung (ZIM) durch einen studentischen Mitarbeiter mit begrenzten personellen Ressourcen umgesetzt und im **Geisteswissenschaftlichen Asset Management System** *(GAMS; Stigler und Steiner 2018)* veröffentlicht.

Die verschiedenen Revisionsschritte sollen mit Hilfe von GPT-4, den vorhandenen Plugins und insbesondere dem Plugin Code Interpreter, sowie einem projektspezifischen Custom GPT unterstützt werden. Das Vorgehen orientiert sich dabei an den Experimenten, die in der Workshopreihe “Angewandte Generative KI in den (digitalen) Geisteswissenschaften” (https://chpollin.github.io/GM-DH ; *Pollin 2023*) erprobt wurden. Die Ergebnisse der GPT-4 assistierten Prozesse werden auf https://github.com/chpollin/Overhaul-UFBAS veröffentlicht.

Angestrebte Verbesserungen

Die in diesem Projekt geplanten Verbesserungen umfassen eine Vielzahl von Aspekten: Die Aktualisierung und Erweiterung der Datenbasis und der Metadaten, die Neugestaltung des Datenmodells durch eine Ontologie, die auf der Top-Level-Ontologie CIDOC-CRM basiert, und die Entwicklung eines TEI XML ODD. Auch das Refactoring des JavaScript- und XSLT-Codes, der für die RDF-Generierung und die Webentwicklung verwendet wird, ist ebenfalls geplant. Ein weiterer Fokus liegt auf der Verbesserung der HTML5-Validierung, der Barrierefreiheit und der Responsivität der Webpräsenz. Darüber hinaus ist die Erstellung prototypischer Datenvisualisierungen geplant, wobei Design und Implementierung durch das ChatGPT-Plugin „Code Interpreter“ wesentlich unterstützt werden können. Diese Auflistung umfasst einige der geplanten Arbeitsschritte, erhebt aber keinen Anspruch auf Vollständigkeit.

TEI XML

Die TEI XML Version vom 31.1.2017 (https://gams.uni-graz.at/o:ufbas.1563) wird grundlegend überarbeitet. Einige dieser Änderungen können als trivial angesehen werden und erfordern kein LLM. Der eigentliche Nutzen liegt jedoch in der unterstützten bzw. automatisierten Generierung eines Scripts durch GPT-4, das basierend auf dem Projektkontext, den Problemen und Beispielen im Prompt bzw. in der Conversation bereitgestellt wird. Dies minimiert manuelle Eingriffe und verbessert die Effizienz und Genauigkeit des gesamten Prozesses. Dies beinhaltet die folgenden beispielhaften Aufgaben:

- Anpassung der @ana-Referenzen zur semantischen Anreicherung des edierten Textes: @ana="#uf_Eintrag" zu @ana=”uf:entry”.
- Überarbeitung des teiHeaders in Anlehnung an das tei-Header Template, das in den letzten Jahren am ZIM als Standard festgelegt wurde.
- Vereinheitlichung im Rahmen der am ZIM erarbeiteten Coding Conventions, z.B. bei der Vergabe von @xml:id.
- Normalisierungen und Verlinkungen von Entitäten, wie beispielsweise Taten oder Strafen.
- Erstellung eines ODD Files auf Basis des TEI XML.

Die GPT-Assistenz umfasst die Analyse und die Fehlersuche im TEI XML, sowie das darauf aufbauende Programmieren der Datentransformation mittels XSLT oder Python.

Kategorien & SKOS

Die im TEI XML referenzierten Kategorien sind derzeit in einem eigenständigen TEI XML Dokument (https://gams.uni-graz.at/o:ufbas.kategorien) abgebildet. Dieses Dokument wird nach SKOS überführt und die Referenzierung der Kategorien im TEI XML auf das SKOS angepasst.

Die Transformation von TEI XML nach SKOS erfolgt mit dem GPT-4 Code Interpreter. Dabei werden die Ein- und Ausgabedaten mittels “Few-Shot Prompting” übergeben und mit diesem " *In-Context-Learning*" ein Python-Skript für die Transformation generiert.

Ontologie

Im Rahmen von UFBAS wurde kein RDFS- oder OWL-File zur formalen Beschreibung des Datenmodells für RDF-Daten realisiert. Im Overhaul wird dafür eine domänenspezifische Ontologie entwickelt, die die assertive Schicht *(Vogeler 2019)* eines Eintrags in einem Urfehdebuch beschreibt. Diese Ontologie beschreibt Tat, Täter und Strafe sowie weitere Phänomene, die in den RDF-Daten repräsentiert werden. Um das Alignment zu gewährleisten, ist eine CIDOC-CRM Ableitung der domänenspezifischen Ontologie enthalten.

GPT-4 wird dafür im Prozess des Ontology Engineerings eingesetzt.

RDF

Beim Ingest-Prozess in GAMS wird RDF aus TEI XML mithilfe von XSLT extrahiert. Dieses RDF (https://gams.uni-graz.at/o:ufbas.1563/RDF) wird an die domänenspezifische Ontologie angepasst, ebenso werden weitere Verbesserungen am RDF XML vorgenommen. Es wird getestet, welche weiteren Normalisierungen und semantischen Anreicherungen der Daten mit Wikidata möglich sind, um die FAIRness *(Wilkinson et al. 2016)* der Daten signifikant zu verbessern. Beispielsweise Anpassungen umfassen:

- Ergänzungen fehlender rdf:type.
- Vereinheitlichung von Klassen und Properties nach der Ontologie: uf:Beruf zu uf:Occupation
- Vereinheitlichung und Integration zwischenzeitlich entwickelter GAMS-spezifischer Ontologien: g2o:inhalt zu gams:textualContent.
- Metadaten zum RDF Datensatz mittels VoID Vocabulary (https://www.w3.org/TR/void/).
- Verwendung anderer bestehender Vokabulare wie schema.org, wo möglich.

GPT-4 wird im Refactoring der Transformation nach RDF eingesetzt.

Webentwicklung

Ein wichtiger Aspekt unserer Verbesserungspläne ist die Überarbeitung der Webrepräsentation. Dies umfasst nicht nur die Sicherstellung von Responsivität, Accessibility und HTML5-Validität, sondern auch das Upgrade auf Bootstrap 5 und andere JavaScript-Bibliotheken, sowie die Überarbeitung des Designs. Dabei werden auch die Such- und Explorations-Funktionalitäten der Website überarbeitet und ggf. erweitert. Insgesamt wird dadurch die FAIRness der Daten verbessert.

GPT-4 wird für die Problemanalyse, das Refactoring und die Neuentwicklungen eingesetzt.

Datenvisualisierung

Schließlich werden die Visualisierungen vollständig überarbeitet. In der UFBAS-Version vom 31.1.2017 wurde eine Netzwerkvisualisierung implementiert, die ein Suchergebnis nach einer Kategorie darstellt. Dabei wurden Täter:innen, ihre Berufe, verknüpfte Orte, Tat und Strafe in einem Graphen abgebildet. Die Aussagekraft dieser Visualisierung sowie die Handhabung sind nicht ideal, da sie eher als Experiment gegen Ende des Projekts umgesetzt wurden (https://tinyurl.com/4nejhksf).

GPT-4 wird während des gesamten Prozesses von der Konzeption über das Prototyping bis hin zur Implementierung eingesetzt.

Fazit

Die Integration von KI, insbesondere von GPT-4, hat das Potenzial, nicht nur das UFBAS-Projekt auf eine neue Ebene zu heben, sondern auch wertvolle Erkenntnisse und Erfahrungen zu liefern, die zur Weiterentwicklung der Anwendung von KI in den digitalen Geisteswissenschaften beitragen können.
Ziel ist es, das Potenzial neuer Technologien wie LLM und generativer KI für die Überarbeitung, Optimierung, Ergänzung und Neugestaltung bestehender digitaler Editionsprojekte auszuloten. Die gewonnenen Erkenntnisse können auch für die Konzeption neuer Projekte genutzt werden. Die Vorgehensweise und der Aufwand werden daher ausführlich dokumentiert und erläutert, was gerade bei so schnelllebigen Technologien unerlässlich ist. Im Zentrum des Überarbeitungsprozesses steht das Custom GPT “ufbasGPT” (https://chat.openai.com/g/g-p8ZMcZK5r-ufbasgpt), das als Überarbeitungsassistent fungiert. Dies geschieht mittels Prompting und einer Wissensbasis aller notwendigen Projektdaten. Für die Reproduzierbarkeit von Forschung, auch im Sinne von Nachnutzbarkeit und Open Science, stellt diese Technologie eine neue Herausforderung dar. Die Dokumentation des GPT-4 Overhauls wird in einem GitHub Repository veröffentlicht (https://github.com/chpollin/Overhaul-UFBAS). Die Ergebnisse des Overhauls werden auf einem nicht veröffentlichten Entwicklungsserver laufend angepasst und bis zur DHd2024 als neue Version in GAMS veröffentlicht.

Bibliographie

AI Explained. 2023. 12 New Code Interpreter Uses (Image to 3D, Book Scans, Multiple Datasets, Error Analysis ...). https://www.youtube.com/watch?v=_njf22xx8BQ .

AI Explained. 2023. GPT 4 Got Upgraded - Code Interpreter (Ft. Image Editing, MP4s, 3D Plots, Data Analytics and More!). https://www.youtube.com/watch?v=O8GUH0_htRM .

Andrae, Magdalena, Susanne Blumesberger, Sonja Edler, Julia Ernst, Sarah Fiedler, Doris Haslinger, Gerhard Neustätter und Denise Trieb. 2020. “Barrierefreiheit für Repositorien. Ein Überblick über technische und rechtliche Voraussetzungen”, Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 73(2), S. 259–277. https://doi.org/10.31263/voebm.v73i2.3640 .

Burghartz, Susanna, Sonia Calvi und Georg Vogeler. 2017. “Urfehdebücher Der Stadt Basel – Digitale Edition”, digital edition, Urfehdebücher der Stadt Basel – digitale Edition. http://gams.uni-graz.at/context:ufbas .

Chen, Zhutian, Chenyang Zhang, Qianwen Wang, Jakob Troidl, Simon Warchol, Johanna Beyer, Nils Gehlenborg und Hanspeter Pfister. 2023. ‘Beyond Generating Code: Evaluating GPT on a Data Visualization Course’. https://doi.org/10.48550/arXiv.2306.02914 .

OpenAI. 2023. ‘GPT-4 Technical Report’. arXiv https://doi.org/10.48550/ARXIV.2303.08774 .

Poldrack, Russell A., Thomas Lu und Gašper Beguš, ‘AI-Assisted Coding: Experiments with GPT-4’ (arXiv, 25 April 2023). https://doi.org/10.48550/arXiv.2304.13187 .

Pollin, Christopher. 2023: Workshopreihe "Angewandte Generative KI in den (digitalen) Geisteswissenschaften" (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.10065626 .

Pollin, Christopher und Georg Vogeler. 2017: Semantically Enriched Historical Data. Drawing on the Example of the Digital Edition of the “Urfehdebucher der Stadt Basel”. WHiSe.

Stigler, Johannes und Elisabeth Steiner. 2018. ‘GAMS–Eine Infrastruktur Zur Langzeitarchivierung Und Publikation Geisteswissenschaftlicher Forschungsdaten’, Mitteilungen Der Vereinigung Österreichischer Bibliothekarinnen Und Bibliothekare 71, no. 1 (2018): 207–16.

Tian, Haoye, Weiqi Lu, Tsz On Li, Xunzhu Tang, Shing-Chi Cheung, Jacques Klein, Tegawendé F. Bissyandé. 2023. ‘Is ChatGPT the Ultimate Programming Assistant -- How Far Is It?’. https://doi.org/10.48550/arXiv.2304.11938 .

Vogeler, Georg. 2019. ‘The “Assertive Edition”’, International Journal of Digital Humanities 1, no. 2 (1 July 2019): 309–22. https://doi.org/10.1007/s42803-019-00025-5 .

Wilkinson, Mark D. 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 .

Quantifying trust towards LLM-based chatbots: A mixed-method approach

Belosevic, Milena
milena.belosevic@uni-bielefeld.de
Bielefeld University, German Linguistics/Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University, Germany

Buschmeier, Hendrik
hbuschme@uni-bielefeld.de
Bielefeld University, German Linguistics/Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University, Germany

1 Introduction

Trust plays a crucial role in human interaction and has therefore been approached from sociological, psychological, philosophical, and the perspective of communication studies (Hendriks et al., 2021). Given the current popularity of large language model/LLM-based chatbots that can interact in a human-like, conversational way (Rudolph and Samson, 2023), it can be stated that trust in automation is gaining increasing importance. Similar to social norms such as politeness (Lumer and Buschmeier, 2022) in human-machine interaction, trust in automation (Lee and See, 2004, Lukyanenko et al., 2022 for an overview) can also be regarded as a shared social value constructed in direct or indirect interaction. However, in linguistic research on trust (Schäfer, 2016, Belosevic 2022), this aspect has rarely been addressed (Schneider et al., 2022, Lotze, 2016, Kabir et al. 2023).

Since trust comprises both cognitive and emotional aspects (Kok and Soh, 2020), this paper focuses on the role of emotional aspects of trust in indirect interaction with chatbots and uses the perceived trustworthiness ascribed to ChatGPT (as one of the most recent LLM-based chatbots) as a testbed. Since little research has focused on how language shapes the evaluation of trustworthiness ascribed to machines in their role as trust objects, the paper aims to show how trust in human interaction with chatbots can be modeled using manual annotation and sentiment analysis. This stands in contrast to recent studies on the role of trust in ChatGPT, which are mainly based on experimentally elicited data and do not consider the role of language (e.g., Funke et al., 2023, Shen et al. 2023, Watters and Lemanski, 2023, Huang et al. 2023, Liu et al. 2023). In particular, we propose a mixed-method approach to quantify the trustworthiness ascribed to ChatGPT in an indirect interaction. In this interaction mode, the distinction between the first- and third-person perspective in the human-machine interaction (Coeckelbergh, 2011) is crucial. Whereas the first-person perspective is concerned with how we interact with chatbots, the third-person perspective is adopted in this paper. It explores how users talk about chatbots and how the perceived trustworthiness of ChatGPT is promoted through the discursive commodification of trust (cf. Krüger and Wilson, 2022) in the public debate about ChatGPT in Germany. To this end, we use qualitative approaches to trust (identification of trust-relevant vocabulary through manual annotation) and qualitative-quantitative methods (sentiment analysis) to account for the emotional aspects of trust in human-chatbot interaction.

Prior to applying these methods to our case study, it is necessary to define trust in automation and specify the properties of trust underlying the annotation scheme and their relation to sentiment values.

2 Methodology and data

Trust is a complex phenomenon that can be operationalized using other more concrete concepts (so-called trust cues or trust indicators) as a proxy. This is also true for trustworthiness (cf. Lewicki and Alister, 1998). To identify emotional aspects of trustworthiness, linguistic units that serve as indicators of perceived trustworthiness must first be detected. We consider manual annotation to be the first step toward modeling linguistic indicators of trustworthiness and narrowing down the complex concept of trust into more concrete aspects. The annotation task is based on an annotation scheme with several annotation categories defined by drawing on existing studies on trust in human-human interaction (cf. Kuhnhenn 2014) and human-robot interaction (cf. de Visser et al. 2020)[1] . The central annotation unit, namely the notion of perceived trustworthiness was adopted from the concept of trust calibration (Lee and See, 2004, Muir, 1994) which is an often applied framework in studies on trust in human-machine interaction (cf. Wischnewski et al. 2023 for an overview). For the purposes of the annotation scheme, it was defined as the perception of users' trust toward the system in contrast to the actual trustworthiness of the trust object.

Based on the results of the manual annotation, sentiment analysis was carried out to account for the role of emotional aspects of perceived trustworthiness in the public discussion about ChatGPT. We tested one machine-learning-based and one lexicon-based model for sentiment analysis of German texts: the model for sentiment classification available on Hugging Face 3 (pre-trained on 1.834 million German-language samples, mainly texts from Twitter, Facebook, movie, app, and hotel reviews, Guhr et al., 2020) and the Python package textblob (Loria, 2020) based on the German polarity lexicon. The German version of textblob can be used to obtain polarity ratings between -1 (negative) and +1 (positive) for words, sentences, and texts. Both the machine-learning-based model and textblob provide ratings based on sentences, phrases, and single words. Trust-related linguistic markers were also annotated manually with regard to the promotion of trust or distrust.

The data (27.138 tokens) were obtained from the DWDS-Webmonitor corpus[2] using the word *ChatGPT* and the period between 2022-11-30 (release of ChatGPT) and 2023-06-30. This dataset comprises some 6.198.349 texts (mostly web pages from German-speaking countries). However, only the intermediate sentence context comprising the search word is available for analysis.

3 Manual annotation and sentiment analysis

As mentioned above, the central part of manual annotation includes the development of an annotation scheme and the definition of annotation categories based on the definition of the main aspects of trust provided in the previous section. The annotation scheme consists of the following annotation categories: interaction mode, trust levels and sentiment values, trust roles in human interaction with ChatGPT, and perceived trustworthiness.

The data were annotated by one annotator using the software MAXQDA Plus[3] . Two types of annotation categories were combined: the annotation with indicators of trust on the word, multi-word, and sentence level as well as sentiment values. Since indicators of trust comprise several aspects (ability, benevolence, and integrity) the annotation of these aspects in terms of positive and negative sentiment values (trustworthiness vs. distrust) is closely related to aspect-based sentiment analysis that goes beyond the formal level of single words and sentences and focuses on properties of aspect categories (cf. Liu 2015: ch. 5 and 6). In the following, the annotation scheme will be described.

For the annotation category ‘interaction mode’, we annotated the following domains in which the perceived trustworthiness ascribed to ChatGPT is discussed: education, science, politics, sports, and industry. To determine the mode, the annotator often checked the whole text in which the example appeared.

Trust roles comprise the society and users in their roles as trustors on the one side and trust objects (here ChatGPT) on the other. Since our case study is concerned with the public debate about perceived trustworthiness ascribed to ChatGPT by the users (i.e., during the interaction), the central aspect of this annotation unit is not the user, but the discourse actors (e.g., journalists, experts) who indicate their perception of the perceived trustworthiness of users.

To identify linguistic cues of perceived trustworthiness, we draw on the linguistic markers of credibility and trustworthiness proposed by Kuhnhenn (2014) and Reinmuth (2006) for human-human interaction, the categories of trust in human-robot interaction (cf. de Visser et al. 2020), and on the three properties of trust (Mayer et al.,1995) widely accepted in the literature, namely, competence (defined as skills, and characteristics that enable the trustee to influence the domain), benevolence (specified as the extent to which the intents and motivations of the trustee are aligned with those of the trustor), and integrity (the degree to which the trustee adheres to a set of principles the trustor finds acceptable). Each linguistic marker was annotated with one of the indicators of trustworthiness (competence, benevolence, or integrity). Manual annotation is necessary as there is no agreement about which linguistic units can be regarded as trust-relevant. Moreover, for each domain in which the role of trust is investigated, trust-relevant aspects should be defined based on indicators underlying the construction of trust in the context.

The annotation category ‘trust levels/sentiment values’ is based on our hypothesis that emotional aspects of trust are related to positive emotions and vice versa so that sentiment values can be regarded as a potential cue of emotional dimensions of trustworthiness. Therefore, the levels of positive trustworthiness, negative trustworthiness (distrust), and ambiguous cases were considered sentiment values and served as annotation units for this category. Specifically, trustworthiness was annotated with 1, negative trustworthiness/distrust with -1, and in cases where there was no clear distinction regarding the sentiment value ‘both/ambiguous' was annotated with 0. To ensure that only trust-relevant aspects and not general emotional aspects are considered for the sentiment analysis, only aspects previously annotated with linguistic trust cues of perceived trustworthiness were annotated with sentiment values. However, we remain agnostic about the exact relation between sentiment analysis and trust(worthiness) because trust comprises further aspects, such as cognitive and attitudinal properties that cannot be completely captured through sentiment analysis and require consideration of further methods.

To illustrate how the annotation scheme was implemented in our dataset, consider the following example extracted from the DWDS WebXL subcorpus:

1. Auch die eloquenten, teilweise charmanten Antworten, die ChatGPT auf bestimmte Fragen gibt, sind manchmal nicht mehr als plausibel klingende Unwahrheiten – man spricht dann davon, so Horn, dass die KI "halluziniert".

‘Even the eloquent, often witty answers that ChatGPT provides to some questions are sometimes nothing more than plausible-sounding untruths – according to Horn, the AI is said to hallucinate.’

In each example, the aspects of trustworthiness (competence, benevolence, and integrity) were annotated with linguistic markers of each category according to the categorization provided in previous studies (Kuhnhenn 2014, Reinmuth 2006, de Visser et al. 2020). Afterward, the linguistic markers were annotated with trust levels/sentiment values. In example (1) the adjectives *eloquent, charmant* , nominal phrase *manchmal nicht mehr als plausibel klingende Unwahrheiten* and the verb *halluziniert* were identified as trust-relevant vocabulary. Next, they were annotated with positive trustworthiness (*eloquent, charmant*), and distrust/negative trustworthiness (*manchmal nicht mehr als plausibel klingende Unwahrheiten* and *halluziniert*). In a further step, the example was considered as distrust. In addition to linguistic cues of trustworthiness and sentiment values, trust roles and the interaction mode were annotated separately. In this case, the interaction mode is ‘industry’ (based on the information provided in the full text[4]), trust roles include ChatGPT as a trust object, and [Dennis] Horn as a trustor. Further examples can be found in the annotation guidelines.

In the next step, we focus on the correlation between trust-related vocabulary obtained by manual annotation and its sentiment values obtained by human sentiment ratings, lexicon-based, and machine-learning-based sentiment models. Positive sentiment scores are related to trustworthiness and vice versa: negative sentiment scores should be related to the erosion of trustworthiness. Neutral scores indicate that both a decrease and increase in trustworthiness can be observed or that there are no sentiment scores. Human sentiment ratings are based on the manual annotation described above. The words, multi-word units, and sentences annotated with human sentiment ratings were imported into Python to obtain their sentiment scores using machine-learning-based and lexicon-based models. We compared the distribution of human ratings with the sentiment scores provided by the pre-trained model (cf. Guhr et al., 2020) and the sentiment analyzer provided in the German language extension for textblob[5] .

4 Results

The annotation with the categories *trust*, *distrust* , and *both/ambiguous* yields that the promotion of trustworthiness occurs more frequently (57.24 %) than the lack of trustworthiness towards ChatGPT. Questions in which trust-relevant aspects could not be identified in the context (e.g., “Ist ChatGPT kostenlos?”) were excluded from the analysis. The analysis indicates that ca. 75 % of data accounts for the aspect *competence* , usually regarding how ChatGPT can be trusted to provide users with accurate information and ensure that the provided information is reliable. The manual annotation yielded some 6480 trust-relevant linguistic markers on the word-, multiword-, and sentence level (25 % of the total number of tokens) that were selected for further analysis. They comprise trust-relevant vocabulary annotated within each trust-relevant utterance.

Regarding the results of the sentiment analysis, the annotated words, multi-word units, and sentences were imported into Python to obtain their sentiment scores. The scores obtained by the trained model are negative (40 %) or neutral (38 %), and only 20 % of the annotated data are positive. Human ratings are 45 % negative and 52 % positive, less than 1 % was rated as neutral. As compared to human and machine-learning-based ratings, the majority of ratings obtained by textblob are neutral. In particular, textblob rated 11 % of trust-related vocabulary as negative and 25.8 % as positive. The results indicate significant differences in sentiment scores between human-, lexicon-based-, and machine-learning-based ratings, especially regarding the amount of neutral ratings in non-human-based models. Regarding the correlation between sentiment scores and human-based evaluation of trustworthiness, preliminary results indicate a higher correlation between negative sentiments and the lack of trustworthiness and vice versa between positive sentiment scores and the assignment of trustworthiness.

5 Conclusions and outlook

The paper explored indirect measures of emotional aspects of trust that go beyond linguistic units such as trust, mistrust, or trustworthiness and, in contrast to direct measures (e.g., scales), require a qualitative approach as a first step toward detecting the role of trust in a particular context. Since indirect measures are highly dependent on human interpretation, they pose a challenge to the research on trust and put the objectivity of qualitative measures into question. In this paper, we argued that Digital Humanities offers appropriate methods to remedy these issues.

The results show how qualitative and quantitative methods used in the Digital Humanities contribute to studies on trust in human-machine interaction. On the other hand, trust in automation as the object of investigation contributes to ongoing debates regarding the reliability of sentiment measures for languages other than English (Kaity and Balakrishnan, 2020) and provides empirical evidence for how sentiment scores can be used for modeling social phenomena like trust.

Fußnoten

1. The annotation guidelines are available online at https://doi.org/10.17605/OSF.IO/FVB7P .
2. https://www.dwds.de/d/korpora/webxl
3. https://www.maxqda.com/
4. https://web.archive.org/web/20230616185754/ https://www.boersenblatt.net/news/boersenverein/digitaler-wandel-nachhaltig-gedacht-289685
5. https://textblob-de.readthedocs.io/

Bibliographie

Belosevic, Milena. 2022. *Vertrauen und Misstrauen in der Flüchtlingsdebatte 2015-2017. Eine diskurslinguistische Untersuchung von Argumentationsmustern* . Hamburg: Buske. https://doi.org/10.46771/978-3-96769-198-6 .

Coeckelbergh, Mark. 2011. „You, Robot: On the Linguistic Construction of Artificial Others“. *AI & SOCIETY* 26 (1): 61–69. https://doi.org/10.1007/s00146-010-0289-z .

De Visser, Ewart J., Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. International Journal of Social Robotics 12 (2): 459–78. https://doi.org/10.1007/s12369-019-00596-x .

Farhat, Faiza, Shahab Saquib Sohail, und Dag Øivind Madsen. 2023. „How Trustworthy is ChatGPT? The Case of Bibliometric Analyses“. Preprint. Social Sciences. https://doi.org/10.20944/preprints202303.0479.v1 .

Funke, Noemi, Katja Stadler, Heidi Vakkuri, Anna Wagner, Marc Lunkenheimer, und Alexander H. Kracklauer. 2023. „‚Your Conversational Partner Is a Chatbot' - An Experimental Study on the Influence of Chatbot Disclosure and Service Outcome on Trust and Customer Retention in the Fashion Industry." https://doi.org/10.25929/JAIR.V1I1.113 .

Guhr, Oliver, Anne-Kathrin Schumann, Frank Bahrmann, und Hans Joachim Böhme. 2020. „Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems". In *LREC 2020 Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Palais Du Pharo, Marseille, France: Conference Proceedings* , herausgegeben von Nicoletta Calzolari, 1627–32. Paris: ELRA - European Language Resources Association. https://aclanthology.org/2020.lrec-1.202 .

Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. De Visser, und Raja Parasuraman. 2011. „A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction". *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53 (5): 517–27. https://doi.org/10.1177/0018720811417254 .

Hendriks, Friederike, Bettina Distel, Katherine M. Engelke, Daniel Westmattelmann, und Florian Wintterlin. 2021. „Methodological and Practical Challenges of Interdisciplinary Trust Research". In *Trust and Communication* , herausgegeben von Bernd Blöbaum, 29–57. Cham: Springer. https://doi.org/10.1007/978-3-030-72945-5_2 .

Huang, Xiaowei, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, et al. 2023. "A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation."

Kabir, Samia, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2023. "Who Answers It Better? An in-Depth Analysis of ChatGPT and Stack Overflow Answers to Software Engineering Questions". arXiv. https://doi.org/10.48550/arXiv.2308.02312 .

Kaity, Mohammed, und Vimala Balakrishnan. 2020. „Sentiment Lexicons and Non-English Languages: A Survey". *Knowledge and Information Systems* 62 (12): 4445–80. https://doi.org/10.1007/s10115-020-01497-6 .

Kok, Bing Cai, und Harold Soh. 2020. „Trust in Robots: Challenges and Opportunities". *Current Robotics Reports* 1 (4): 297–309. https://doi.org/10.1007/s43154-020-00029-y .

Krüger, Steffen, und Christopher Wilson. 2023. „The Problem with Trust: On the Discursive Commodification of Trust in AI". *AI & SOCIETY* 38 (4): 1753–61. https://doi.org/10.1007/s00146-022-01401-6 .

Kuhnhenn, Martha. 2014. *Glaubwürdigkeit in der politischen Kommunikation Gesprächsstile und ihre Rezeption* . Konstanz; München: UVK-Verl.-Ges.

Lee, J. D., und K. A. See. 2004. „Trust in Automation: Designing for Appropriate Reliance". *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 .

Lewicki, Roy J., Daniel J. McAllister, and Robert J. Bies. 1998. „Trust and Distrust: New Relationships and Realities". *The Academy of Management Review* 23 (3): 438. https://doi.org/10.2307/259288 .

Liu, Yang, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." https://doi.org/10.48550/ARXIV.2308.05374.

Liu, Bing. 2015. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge Univ. Press.

Loria, Steven. 2020. „textblob Documentation. Release 0.16.0". https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf .

Lotze, Netaya. 2016. *Chatbots* . Bern: Peter Lang. https://doi.org/10.3726/b10402 .

Lukyanenko, Roman, Wolfgang Maass, und Veda C. Storey. 2022. „Trust in Artificial Intelligence: From a Foundational Trust Framework to Emerging Research Opportunities". *Electronic Markets* 32 (4): 1993–2020. https://doi.org/10.1007/s12525-022-00605-4 .

Lumer, Eleonore, und Hendrik Buschmeier. 2022. „Modeling Social Influences on Indirectness in a Rational Speech Act Approach to Politeness". In *Proceedings of the 44th Annual Conference of the Cognitive Science* , herausgegeben von Jennifer Culbertson, Andrew Perfors, Hugh Rabagliati, und Veronica Ramenzoni, 2796–2802. Toronto.

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. „An Integrative Model of Organizational Trust". *The Academy of Management Review* 20 (3): 709. https://doi.org/10.2307/258792 .

Muir, Bonnie M. 1994. „Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems". *Ergonomics* 37 (11): 1905–22. https://doi.org/10.1080/00140139408964957 .

Reinmuth, Marcus. 2006. Vertrauen schaffen durch glaubwürdige Unternehmenskommunikation - Von Geschäftsberichten und den Möglichkeiten und Grenzen einer angemessenen Sprache. Dissertation. Düsseldorf. https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=3547

Rudolph, Jürgen, and Tan Samson. 2023. „ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?" *Journal of Applied Learning & Teaching* 6 (1). https://doi.org/10.37074/jalt.2023.6.1.9 .

Schäfer, Pavla. 2016. *Linguistische Vertrauensforschung: Eine Einführung* . Berlin: De Gruyter. https://doi.org/10.1515/9783110451863 .

Schneider, Britta, Bettina Migge, Doris Dippold, Iker Erdocia, Marie-Theres Fester-Seeger, Sviatlana Höhn, Ledia Kazazi, u. a. 2022. „Changing Language Ideological Concepts in the Human-Machine Era.

Questions, Themes and Topics“. https://doi.org/10.13140/RG.2.2.25867.36649 .

Shen, Xinyue, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. „In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT“. https://doi.org/10.48550/ARXIV.2304.08979 .

Watters, Casey and Michal Lemanski. 2023. „Universal Skepticism of ChatGPT: A Review of Early Literature on Chat Generative Pre-Trained Transformer“. *Frontiers in Big Data* , Nr. 6.

Wischnewski, Magdalena, Nicole Krämer, und Emmanuel Müller. 2023. „Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-of-the-Art and Future Directions“. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* , 1–16. Hamburg Germany: ACM. https://doi.org/10.1145/3544548.3581197 .

Status Quo der Entwicklungen von Ontologien Rhetorischer Figuren in Englisch, Deutsch und Serbisch

Kühn, Ramona
ramona.kuehn@uni-passau.de
Universität Passau, Deutschland

Mitrović, Jelena
jelena.Mitrovic@uni-passau.de
Universität Passau, Deutschland

Einleitung und Motivation

Rhetorische Figuren sind seit Jahrhunderten Gegenstand der Untersuchung in den Bereichen der Linguistik und der Rhetorik. Diese Figuren, wie zum Beispiel Metapher, Ironie, Alliteration und viele weitere, können als Abweichung vom normalen Sprachgebrauch verstanden werden (Fahnestock, 1999). Die daraus resultierende Wirkung ist, dass Texte oder Worte, die rhetorische Figuren enthalten, einprägsamer, überzeugender oder auch beeinflussender sein können. Wir möchten dabei Figuren unterscheiden, die sprachunabhängig funktionieren, wie zum Beispiel die Wiederholung eines Wortes, oder die Verwendung des gleichen Anfangslauts (z. B. bei der Figur Alliteration) und Figuren, die auf einem gemeinsamen kulturellen Verständnis und Hintergrundwissen basieren (z. B. Metaphern wie „jemandem das Herz brechen”) und oft nicht wörtlich, sondern nur im übertragenen Sinne verstanden werden können. Besonders bei Übersetzungen zwischen zwei Sprachen ist es immens wichtig, die Nuancen zu erkennen und richtig zu interpretieren.
Bei der Analyse dieser übertragenen Figuren können kultur- und sprachübergreifende Vergleiche angestellt werden, um Gemeinsamkeiten und Unterschiede in ihrer Verwendung und Wirkung über verschiedene kulturelle Kontexte hinweg zu untersuchen. Auch können neue Einblicke in menschliches Verhalten oder die soziale Dynamik innerhalb bestimmter Gemeinschaften gewonnen werden. Um dies zu ermöglichen, benötigt man große Datenmengen, die von Expert:innen annotiert werden. Diese manuelle Arbeit ist jedoch sehr zeit- und kostenintensiv. Darüber hinaus sind versierte Fachkräfte auf diesem Forschungsgebiet vergleichsweise selten anzutreffen. Daher ist es wichtig, dass rhetorische Figuren maschinell annotiert werden können. Eine maschinenlesbare Darstellung und Beschreibung rhetorischer Figuren ist daher notwendig. Häufig werden dafür formale, domänenspezifische Ontologien verwendet.
In diesem Artikel gehen wir der Frage nach, wie der aktuelle Status Quo der Ontologie-Entwicklung rhetorischer Figuren ist. Unser Fokus liegt dabei vor allem auf englisch-, serbisch- und deutschsprachigen Modellen. Wir geben zuerst einen Überblick über die Entwicklung und Kategorisierung von rhetorischen Figuren, um die Notwendigkeit für Ontologien hervorzuheben. Anschließend präsentieren wir aktuelle Ontologien in Serbisch, Deutsch und Englisch, sowie eine mehrsprachige Ontologie, die eine Vereinigung der vorigen Ontologien darstellt. Am Ende des Artikels geben wir einen Ausblick, wie die Zukunft bezüglich automatisierter Erkennung rhetorischer Figuren im Kontext großer Sprachmodelle („Large Language Models“) aussehen könnte.

Einblicke in die Beschreibungen Rhetorischer Figuren und ihrer Kategorisierung

Bereits Aristoteles, Sokrates und Plato beschäftigten sich mit den Grundlagen der Rhetorik und ihrer Überzeugungsfähigkeit in Argumentationen. Die erste konkrete Beschreibung rhetorischer Figuren findet sich im Buch IV der Schrift „Rhetorica ad Herennium“, dessen unbekannter Autor das Werk ca. 80 v. Chr. verfasst hat. Die Figuren werden dort in die zwei Gruppen „Figuren der Diktion“ (zum Beispiel Epanaphora, Antithesis, Klimax, Metapher) und „Gedankenfiguren“ (zum Beispiel Vergleich, Personifikation, Untertreibung, oder lebendige Beschreibung) eingeteilt (Caplan und Winterbottom, 2016). Man kann jedoch in dieser Einteilung kein einheitliches Schema erkennen. Fahnestock (1999) kritisiert, dass zusätzlich indirekt eine dritte Gruppe mit Ausnahmen eingeführt wird, sodass diese Einteilung in Gruppen inkonsistent und sogar verwirrend erscheint.

Etwa zur gleichen Zeit entstand Ciceros Werk „De inventione“, welches sich zum einen mit Grundbegriffen der Rhetorik und zum anderen mit Argumentationstechniken befasst. Es ähnelt in vielen Bereichen der Schrift Rhetorica ad Herennium. Quintilian wagt in seiner Schrift „Institutio Oratoria” (Quintilian, ca. 92 n. Chr. / 1996) einen neuen Versuch, rhetorische Figuren zu kategorisieren. Er verwendet dabei die Kategorien aus dem Werk Rhetorica ad Herennium, führt aber eine Unterscheidung zwischen „Figuren“ und „Tropen“ ein. Bereits in dieser Schrift fällt auf, dass es verschiedene Kategorisierungen und Meinungen gibt, die sich nur schwer in ein einheitliches Format bringen lassen. Dieses Dilemma zieht sich durch die Jahrhunderte hinweg, von Richard Whatelys „Elements of Rhetoric“ im Jahre 1828 über Kenneth Burkes rhetorische Theorien (1945-1966), bis Heinrich Lausberg 1960 mit seinem „Handbuch der literarischen Rhetorik“ einen neuen Versuch wagt. Das Buch ist eine enorme Sammlung von Erkenntnissen der vorigen Jahrhunderte und versucht, eine Systematik zu etablieren. Mit Hilfe mehrsprachiger Beispiele sollen den Leser:innen rhetorische Figuren und ihre Eigenheiten verdeutlicht werden. Jedoch merkt man in diesem Werk, dass es schwierig ist, allgemeingültige Standards in einem Gebiet zu etablieren, das es seit nahezu 2000 Jahren nicht geschafft hat, Kategorien ohne Widersprüche zu formulieren. Möchte man dies modellieren, wäre es notwendig, die verschiedenen Datenquellen und Definitionen zu integrieren. Außerdem müsste eine Modellierung gerade abstrakt genug sein, um auch konkurrierende Kategorien abbilden zu können, aber dennoch spezifisch genug, um die Eigenheiten der verschiedenen Figuren reflektieren zu können.

Ontologien zur Standardisierung und ihre Relevanz für die Digital Humanities

Wie soll man es nun schaffen, aus diesem Wirrwarr der Kategorien einen einheitlichen Standard zu entwickeln, der zugleich maschinenlesbar ist? Warum sollte dies überhaupt nötig sein? Rhetorische Figuren spielen auch in unserer heutigen Zeit eine große Rolle, sei es bei politischen Reden, in der Werbung oder in sozialen Medien. Denn auch heute haben sie die gleiche Wirkung wie vor tausenden Jahren: Sie können die Leserschaft überzeugen, Argumente plausibler erscheinen lassen, oder eine Botschaft einprägsamer machen. Da eine manuelle Analyse zu zeit- und kostenintensiv ist, wird eine computergestützte Erkennung immer wichtiger. Dafür sind allerdings eine einheitliche Benennung, ein gemeinsames Verständnis, sowie ein Set an Regeln erforderlich.

Ontologien sind hierfür das geeignete Mittel, da sie eine formale, einheitliche Struktur darstellen, die sowohl von Menschen als auch Maschinen lesbar und verarbeitbar ist. Sie können Daten aus heterogenen Quellen integrieren (Rehbein, 2017) und bestehende Wissensbestände vereinen (Mladenović et al., 2014). Sie dienen der Wissensrepräsentation und -organisation. Eine semantische Verarbeitung wird dadurch ermöglicht (Kaiya und Saeki, 2006). Zudem helfen sie, ein gemeinsames Verständnis der vorhandenen domänenspezifischen Informationen zu vermitteln (Noy und McGuiness, 2001). Zusammenhänge können durch die graphische Darstellung der Relationen in Form eines Netzwerkes noch leichter erkannt werden.

Konsistenz, Validität und das Einhalten logischer Regeln lassen sich mit Ontologien ebenfalls überprüfen. Ontologien haben bereits in vielen Bereichen bewiesen, dass sie unterschiedliche Definitionen und Beschreibungen vereinheitlichen und Sachverhalte konsistent modellieren können, zum Beispiel in der Medizin (Hu, 2006) oder im Katastrophenschutz (Bu Daher, 2022). Die wohl bekannteste Ontologie im Bereich der Digital Humanities ist das CIDOC Conceptual Reference Model (Bruseker et al., 2017), welches Objekte im Bereich des kulturellen Erbes beschreibt. Ein anderes Projekt in diesem Gebiet ist beispielsweise „GOLEM“ (Pianzola et al., 2023), welches mit Hilfe von Ontologien kulturelle Merkmale der Online-Fiktion in verschiedenen Sprachen miteinander vergleicht. Langmead et al. (2016) untersuchen die Relevanz von Ontologien in den Digital Humanities mit einem Fokus auf historische Netzwerke der Frühen Neuzeit.

Im Bereich rhetorischer Figuren sind Ontologien wichtige Werkzeuge für Linguist:innen und Informatiker:innen, die ein tieferes Verständnis dafür entwickeln wollen, wie Sprache in bestimmten Kontexten oder Genres verwendet wird. Rhetorische Figuren sind komplexe sprachliche Phänomene, die in verschiedenen Texten und Sprachen auftreten. Die Ontologien vereinen dabei die Vielzahl an heterogenen Datenquellen, die es im Bereich der rhetorischen Figuren gibt. Eine Ontologie rhetorischer Figuren ermöglicht die systematische Erfassung und Strukturierung des Wissens über diese Figuren. Sie hilft dabei, die verschiedenen Figuren, ihre Definitionen, Eigenschaften und Beziehungen zueinander in einer formalen und einheitlichen Weise darzustellen. Ontologien, in denen verwandte rhetorische Figuren nach ihren Konstruktionsregeln oder Eigenschaften gruppiert sind, dienen dabei unter anderem als Nachschlagewerk und Sammlung des Wissens über rhetorische Figuren. Darüber hinaus können sie in der Erkennung rhetorischer Figuren in Texten unterstützen und dabei helfen, Annotationsrichtlinien zu erstellen (Reiter, 2020). Diese Annotationen können dann für das Training von großen Sprachmodellen verwendet werden, um eine automatische Identifizierung in Texten zu ermöglichen.

Die Bedeutung von Ontologien für rhetorische Figuren in den Digital Humanities liegt in ihrer Fähigkeit, semantische Strukturen und Beziehungen zwischen den Figuren zu erfassen und zu organisieren. Zudem können solche Ontologien dazu beitragen, das Textverständnis zu steigern, da die darin erkannten Figuren direkt mit ihrer Wirkung verknüpft werden können, die sie auf die Zielgruppe haben. Zudem können kultur- und sprachübergreifende Analysen durchgeführt werden, da bestimmte Figuren in manchen Spra-

chen eventuell häufiger verwendet werden als in anderen. Hinzu kommt, dass Definitionen mancher Figuren oder sogar Figuren selbst sprachabhängig existieren und dadurch Gemeinsamkeiten und Unterschiede in den Sprachen und über verschiedene kulturelle Kontexte hinweg untersucht werden können.

Status Quo der Entwicklungen von Ontologien Rhetorischer Figuren und ihre Anwendungen

Im Folgenden stellen wir verschiedene Ontologien rhetorischer Figuren und ihre Entwicklung vor. Dabei gehen wir nur auf Ontologien ein, die mehrere verschiedene rhetorische Figuren abbilden und sich nicht auf eine einzelne Figur fokussieren. Den Beginn der Entwicklung solcher Ontologien markiert das sogenannte RhetFig Projekt (Kelly et al., 2010), dessen Ziel es war, ein theoretisches Konzept für eine mögliche Modellierung von Ontologien rhetorischer Figuren zu erstellen. Dabei werden rhetorische Figuren in drei Gruppen kategorisiert, nämlich Tropen, Schema und Chroma. Zudem werden linguistische Domänen (z. B. morphologisch, phonologisch) eingeführt, die jeder Figur zugeordnet werden. Dieses Konzept diente als Basis für die serbische RetFig (ohne „h“) Ontology (Mladenović und Mitrović, 2013). In dieser wird die Einteilung in linguistische Domänen und Gruppen fortgeführt. 98 serbische Figuren wurden dabei formal in der W3C Web *Ontology* Language (OWL) modelliert und beschrieben. Wang et al. (2021) entwickelten dann die Ploke Ontologie, die Figuren der perfekten lexikalen Wiederholung modelliert. Diese Ontologie basiert ebenfalls auf der Grundstruktur des RhetFig Projekts. In dieser Ontologie werden zusätzlich neurokognitive Eigenschaften eingeführt, die von einer Wiederholungsfigur ausgelöst werden (z. B. „Attention Effect“). Eine deutsche Ontologie für rhetorische Figuren stellt die GRhOOT Ontologie (Kühn et al., 2022) dar, die wiederum auf der serbischen RetFig Ontologie basiert, aber insgesamt 110 verschiedene Figuren modelliert. Dafür wurde für jede Figur einzeln untersucht, ob sie im Deutschen ebenfalls existiert und ob die Definitionen gleich sind. Dabei wurde festgestellt, dass 12 serbische Figuren so nicht im Deutschen existieren (z. B. die serbische rhetorische Figur „Wunsch“ hat kein Pendant im Deutschen). Circa 60 Figuren konnten gleich modelliert werden. Allerdings gab es auch Fälle wie die Figur „Syllepsis“, deren deutsche Definition besagt, dass ein Wortelement weggelassen wird, während bei der serbischen Figur „Silepsa“ ein Wortelement hinzugefügt wird. Inspiriert durch diese sprachlichen Unterschiede wurde die „Multilingual Ontology” (Wang et al., 2022) modelliert, die versucht, Ploke Figuren – also Figuren mit perfekter lexikaler Wiederholung – der serbischen, deutschen und englischen Ontologie zu vereinen. Wie wir bereits anfangs festgestellt haben, sollten diese Figuren sprachlich unabhängig sein. Jedoch wurden dabei ebenfalls Unterschiede festgestellt. Beispielsweise lautet die Definition der Figur Epizeuxis im Deutschen, dass ein Wort oder Wortgruppe mindestens drei Mal wiederholt wird. Die englische Definition spezifiziert jedoch nicht näher, wie oft ein Wort wiederholt werden muss, wodurch bereits eine zweimalige Wiederholung auf die Figur Epizeuxis hindeuten kann.

Diese Beispiele zeigen deutlich, dass existierende Ontologien rhetorischer Figuren nicht einfach in andere Sprachen übersetzt werden können. Für jede einzelne Figur muss verglichen werden, ob sie so in der Zielsprache existiert und die Definition gleich ist. Zudem müssen die Figuren gefunden werden, die es in der Sprache der existierenden Ontologie nicht gibt, allerdings in der Zielsprache als rhetorische Figur angesehen wird. Genau so wurde verfahren, um eine englische Ontologie zu modellieren, die mehrere verschiedene Figuren abbildet. Mit dieser sogenannten ESTHER Ontologie (Kühn et al. 2023) wurde das englische Äquivalent zur deutschen und serbischen Ontologie geschaffen. Sie ist eine Weiterentwicklung der deutschen GRhOOT Ontologie und beinhaltet zusätzlich Wissen aus weiteren englischen Quellen. Auch wurden hierarchische Abhängigkeiten und Co-Lokationen zwischen verschiedenen Figuren modelliert, zum Beispiel, dass die Figur Symploke stets auch eine Anapher und Epipher ist, oder dass eine Klimax aus der Kombination von Gradatio und Inkrementum gebildet wird (O’Reilly et al. 2018). Eine zeitliche Übersicht der hier genannten Ontologien findet sich in Abbildung 1.

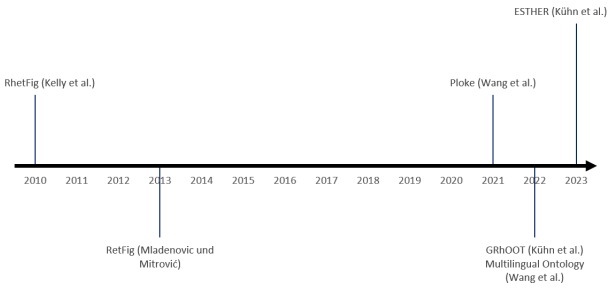


Abbildung Zeitliche Übersicht der Ontologie-Entwicklungen.

Doch wie sollen diese Ontologien nun dazu beitragen, große Datenmengen zu analysieren und rhetorische Figuren darin automatisch zu identifizieren? Kühn und Mitrović (2023) zeigen dafür einige Anwendungsbeispiele auf. Sie nennen als größtes Problem für die automatisierte Erkennung das Fehlen annotierter Datensätze. Die Ontologien sollen dabei unterstützen, diese Hürde zu überwinden, um dann mit annotierten Daten große Sprachmodelle für diese spezielle Aufgabe zu trainieren. Konkret können die Ontologien dazu genutzt werden, Personen in der manuellen Annotation zu schulen, um die wenigen Expert:innen darin zu unterstützen. Da die Ontologien in OWL geschrieben sind, können die Daten mit der Anfragesprache SPARQL angefragt werden. Zudem beschreiben Kühn und Mitrović (2023) das theoretische Konzept einer grafische Benutzer-

oberfläche, die ein niedrigeschwelliges Angebot zur Annotation rhetorischer Figuren für viele Nutzer:innen darstellt. Dabei handelt es sich um eine Website mit Auswahlmöglichkeiten, die die Nutzer:innen aktiv dabei unterstützt, Figuren in einem Text bestimmen zu können. Die Anfragen im Hintergrund basieren dabei auf SPARQL-Anfragen an die jeweilige Ontologie. Diese Oberfläche befindet sich allerdings noch in der Entwicklung. Da die serbische RetFig, die deutsche GRhOOT und die englische ESTHER Ontologie auf der gleichen Architektur basieren, kann dieses Tool jedoch dann sprachübergreifend eingesetzt werden. Zusätzlich können Gamification Elemente das Angebot noch attraktiver für eine größere Nutzerschaft machen.

Zusammenfassung und Ausblick

Wir haben gezeigt, wie aktuelle Entwicklungen bezüglich maschinenlesbarer und formaler Ontologien im Bereich rhetorischer Figuren voranschreiten. Dadurch, dass die Ontologien alle aufeinander aufbauen aber dennoch unterschiedliche Sprachen betrachten, ist eine strukturelle Kompatibilität gegeben.

Die in den Ontologien spezifizierten Konstruktionsregeln rhetorischer Figuren können Sprachmodelle trainiert werden. Diese können dabei auch dazu dienen, Sätze oder ganze Texte mit gewünschten rhetorischen Figuren zu generieren. Somit könnte der Mangel an Daten von gewissen Figuren in diesem Bereich überwunden werden. Außerdem muss erforscht werden, inwiefern Sprachmodelle bereits jetzt rhetorische Figuren in Texten oder Sätzen erkennen und korrekt annotieren können. Sie würden damit auch eine kostengünstige Alternative zur manuellen Annotation schaffen. Bis dies jedoch möglich ist, brauchen wir unsere Ontologien und manuelle Bemühungen, denn die Sprachmodelle sind nur so gut wie die verfügbaren Trainingsdaten. Dazu gehören ethische Überlegungen wie die Vermeidung von Bias und die Gewährleistung der Transparenz bei der Generierung von Texten. Außerdem können Sprachmodelle nicht ohne menschliche Überprüfung als alleinige Quelle der Wahrheit betrachtet werden, da sie auf vorherigen Textdaten trainiert werden und daher deren Vorurteile und Einschränkungen reflektieren können.

Wenn wir jedoch in der Lage sind, rhetorische Figuren automatisiert zu erkennen und zu identifizieren, bekommen wir einen besseren Einblick in die Art und Weise, wie wir sowohl in schriftlicher als auch in mündlicher Form effektiv kommunizieren - etwas, das zweifellos im Laufe der Zeit relevant bleiben wird, unabhängig von den technologischen Fortschritten, die in den kommenden Jahren gemacht werden.

Acknowledgment

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049. The author is responsible for the content of this publication.

Bibliographie

Bruseker, George, Nicola Carboni and Anais Guillem. 2017. “Cultural heritage data management: The role of formal ontology and CIDOC CRM.” *Heritage and archaeology in the digital age: acquisition, curation, and dissemination of spatial cultural heritage data*: 93-131.

Bu Daher, Julie, Tom Huygue, Patricia Stolf and Nathalie Hernandez. 2023. “An Ontology and a reasoning approach for Evacuation in Flood Disaster Response.” *Journal of Information & Knowledge Management*, 2350042.

Caplan, Harry and Mark Winterbottom. 2016. *Rhetorica ad herennium.* Oxford Research Encyclopedia of Classics.

Fahnestock, Jeanne. 1999. *Rhetorical figures in science.* Oxford University Press, USA.

Hu, Xiaohua. 2006. “Natural language processing and ontology-enhanced biomedical literature mining for systems biology.” In *Computational systems biology*: 39–56. Elsevier.

Kaiya, Haruhiko and Motoshi Saeki. 2006. “Using domain ontology as domain knowledge for requirements elicitation.” In *14th IEEE International Requirements Engineering Conference (RE'06):* 189-198. IEEE.

Kelly, Ashley R., Nike A. Abbott, Randy Allen Harris, Chrysanne DiMarco and David R Cheriton. 2010. “Toward an ontology of rhetorical figures.” *Proceedings of the 28th ACM International Conference on Design of Communication.*

Kühn, Ramona and Jelena Mitrovic. 2023. “Multilingual Domain Ontologies of Rhetorical Figures and Their Applications.” *UniDive General Meeting, Paris.*

Kühn, Ramona, Jelena Mitrović and Michael Granitzer. 2022. “GRhOOT: Ontology of Rhetorical Figures in German.” *Proceedings of the Thirteenth Language Resources and Evaluation Conference.*

Kühn, Ramona, Jelena Mitrović and Michael Granitzer. 2023. “ESTHER: Ontology of Rhetorical Figures in English.” *Proceedings of the Joint Ontology Workshops*. To be published.

Langmead, Alison, Jessica M. Otis, Christopher N. Warren, Scott B. Weingart and Lisa D. Zilinksi. 2016. “Towards interoperable network ontologies for the digital humanities.” *International Journal of Humanities and Arts Computing, 10 (1)*: 22-35.

Lausberg, Heinrich. 1960. *Handbuch der literarischen Rhetorik.* Vol. 2. München. Hueber.

Mladenović, Miljana and Jelena Mitrović. 2013. “Ontology of Rhetorical Figures for Serbian.” *Text, Speech, and Dialogue: 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5. Proceedings 16*. Springer Berlin Heidelberg.

Mladenović, Miljana, Jelena Mitrović and Cvetana Krstev. 2014. “Developing and maintaining a wordnet: Procedures and tools.” In *Proceedings of the Seventh Global Wordnet Conference:* 55-62.

Noy, Natalya F. und Deborah L. McGuinness. 2001. *Ontology development 101: A guide to creating your first ontology*.

O’Reilly, Cliff, Yetian Wang, Katherine Tu, Sarah Bott, Paulo Pacheco, Tyler William Black and Randy Allen Harris. 2018. “Arguments in gradatio, incrementum and climax; a climax ontology.” In *Proceedings of the 18th workshop on Computational Models of Natural Argument. Academic Press*.

Pianzola, Federico, Xiaoyan Yang, Noa Visser, Michiel van der Ree and Andreas van Cranenburgh. 2023. “Constructing the GOLEM: Graphs and Ontologies for Literary Evolution Models.” *Digital Humanities 2023*. Collaboration as Opportunity (DH2023), Graz, Austria. https://doi.org/10.5281/zenodo.8107749

Quintilian, Marcus Fabius. Ca. 92 n. Chr. *Institutio oratoria.* Übersetzt von HE Butler. 1996.

Rehbein, Malte. 2017. “Ontologien”. *Digital Humanities: Eine Einführung*: 162-176.

Reiter, Nils. 2020. “Anleitung zur Erstellung von Annotationsrichtlinien.“ In Nils Reiter/Axel Pichler/Jonas Kuhn (Hg.), *Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*: 193-202. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110693973-009

Wang, Yetian, Ramona Kühn, Randy Allen Harris, Jelena Mitrović and Michael Granitzer. 2022. “Towards a Unified Multilingual Ontology for Rhetorical Figures.” *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Valletta, Malta: SCITEPRESS-Science and Technology Publications.*

Wang, Yetian, Randy Allen Harris and Daniel M. Berry. 2021. “An Ontology for Ploke: Rhetorical Figures of Lexical Repetitions.” *JOWO.*

Synergieeffekte zwischen Henze-Digital und der Carl Maria von Weber-Gesamtausgabe durch die bilaterale Weiterentwicklung der WeGA-WebApp

Ried, Dennis
dennis.ried@uni-paderborn.de
Universität Paderborn, Deutschland

Bei »Hans Werner Henzes künstlerisches Netzwerk« (kurz: »Henze-Digital«/»HenDi«) handelt es sich um ein DFG-gefördertes Editionsprojekt[1], in dessen Zentrum die Erschließung der Korrespondenz Hans Werner Henzes (1926–2012) steht. Seit August 2021[2] am Musikwissenschaftlichen Seminar Detmold/Paderborn angesiedelt, wird eine Digitale Edition der postalischen Dokumente angestrengt. Aufgrund des Umfangs der überlieferten Dokumente (mehr als 11.000) werden in der ersten Förderphase die Korrespondenzen (d. h. Hin- und Rückbriefe) zwischen Henze und ausgewählten Librettisten[3] und Henzes Mäzen Paul Sacher erschlossen.

Für die digitale Erschließung der Korrespondenzen wird der Codierungsstandard TEI verwendet. Die semantische Anreicherung des edierten Texts erfolgt durch textkritische Auszeichnung und die Identifikation von Entitäten (z. B. Personen, Organisationen, Orte, Werke usw.) sowie deren Verknüpfung mit projekteigenen Referenzdatensätzen. Diese internen Authority Files werden wiederum mit Normdatenbanken wie der GND, VIAF oder der geonames-Datenbank verknüpft, um die Vernetzbarkeit und Nachnutzbarkeit der Inhalte sicherzustellen.

Bezüglich der Codierungsrichtlinien hat die Weber-Gesamtausgabe (WeGA) gegenüber Henze-Digital einen enormen Vorsprung. Seit 2011, d. h. seit mehr als einer Dekade werden Briefe in der WeGA digital ediert und bereitgestellt (WeGA 2023), sodass ausdifferenzierte Codierungs- und Editionsrichtlinien geschaffen wurden, die Expertise im Umgang mit TEI-basierten Briefeditionen widerspiegeln und auf denen Henze-Digital nun aufbauen kann. Unproblematisch war die Adaption jedoch nicht, da sich das Projekt Henze-Digital mit einem ganz anderen Forschungsgegenstand auseinandersetzt als die WeGA. Die Unterschiede sind nicht nur durch den historisch-zeitlichen

Abstand bedingt – etwa 200 Jahre, in denen bspw. der Wechsel von Kurrent zu lateinischer Schreibschrift vollzogen wird –, sondern vor allem auch durch den technischen Fortschritt. In den Korrespondenzen Webers finden sich zum Beispiel keine Typoskripte oder Telegramme, da solche Medien noch nicht verfügbar waren. Dennoch kann ein Großteil der Codierungsrichtlinien ohne größere Anpassung auch auf Korrespondenzdokumente angewandt werden, die im 20. Jahrhundert verfasst wurden, weshalb der Entschluss gefasst wurde, die TEI-Schemata der WeGA nachzunutzen.

Die TEI-ODDs der WeGA enthalten die Spezifikationen und die Dokumentation zur entsprechenden Entität, für die das Schema definiert ist.[4] Hieraus werden dann u. a. die RelaxNG-Schemata abgeleitet, die wiederum für die Validierung der Daten verwendet werden. Diese Methode hat sich in der WeGA seit vielen Jahren bewährt und stellt nicht zuletzt durch die stetige Verbesserung und Erweiterung eine solide Basis für die Nachnutzung dar. Durch die Publikation der ODDs als open source bei Github und Zenodo (Stadler 2023) ist die Nachnutzbarkeit sichergestellt.

In der von der WeGA angefertigten TEI-Customization werden die Anpassungen zum TEI-All-Schema definiert. Die Methode des ODD-Chaining (Burnard 2016), die Henze-Digital zur Nachnutzung verwendet, setzt in zweiter Instanz an und definiert die Anpassungen zu den Anpassungen der WeGA-Customization – ähnlich einem Zug, bei dem ein weiterer Wagen angehängt wird (vgl. auch Stadler/Bohl/Viglianti 2020). Technisch betrachtet wird im HenDi-ODD lediglich eine andere Quelle angegeben als bei der WeGA. In erster Instanz (WeGA) ist der TEI-Standard die Quelle. In den ODDs von Henze-Digital werden jedoch die kompilierten ODDs der WeGA als Quellen eingebunden. Beim Generieren der Schemata wird dann die gesamte Kette an ODDs (daher ODD-Chaining) ausgelesen und zu einem neuen Schema kompiliert. So können bestehende Datenmodelle nachgenutzt werden, wobei Nachnutzung hier konkret die Verwendung und Anpassung der vorhandenen Customization meint.

Dieser Prozess ist überaus abstrakt, da beim Hinzufügen, Löschen und Modifizieren von Elementen, Attributen usw. nicht immer offensichtlich ist, worauf aufgebaut wird. Es ist immer zu prüfen, ob die Modifikation an (Standard-)TEI oder dem WeGA-TEI ansetzt. Darüber hinaus ist ODD-Chaining vor allem in den Grundsätzen gut dokumentiert, jedoch weniger für die fortgeschrittene Anwendung, sodass sich zusätzlich Unsicherheiten ergeben: Habe ich das ODD falsch codiert? Handelt es sich hier um einen Bug in den TEI-Stylesheets?

Ferner ist man beim ODD-Chaining immer von der Codierung des zugrunde liegenden ODDs abhängig. Werden beispielsweise Attribute vermehrt direkt verändert, kann es zu Problemen kommen, wenn Attributklassen aus TEI übernommen werden. Solche Problematiken, die sich aus der Codierung der ODDs herleiten, schlagen sich auch in der Modifikation von bereits modifizierten Attributen und Elementen nieder.

Dennoch ist die Verwendung von ODD-Chaining hier lohnend, da auf die jahrelange Erfahrung im Umgang mit TEI zurückgegriffen werden und ein Teil der Dokumentation weiterverwendet werden kann. Letzteres betrifft vor allem die Codierungsbeispiele, die – sofern nicht modifiziert – automatisch in die neue Dokumentation übernommen werden.

Während in den WeGA-ODDs allgemeine Modifikationen in einer Datei zentral definiert sind, sodass sie in diverse Schemata eingebunden werden können, sind die Schemadefinitionen bei Henze-Digital fast ausschließlich zentral definiert. Dies führt zu einer Verschlankung der einzelnen ODDs, kann aber hinsichtlich des ODD-Chaining Probleme verursachen, wenn die Datenbasis eigentlich eine andere Handhabung erfordert.

Für die Publikation und Vermittlung der Forschungsdaten wird im Projekt Henze-Digital bewusst keine eigene Softwarelösung erarbeitet. Dies wäre in der dreijährigen Förderphase nicht umsetzbar gewesen. Stattdessen werden die vorhandenen Ressourcen in die Weiterentwicklung der WeGA-WebApp, eine im Rahmen der WeGA entwickelte Forschungssoftware, investiert. Da die WeGA-WebApp bereits die Forschungsdaten der WeGA verarbeiten kann, müssen also ›nur‹ die neuen Anpassungen für Henze-Digital implementiert werden – so die Theorie. Ausgewählt wurde diese Forschungssoftware aber vor allem aufgrund der Tatsache, dass sie v. a. im Bereich der Musikwissenschaft nach wie vor als State-of-the-Art gilt. Zu diesem Schluss gelangt auch Thorsten Röder (2020) in einer 2020 angefertigten Rezension für RIDE: »Ohnehin definiert die Weber-Briefausgabe in vielerlei Hinsicht den ›State of the Art‹ der digitalen Briefedition«.

Dass die aus der Weiterentwicklung entstehende HenDi-WebApp sowie deren Ursprung (WeGA-WebApp) eine stehende Verbindung aufweisen, ist Teil des methodischen Konzepts. Um diese Verbindung gewährleisten zu können, musste ein Workflow erarbeitet werden, der es erlaubt, die Nachnutzung der WeGA-WebApp ähnlich dem Prinzip der ODD-Chaining-Methode zu gestalten. Schließlich sollten die Anpassungen und Weiterentwicklungen, die im Rahmen von Henze-Digital nötig waren, auch einen Synergieeffekt für in die WeGA-WebApp erzeugen. In diesem Sinne ist Nachnutzung hier als Mitnutzung und Weiterentwicklung der Software zu verstehen. Wie aber kann das bewerkstelligt werden?

Ein erster Ansatz war, den Quellcode der WebApp als Fork (git) zu modifizieren, um die Abhängigkeit zum Originalcode zu erhalten. Dies hat sich recht schnell als untauglich erwiesen, da ein Fork nicht die nötigen Rahmenbedingungen liefert; der Fork konnte sich selbstständig weiterentwickeln und zu weit vom Ursprung (origin) entfernen. Dies sowie tiefergehende Anpassungen in der Infrastruktur machten Updates aus der WeGA bereits nach kurzer Zeit in einem arbeitsökonomischen Sinne unmöglich.

Als zweiten Ansatz wurde versucht, eine Methode zu entwickeln, die dem ODD-Chaining-Prinzip ähnelt: Zunächst wurde der Quellcode der WeGA-WebApp kompiliert (.xar), um eine vollfunktionsfähige Applikation zu er-

halten.[5] In einem nächsten Schritt wurde der Quellcode der Applikation wieder entpackt und die Teile ersetzt, die in der HenDi-WebApp in modifizierter Form nötig sind, bevor das Paket wieder geschnürt wurde. Damit lag eine vollfunktionsfähige WebApp vor, mit den Spezifikationen für die Daten aus dem Projekt Henze-Digital. Doch auch dieser Weg hat sich nur bedingt als brauchbar erwiesen, da Modifikation an der Grundstruktur der Applikation auf diesem Weg einen enormen Aufwand bedeuteten, sofern sie überhaupt möglich waren. Daher wurde bei Henze-Digital in einem dritten Ansatz dazu übergegangen, den Quellcode der App direkt zu modifizieren.

In der Entwicklung bedeutet das, dass zunächst alle Dateien, die modifiziert werden müssen, in einem eigenen HenDi-Repository vorgehalten werden. Beim build-Prozess der Applikation wird der Quellcode der WeGA-WebApp dann gezielt überschrieben, sodass die Applikation nur ein einziges Mal kompiliert werden muss.

Weiterentwicklungen seitens der WeGA haben natürlich auch Auswirkungen auf die WebApp von Henze-Digital, weshalb der Austausch der Entwickler aus beiden Projekten von großer Bedeutung ist. Der Quellcode der WeGA-WebApp liegt in einem öffentlich zugänglichen Github-Repository. Der modifizierte Quellcode für die HenDi-WebApp wird getrennt in einem separaten Repository organisiert, um die Unabhängigkeit beider Datenbestände sicherzustellen.[6] Verbunden sind die beiden Repositories dadurch, dass das WeGA-WebApp-Repository als Submodule (git) im HenDi-WebApp-Repository initialisiert ist. Somit ist ein stabiler Zustand des Quellcodes der WeGA-WebApp lokal vorhanden und kann direkt in den build-Prozess eingebunden werden. Durch die Einbindung als Submodule kann gesteuert werden, welcher Zustand der WebApp zugrunde gelegt wird, unabhängig vom Releasezyklus. Dies erleichtert auch das Testen von Quellcode-Updates.

Über die lokale Zusammenführung von Quellcode (Submodule) und modifiziertem Quellcode werden, ist es möglich, die Daten zur WeGA-WebApp im Submodule lokal zu überschreiben und dann mit den dort vorhandenen Routinen zu kompilieren. Durch Modifikationen, die im Rahmen von Henze-Digital nötig waren, ist das build-Skript der WeGA (ANT) jedoch nicht mehr eins-zu-eins anwendbar. Für Henze-Digital wurde folglich ein eigenes build-Skript aufgesetzt, dass das der WeGA jedoch inkludiert und soweit möglich auf bereits bestehende Infrastruktur zurückgreift. Auf diese Weise kann jede einzelne Stufe beim Kompilieren kontrolliert und bei Bedarf angepasst werden.

Doch wie wird mit den Anpassungen am Code konkret umgegangen? Schließlich ist ein Bugfix in der HenDi-WebApp nur bedingt sinnvoll, wenn dieser Fehler in der WeGA-WebApp bestehen bliebe.

Um diesem Problem entgegenzuwirken, arbeiten die WeGA und Henze-Digital eng und bei Bedarf auch projektübergreifend zusammen. Wird beispielsweise ein Bug in der WebApp bei Henze-Digital festgestellt, so wird geprüft, ob dieser auch im Kontext der WeGA relevant ist. Wenn ja, dann behebt eines der beiden Projekte das Problem im Quellcode der WeGA-WebApp.[7] Die Änderung finden dann nach einem Update der zugrunde liegenden WebApp-Version über den build-Prozess automatisch den Weg in die HenDi-WebApp. Dies hat den Vorteil, dass sowohl die WeGA als auch alle Derivate hiervon profitieren können. Dieses Vorgehen nützt also nicht nur WeGA und Henze-Digital.

Ähnlich wird auch bei der Programmierung neuer Features verfahren. Entsteht z. B. bei Henze-Digital ein Feature, das nicht nur im projektspezifischen Kontext einsetzbar ist, wird projektübergreifend beraten, wie dieses Feature weiter generalisiert werden und Eingang in die WeGA-WebApp finden kann. Anschließend wird das Feature quasi projektfremd weiterentwickelt: Es wird als neue Funktionalität in die WebApp der Weber-Gesamtausgabe implementiert. Durch das verwendete Nachnutzungsverfahren kann der entsprechende Quellcode dann im Repository von Henze-Digital entfernt werden, da der zugrunde liegende Quellcode das Feature dann bereits enthält und die Funktionalität nicht überschrieben bzw. eingeschrieben werden muss.

Dadurch, dass alle Arbeiten für die WeGA-WebApp auch wieder zu Henze-Digital zurückfließen, ist die Arbeit am Ende nicht wirklich projektfremd. Dieses Vorgehen erzeugt Synergieeffekte, die beiden Projekten zugutekommen und Ressourcen sparen, da eine Software aus zwei Perspektiven gepflegt und erweitert wird. Ferner kann die WeGA-WebApp dadurch optimal nachgenutzt werden und es entsteht ein erweiterter Support für diese Software.

Fußnoten

1. Digitale Briefedition: Hans Werner Henzes künstlerisches Netzwerk, Kurztitel: Henze Digital, https://web.archive.org/web/20230719064848/https://gepris.dfg.de/gepris/projekt/459602398?context=projekt&task=showDetail&id=459602398& (zugegriffen: 19. Juli 2023).
2. Erste Förderphase 2021–2024.
3. Hans Magnus Enzensberger, Wystan Hugh Auden und Chester Kallman, Grete Weil und Walter Jockisch, Miguel Barnet, Friedrich Hitzer.
4. Für jeden Dokumenttyp (bspw. Personen, Briefe, Schriften, Werke, Orte) sind eigene Schemata definiert.
5. Alternativ hätte auch ein stabiles Release verwendet werden können, jedoch war es wichtig auf der aktuellen Entwickler-Version aufzubauen, um die Unterschiede bei Updates möglichst gering zu halten.
6. Der Quellcode der HenDi-WebApp ist auf Github öffentlich zugänglich. Jedes Release wird zudem auf Zenodo (z. B. Ried 2023) publiziert.
7. Veränderungen an der WeGA-WebApp werden stets als Pull Request gestellt, sodass die Funktionsweise der App nicht durch Außeneinwirkung negativ beeinflusst werden kann. Nach einem Code-Review, Tests und der Freigabe durch die Hauptverantwortlichen, wird der Code dann ins System übernommen.

Bibliographie

Burnard, Lou. 2016. Lou "ODD chaining for Beginners". https://web.archive.org/web/20230719071637/http://teic.github.io/PDF/howtoChain.pdf (zugegriffen: 19. Juli 2023).

Ried, Dennis. 2023 "HenDi-WebApp". Version 1.0.1. 10.5281/zenodo.8330256.

Roeder, Torsten. 2020 "Die offene Editionswerkstatt: Carl Maria von Webers Briefe in der digitalen WeGA". RIDE 12. 10.18716/ride.a.12.4.

Stadler, Peter/Bohl, Benjamin W./Viglianti, Raffaele. 2020 "Einführung in TEI-ODD". 10.5281/zenodo.4621940

Stadler, Peter/et. al. 2023. "Edirom/WeGA-ODD: WeGA ODD files release 4.7.0". 10.5281/zenodo.7652568.

[WeGA]. 2023. " Carl-Maria-von-Weber-Gesamtausgabe. Digitale Edition", http://weber-gesamtausgabe.de/A070006 (Version 4.7.0 vom 19. Februar 2023).

The Future of Philosophy In the Digital Humanities

Heßbrüggen-Walter, Stefan
early.modern.thought.online@gmail.com
Universitäten Tours / Orléans, Frankreich

When thinking about the relation between philosophy and the digital humanities, a number of interdependent problems need to be resolved:

1. Is there a unified method for the digital humanities and, if so, is it applicable to philosophy?

2. If we believe that the digital humanities are committed to 'the scientific method', must philosophy be regarded as a science in order to be part of the digital humanities?

3. Conversely, if DH is not committed to 'the scientific method', does that mean that philosophy must be practised as a 'humanist discipline' in order to be part of the digital humanities?

This contribution will not present conclusive answers to these questions. It will rather use them as background for the introduction for an alternative understanding of the relation between philosophy and the digital humanities that is inspired by the founder of the Vienna Circle, Moritz Schlick, a conception that, as I believe, can provide useful orientation for the future of philosophy in the digital humanities. The first section of this contribution contains some preliminary reflections on the three questions regarding method that I introduced above. This will lead to a brief discussion of Schlick's understanding of the relation between philosophy and other disciplines, in particular the sciences. I will close with some indications how to apply lessons to be learned in Schlick to understanding the relation between philosophy and the digital humanities.

'Methods' And 'the Scientific Method'

When we talk about 'method' in the digital humanities, we can mean two different things:

1. a specific 'research method', e. g. topic modeling or network analysis, dedicated to a specific group of problems (probability distributions of co-occurring terms or the relative importance of nodes in a network). Methods in this sense exist in the sciences, too (e. g. radio astronomy, spectroscopy).

2. If we use 'method' in the singular, we do not refer to individual methods, but an overarching understanding of how to proceed in a given discipline or field of study – or, even more ambitiously, in 'science' as such, i. e. 'the scientific method' in the singular. Here, we will concern ourselves only with the latter. The question is then: are the digital humanities a field that is defined by adhering to 'the scientific method' in much the same way as the natural and (quantitative) social sciences? Some think so (Roller 2021, Barzen / Leymann 2017), others are more sceptical (Durlacher 2022).

However, I will not address this problem directly, not least because it is not that simple to understand how to apply the notion of 'the scientific method' to what we do in the digital humanities. Instead, I want to ask what adoption of the thesis 'DH practices the scientific method' would mean for the prospects of 'digital philosophy', i. e. philosophy as part of the digital humanities. If research questions in philosophy can be answered through the application of DH methods (in the plural), and DH methods are instantiations of 'the scientific method', this would mean that philosophy – or the parts of philosophy that are amenable to such approaches – must be considered as a science.

Conversely, if we believe that DH methods are not exemplifications of 'the scientific method', we may feel tempted to regard philosophy as being fundamentally different from the sciences. This raises the question whether it may then count as one of the humanities (to be conceived in a way that makes them fundamentally different from the sciences) or whether we should understand philosophy as something that is neither a scientific nor a humanist discipline. One way to conceive philosophy as one of the humanities emphasises the role of the history of philosophy for philosophy as a whole, e. g. as history of philosophical thinking in all periods of history, all countries of the world, and for infinitely many 'thinking humans' and their ways of life, in other words a global history of philosophy in the true meaning of the term: a research program that is inspired both by Dilthey's conception of Geistesgeschichte and the potential of digital methods to process large quantities of multilingual texts (Hartung 2023, 102).

Schlick And Philosophy As Science

Some orientation in this complex and disputed area can be gleaned from an article first published by Moritz Schlick, the founder of the Vienna Circle, in 1932 and republished in a collection of his papers (Schlick 1938). As I have argued elsewhere (Heßbrüggen-Walter 2020), the rubrication of thinkers of the Vienna Circle as 'neopositivists' should not stand in the way of a deeper appreciation of their contributions, since it seems that many of the problems they grappled with resurface when thinking about the place of digital humanities in the contemporary landscape.

Schlick resolves the question whether or not to count philosophy as part of the humanities by making a distinction between the perspective of the historian and the perspective of the philosopher proper. While the historian assesses extrinsic values like the beauty or historical relevance of a philosophical text or author, the philosopher is primarily interested in whether or not it contains truths (Schlick 1938, 118). The perspective of the historian is, according to Schlick, not per se illegitimate. It only becomes misleading when we are tempted to draw philosophical conclusions from it, i. e. when we derive from the many fruitless attempts to put forward philosophical truths the sceptical conclusion that progress in philosophy is impossible (Schlick 1938, 120). If we assume that history of philosophy is part of the humanities in this understanding, digital approaches in this domain are on a part with the role of digital methods in the humanities at large. But we must accept the limitation that their results will not contribute to a better understanding of what Schlick takes to be philosophy strictu sensu.

The relationship between philosophy and the sciences is more complex. Schlick reminds the reader that the opposition between both is a product of the late 18th and 19th century when the disciplines now comprising the sciences emancipated themselves from philosophy and, conversely, philosophers began to propagate philosophy as a science in its own right: the philosopher "sits in his library, he consults innumerable books, he works at his desk and studies various opinions of many philosophers as a historian would compare his different sources, or as a scientist would do while engaged in some particular pursuit in any special domain of knowledge; he has all the bearing and really believes that he is using in some way the scientific method, only doing so on a more general scale." (Schlick 124) Schlick provides mainly two reasons why this self-image of the philosopher as a scientist sui generis is misguided:

1. Philosophy in this sense has no definite domain: it concerns itself with 'most general truths', but what these truths are about is again undecided.

2. Philosophy differs from the sciences in that it is incapable of aggregating knowledge in a cooperative manner: "Scientific results go on developing, combining themselves with other achievements, and receiving general acknowledgment, but there is no such thing to be discovered in the work of the philosopher." (Schlick 1938, 124)

Where does this leave digital philosophy? If we understand its role – in analogy to this understanding of philosophy as science sui generis – as contributing to insights into philosophical problems, i. e. if we believe we could use the computer to assess the truth or falsity of philosophical propositions, such attempts might fail. This failure, however, would not be due to some deficiency in our technical solutions (methods in the first of the senses distinguished above), but due to a misunderstanding of what philosophy is about. I would not go so far as to claim that such a non-empirical a priori argument based on a certain understanding of philosophy is in itself conclusive. Rather, I think that Schlick's constructive proposal for how to understand and practice philosophy provides helpful directions for the future of digital philosophy.

Digital Philosophy As an Activity

Schlick does not think that philosophy should be abandoned. We just have misunderstood its proper place in the overall order of knowledge. Instead of trying to turn it into a science, he envisions a division of labour between philosophy and science, taking into account that both endeavours are fundamentally different. Schlick understands philosophy as a peculiar form of activity, namely the activity of 'finding meaning' or 'clarification', while science consists in the pursuit of truth. Nevertheless, philosophy and science are rather intimately related, so that at times the scientist must turn into a philosopher:

> […] sometimes in the course of their work they [sc. scientists] are surprised to find, by the contradictory results at which they arrive, that they have been using words without a perfectly clear meaning, and then they will have to turn to the philosophical activity of clarification, and they cannot go on with the pursuit of truth before the pursuit of meaning has been successful. (Schlick 1938, 130)

Digital humanists engage in clarification as soon, as they aim to translate foundational concepts of their background discipline (e. g. history or literary studies) into a form that is amenable to digital processing (see e. g. Bosse 2019 for a comprehensive analysis of what is involved in the historical concept of 'place'). But, next to that, they need to determine the meaning of concepts related to their own research practice (AG Digital Humanities Theorie 2023, Ciula et al. 2018). These practices qualify as philosophical in Schlick's sense regardless of whether practitioners are taken to be philosophers in the academic sense of the term. In fact, one can suspect that the practice of clarifying the terminology of background disciplines is closely related to how Schlick describes the situation of ethics and aesthetics, philosophical subdisciplines that "[…] do not yet possess sufficiently clear concepts, most of their work is still devoted to clarifying them, and therefore it may justly be called philosophical." (Schlick 1938, 132)

Conclusion: The Future of Digital Philosophy

We have distinguished a ‘wide’ and a ‘strict’ conception of philosophy in Schlick. Both can be meaningfully applied to philosophy as a part of the digital humanities and are not mutually exclusive. We can understand philosophy in relation to the digital humanities as a discipline of the digital humanities that engages with philosophical texts from the perspective of the humanities at large, i.e. without an interest in their truth, akin to other historical disciplines. Or we can take it to be an activity that aims to clarify our use of terms through ‘operationalisation’, i. e. through a transformation that makes them amenable to digital processing using formal (i. e. programming) languages. In this sense, every digital humanist is a philosopher. Besides that, digital humanities uses terminology that expresses specific concerns of the discipline. Such concepts are in need of clarification and ‘operationalisation’ as well.

But we do not need to stop here. Why should it not be possible to apply a strict understanding of digital philosophy to the history of philosophy at large, using tools of the digital humanities to clarify philosophical terms in their historical development? An exploration of this approach would be a worthwhile project for the future of digital philosophy (Heßbrüggen-Walter 2023).

Bibliographie

AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum e. V. (ed.). 2023. Begriffe der Digital Humanities. Ein diskursives Glossar. In Zeitschrift für digitale Geisteswissenschaften / Working Papers, 2. Wolfenbüttel. 10.17175/wp_2023_001

Ciula, Arianna and Øyvind Eide, Christina Marras, Patrick Sahle (eds.). 2018. *Models and Modelling between Digital and Humanities: A Multidisciplinary Perspective.* URL: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-62883, last access: 2023-07-19.

Barzen, Johanna and Frank Leymann. 2017. “Patterns as Formulas: Patterns in the Digital Humanities.” In *Proceedings of the Ninth International Conferences on Pervasive Patterns and Applications* (PATTERNS), 17-21.

Bosse, Arno. 2019. “Place”. In *Reassembling the Republic of Letters in the Digital Age: Standards, Systems, Scholarship,* eds. Howard Hotson, Thomas Wallnig, Göttingen: Göttingen University Press, 89-95.

Durlacher, Thomas. “Philosophical perspectives on computational research methods in digital history.” In *Digital History and Hermeneutics*, eds. Andreas Fickers, Juliane Tatarinov, Berlin, Boston. De Gruyter Oldenbourg, 109-127.

Hartung, Gerald. 2023. “Philosophiegeschichtsschreibung und die Idee der Geistesgeschichte – das Dilthey-Projekt.” In *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 97: 95-104.

Heßbrüggen-Walter, Stefan. 2020. “Positivismus der geistigen Gegenstände: Carnap und die Digital Humanities.” In *Book of Abstracts, DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation (DHd 2020)*. 10.5281/zenodo.4621804

Heßbrüggen-Walter, Stefan. 2023. A Philosophical View of the Digital History of Concepts: Four Theses And a Postscript. In *Digital Humanities 2023. Collaboration as Opportunity (DH2023)*, Graz, Austria. 10.5281/zenodo.8107790

Roller, Ramona. 2021. “Theory-Driven Statistics for the Digital Humanities: Presenting Pitfalls and a Practical Guide by the Example of the Reformation.” In *Journal of Cultural Analytics* 7. 10.22148/001c.57764.

Schlick, Moritz. 1938. “The Future of Philosophy”. In *Gesammelte Aufsätze 1926-1936*. Wien: Gerold & Co., 117-133.

Towards a Method for Automatic Detection of Textual Comparisons. A DH-Case Study on the Construction of “Swissness”

Aust, Robin-M.
robin-martin.aust@uni-bielefeld.de
Universität Bielefeld, Germany
ORCID: 0009-0002-2245-281X

Kababgi, Daniel
daniel.kababgi@uni-bielefeld.de
Universität Bielefeld, Germany
ORCID: 0009-0002-0990-6418

Herrmann, Berenike
berenike.herrmann@uni-bielefeld.de
Universität Bielefeld, Germany
ORCID: 0000-0002-5256-0566

Introduction

The concept 'national literature' presupposes and co-constructs distinct ‘literatures’ and invites their comparison

across co-constructed ‘nations’(Anderson, 1998). In German-Swiss literary discourse from 1850-1950, the (self-)conceptualization of ‘Swiss’ nationality and the shifting relationship to its neighboring countries are specifically challenged. Conceptualizing a corpus of texts *about* literature, our paper examines implicit textual practices and semantics of comparing (Epple et al., 2020) that evoke different practices of comparing and evaluation.

Practices of comparing relate two or more *comparata* to each other, conveying a *tertium comparationis* – e.g., when comparing the comparata ‘apples’ to ‘pears’, one conventional tertium is ‘fruits’.Textual comparisons cover a broad range between implicit and explicit forms of comparing. In our CLS research project, we compressed a scale of 5 grades of explicitness (Davy et al. 2019) to a binary opposition: we consider sentences as either explicit comparisons or implicit comparisons. We define implicit comparisons as language use that states something about *comparata* without addressing any similarity/dissimilarity relation within the utterance itself.

In this proof-of-concept study, we focus on implicit comparisons in literary histories: We assume that creating and delineating literary production as ‘national’ was enabled by the networked usage of ‘national literature’ as an overarching *tertium.* It follows that single-sentence statements about a national entity may be considered part of a vast network of comparisons spanning works about literature. Implicit comparisons rely strongly on the situative, material, and discursive context, which complicates algorithmic detection. We address this challenge by considering single utterances as part of networked implicit, intratextual, or supra-textual formations of comparing that are dynamically initiated by the construction of ‘national literature’ (Hillebrandt 2014, p. 103-4). Our vantage point for formalized identification of implicit comparisons is their evaluative dimension in the ‘national’ literary history ‘around 1900’: not only do these longer texts feature comparisons at sentence level that are likely to be implicit and evaluative.

To detect potential implicit comparisons, we combine rule-based entity detection and manual annotation with a deep learning approach to sentiment analysis: (1) identification of discrete entities as potential *comparata* based on a lexicon of nation-related entities; (2) manual annotation of the resulting sentences for *valence* (for reasons of focus, ‘explicit comparison’ sentences will not be addressed in the present paper, but will be subject of further studies). (3) we use the annotated data to train a deep learning sentiment classifier.

We utilize sentiment analysis as a marker of evaluative assessment of discourse entities: The evaluative dimension of text can be made visible by looking at the sentences’ *valence*, detected by sentiment analysis (SA), a process for quantifying the subjective and emotional undercurrent in a given text (Lei and Liu, 2021). It finds wide application in recent studies in Computational Literary Studies (Grisot and Herrmann, 2023; Salgaro, 2023; Vlachos and Husen, 2023), where many studies classify ‘ *valence*’ on a positive/neutral/negative axis (cf. Kim and Klinger, 2019). Approaches to computing sentiment range from dictionary-based ones to deep learning (Ligthear, Catal, and Tekinerdogan, 2021). By expanding the analysis to sentences, we test and expand the obtained results.

We aim to provide a first step into the operationalization of textual manifestations of comparison for a mixed-methods approach, following the typology provided in the works of the CRC1288, especially Davy et al 2019, Kramer et al 2020, Epple et al 2020. More information on the full project can be found on https://www.uni-bielefeld.de/sfb/sfb1288/projektbereiche/e06/. The annotation guidelines, code, and the data can be found here: https://github.com/DanielKababgi/Swissness_Comparison/.

The contributions of this paper are threefold: (1) we develop an annotation process for *valence* of implicit comparisons; (2) evaluate a method for detecting implicit value assessment for relevant nation-related entities; and (3) perform a pilot study into the usage of implicit comparisons in the valuation of ‘Swiss’ and other nations in a corpus of German-Swiss literary histories in between Federal state foundation and the end of the Second World War.

Data

Table 1: Current corpus of literary histories, ranging from 1861 to 1951.

Year	Autho r	Title	N Senteces	N Tokens	Place, Publisher
1861	Mörikofer	Die Schweizerische Literatur des achtzehnten Jahrhunderts	7586	172934	Leipzig, Hirzel
1866	Weber	Die poetische Nationalliteratur der deutschen Schweiz; Musterstücke aus den Dichtungen der besten schweizerischen Schriftsteller von Haller bis auf die Gegenwart, mit biographischen und kritischen Einleitungen	11035	145046	Glarus, Vogel
1892	Bächtold	Geschichte der deutschen Literatur in der Schweiz	30557	398525	Frauenfeld, Huber
1910	Jenny/Rossel	Geschichte der schweizerischen Literatur	9616	167673	Bern, Francke
1914	Frey	Schweizer Dichter	2742	55839	Leipzig, Quelle & Meyer
1918	Korrodi	Schweizerische Literaturbriefe.	804	15106	Frauenfeld, Huber
1924	Korrodi	Schweizerdichtung der Gegenwart	906	17433	Leipzig, Haessel
1924	Nadler	Der geistige Aufbau der deutschen Schweiz (1798-1848)	817	18384	Leipzig, Haessel
1933	Ermatinger	Dichtung und Geistesleben der deutschen Schweiz	15312	280703	München, Beck
1943	Korrodi	Geisteserbe der Schweiz. Schriften von Albrecht von Haller bis zur Gegenwart	5316	109865	Zürich, Rentsch
1951	Zäch	Die Dichtung der deutschen Schweiz	3426	66301	Zürich, Speer

We created a corpus consisting of eleven academic works about Swiss literature, published in German 1861-1951, as shown in Table 1.

It is important to note that while this corpus is comprehensive for the period and text type, it is not a representative sample of the 'German-Swiss literary discourse' as a whole. Any results obtained will be assertions about a set of 'literary histories' as a historically situated text type, reflecting the included authors'/publishers' perspectives, and do not assess historical change in a more general way. This study on literary histories however provides a starting point, with N=1.44 mio words and N=88,116 sentences, being situated in between a qualitative and a quantitative analysis.

Method

For the automatic detection of nation-related references, we split all texts into single sentences and use a lexical approach to select the relevant ones. To operationalize nation-related *comparata*, we created a 'whitelist' containing names of countries, cities, and synonyms. Sentences containing such an entity were selected for analysis. Out of roughly N= 8,000 filtered sentences, n= 2,000 randomized sentences were used, assuring a matching distribution between random sample and complete dataset. After dropping all false-positive matches, n=1,655 sentences remained. These were then manually annotated sentences, the rest were automatically annotated with the deep learning classifier. Fig. 1 shows the count of occurrences of sentences with nation-related entities in our dataset:

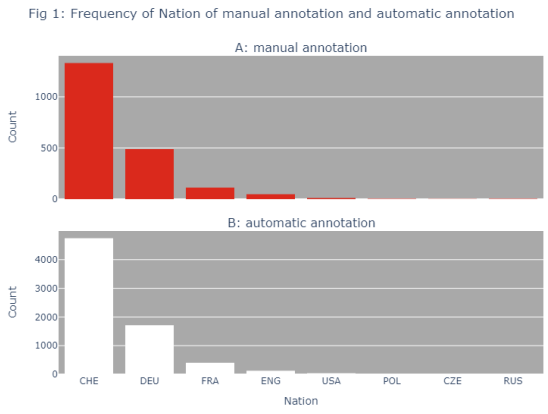


Figure 1: Occurrences of sentences with nation-related entities in (A) the dataset for manual annotation, and (B) the dataset for automatic annotation.

Since only Switzerland, Germany, and France are referenced frequently, the following analysis focuses on the evaluative assessment of those nations. An overview of the relative distribution of sentences containing nation-related entities can be seen in Fig 2:

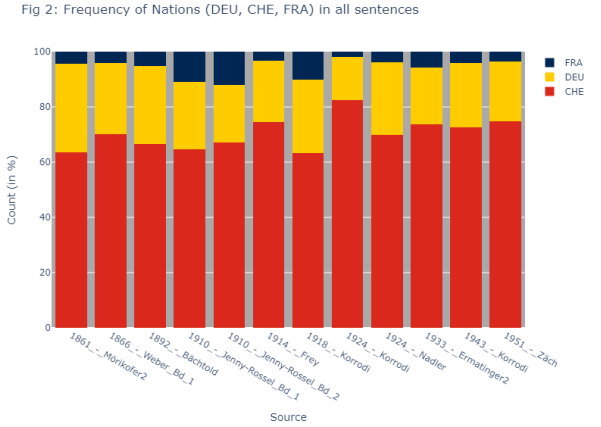


Figure 2: Relative frequency of CHE, DEU, and FRA in both the datasets for manual and automatic annotation.

As was expected from a corpus of Swiss literary histories, references to Switzerland are dominant (and over time appear to slightly rise). The proportion of references to Germany is relatively stable (but appears to slightly drop). Some oscillations are visible regarding France, but they may be quite clearly attributed to the bi-lingual two-tome publication by Jenny/Rossell, and Korrodi's 1918 book. It remains to be seen in future analyses, whether diachronic trends more generally may be observed.

Assuming that in literary histories, lexical evocation of the 'national dimension' indicates a potential comparison, statements without explicit markers of comparison (e.g., "verglichen mit", "ist schöner als") were considered as potential context-based 'implicit comparisons'. Those sentences were annotated for valence on a categorial scale, as shown in Table 2.

Table 2: Examples of different valences.

Valence	Example
+2	" Amiel [...] war entzückt, gerührt, ergriffen."
+1	" Er wirkte redlich mit, die Poesie der Engländer und der Alten zum Gemeingute der deutschen Nation zu machen."
0	" Später wandte er sich nach München."
-1	" Das Machtbedürfnis der Kantone [...] rächte sich durch Schikanen im Inneren und durch den üppigen Emportrieb nörgelnden Philistergeistes und öder Fraubaserei, die in den kleinen Verhältnissen der Schweiz immer gedeihen."
-2	" Ganz arm ist die Schweiz des siebzehnten Jahrhunderts an Werken der Prosa, die in den Bereich der Literaturgeschichte fallen würden."

The n= 2,000 randomly selected sentences were annotated by two sets of annotators, achieving an Intraclass correlation coefficient (ICC31) of 0.55, with a lower bound of 0.52 and an upper bound of 0.58 for the 95% confidence interval. The moderate reliability (add ref to Landis / Koch 1977, p. 165), which given the historical alterity and complexity of the scientific prose, is a good basis for continued analysis.

A deep learning classifier was trained on the manually annotated data. We reduced the number of categories from five to three (-1, 0, +1), corresponding to negative, neutral,

or positive sentiment. Using the gbert-large model by Deepset (Chan et al., 2020) as input for the multi-class classification head, we achieved a final F1 score of 0.5921. Further thoughts about this score are given in the discussion.

To analyze evaluative assessment of nation-related references, we used the trained model to compute the valence of each sentence contributing to the text-spanning networked implicit comparisons. We utilized the manually annotated valence data to (1) discern the evaluative assessment of 'national references' in the comparative statements, (2) to evaluate automatic sentiment analysis for previously unseen texts with the same goal.

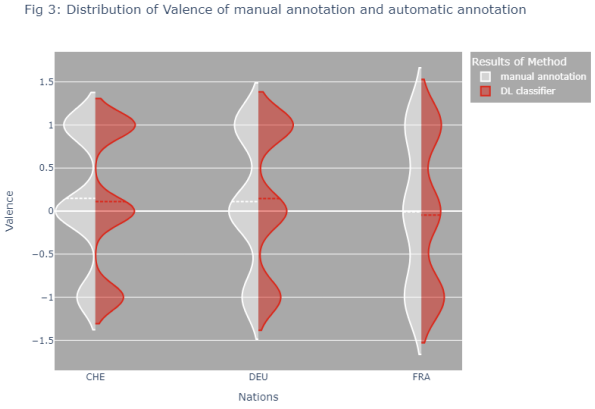


Figure 3: Distribution of valence for manual in white and automatic annotation in red for CHE, DEU, and FRA.

The distribution of valence for the manually annotated data and the results of the automatic classifier are shown in Fig 3. The results show similar distributions with some slight differences. These can be explained by the different sets and amounts of sentences in the datasets. The means of the distributions are very similar, and in both datasets Switzerland and Germany have slightly positive sentiment compared to neutral sentiment for France. To evaluate the method more formally, we use a ranked-sum Wilcox test and compare the valence distributions between the individual nations for the manually and automatically annotated data. The results are shown respectively in Fig 4 (A) and (B).

For both datasets the differences between Switzerland and Germany are not statistically significant, which corresponds to the trend depicted in Fig 3. For the manually annotated dataset neither the differences between France and the two other nations, respectively, are statistically significant. For the automatically annotated dataset, both the distributions of FRA are highly significantly different from CHE and DEU. The trend seen in the manually annotated dataset could not be replicated fully in the automatically annotated dataset.

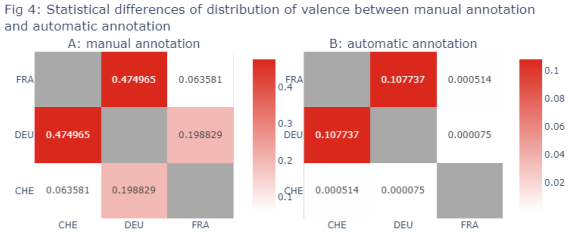


Figure 4: Statistical differences of distribution of valence between (A) manual annotation and (B) automatic annotation for CHE, DEU, and FRA.

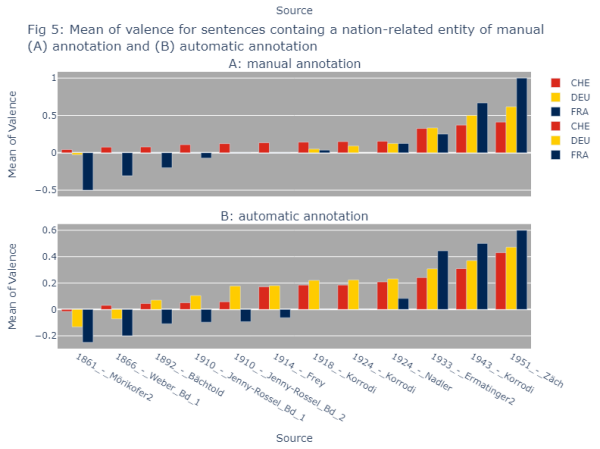


Figure 5: Mean of valence for sentences containing a nation-related entity per literary history for manual and automatic annotation.

As Fig 5 shows, positive valuation of Switzerland, Germany, and especially France is generally increasing in our corpus in both datasets. These findings may very cautiously be interpreted: The initially negative valuation of Germany and France may indicate a tendency to differentiate 'Swiss culture' from that of other countries after the Bundestaatsgründung in 1848. Negative sentiment towards French culture might be attributed to both countries' strained relationship during the 19th Century. The overall neutral and later on positive valuation towards German culture might be explained by the linguistic, historical, and cultural proximity of both countries. The rapid increase of positive sentiment towards France after 1933 can be related to the ' Geistige Landesverteidigung ' (Mooser, 1997; Jorio, 2006; Sandberg, 2007) against The Third Reich, propagating positivity towards non-German neighbors, and to the intention to strengthen Switzerland from within by integrating the German-speaking parts with its other language communities. Striking is the increasing positive valuation of Germany despite the Geistige Landesverteidigung . This might be due to individual authors' preference to seek closeness to Germany and/or The Third Reich. Both points will be further tested against a planned larger dataset.

Our approach produces insights regarding the diachronic distant reading perspective. The analyzed sentences are also viable candidates for close reading, examining comparison type and rhetorical strategies. Our method yields findings that can be assessed further. E.g., the following sentences identified by the classifier show attributions and valuations made by the different authors with regard to the different nations:

> “ Mit scharfem Verstand erkennt und faßt der Schweizer sein Ziel, wählt mit ruhiger Umsicht die Mittel zu demselben und erreicht es durch Entschlossenheit und zähe Willenskraft. ” (Mörikofer 1861, 50)

> “ Die reiche und große Natur der Schweiz regte daher vor Allem die Natur Forschung an und bildete und erzog zu allen Zeiten berühmte Natur forscher, unter denen Konrad Geßner, Scheuchzer, Haller u.s.w. nicht nur Zierden der Wissenschaft , sondern auch des ganzen Vaterlandes sind.” (Mörikofer 1861, 58)

Discussion and Outlook

It is important to note that the moderate ICC (0.55) points to some remaining issues with the annotation guidelines and the data. Some sentences prove especially difficult to classify by our annotators, due to the difficulty of the elevated style and historicity of corpus. The following sentences were annotated both as positive and negative. Depending on their context and the readers’ perspective, both interpretations can be seen as viable:

> “Auf diesem schattenlosen Voten nun trinkt man des Morgens die Ziegenmolken oder Geißschotten, wie die Schweizer sprechen, die täglich aus dem Gebirge drei Stunden weit noch ganz heiß gebracht wird, sofern es wahr ist, dass sie nicht unterwegs gewärmt werde - und bratet dabei an der Sonne, deren Strahlen nun schon wieder brennen, als könnte es hier nie Winter werden.” (Weber 1866, 15693)

> “Während sonst die Dichter jener Zeit sich den Preis der Fürsten zur Aufgabe machten, so singt der Schweizer die Eitelkeit äußerer Größe, und während jene Dichter sich mit ihren Reimen den Großen der Erde zu Füßen legen, kennt Haller keine höhere Befriedigung als das Glück seines Freundes und den treuen Freundschaftsbund mit ihm.” (Mörikofer 1861, 396)

Similar problems arise from sentences containing multiple entities and subjects. Depending on the entity focalized, sentiment can vary widely:

> “ Es gibt keinen Lyriker der Schweiz und in Deutschland nur sehr wenige, die ihn [Leuthold] an virtuoser Herrschaft über die Sprache und an Wohlklang der Verse erreichen.” (Ermatinger 1933, 81596)

Still, the method for operationalizing implicit comparisons renders promising results, pointing to a contrast in the evaluation of national comparata by language, with the German-speaking countries on the one hand and especially France on the other. Additionally, the preliminary diachronic analysis may indicate a tendency of valuating German entities more positively than Swiss ones, which is surprising and needs more analysis on qualitative and quantitative grounds.

Our approach yielded various interpretable data, yet will be refined iteratively. The identification of a large amount of nation-related comparisons proved to work. The whitelist approach for finding relevant sentences is easily implemented and expanded, but may be made more effective.

Provided more analyses, our manual and explorative approaches thus pave the road for further applying and refining the taxonomy proposed by the CRC1288. The next steps include expanding our corpus, operationalizing explicitness of comparisons, and assessing the tertia employed, enabling more fine-grained analysis focused on diachronic differences.

Bibliographie

Anderson, Benedict. 1998. *Die Erfindung der Nation. Zur Karriere eines folgenreichen Konzepts*. Berlin: Ullstein.

Chan, Branden, Stefan Schweter, and Timo Möller. 2020. German’s Next Language Model. arXiv. http://arxiv.org/abs/2010.10906 .

Davy, Ulrike, Johannes Grave, Marcus Hartner, Ralf Schneider, and Willibald Steinmetz. 2019. " *Grundbegriffe für eine Theorie des Vergleichens. Ein Zwischenbericht. Working Paper des SFB 1288, No. 3*". https://doi.org/10.4119/unibi/2939563.

Epple, Angelika, Antje Flüchter, and Thomas Müller. 2020. " *Praktiken des Vergleichens: Modi und Formationen. Ein Bericht von unterwegs. Working Paper des SFB 1288, No. 6*". https://doi.org/10.4119/unibi/2943010.

Ermatinger, Emil. 1933. *Dichtung und Geistesleben der deutschen Schweiz*. München: Beck.

Frey, Adolf. 1914. " *Schweizer Dichter"* . Leipzig: Quelle & Meyer.

Grisot, Giulia and Berenike Herrmann. 2023. "Examining the representation of landscape and its emotional value in German-Swiss fiction between 1840 and 1940". *Journal of Cultural Analytics* 8. https://doi.org/10.22148/001c.84475 .

Hillebrandt, Frank. 2014. *Soziologische Praxistheorien*. Wiesbaden: Springer Fachmedien. https://doi.org/10.1007/978-3-531-94097-7.

Im Hof, Ulrich. 1997. " *Geschichte der Schweiz "* . Stuttgart: Kohlhammer.

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 .

Jorio, Marco. 2006. " Geistige Landesverteidigung". Historisches Lexikon der Schweiz, Bd. 5, S. 163–165, accessed Dec, 2023, https://hls-dhs-dss.ch/de/articles/017426/2006-11-23/.

Kim, Evgeny, and Roman Klinger. 2019. „A Survey on Sentiment and Emotion Analysis for Computational Literary Studies“. *Zeitschrift Für Digitale Geisteswissenschaften*. https://doi.org/10.17175/2019_008 .

Kramer, Kirsten, Martin Carrier, Joris Corin Heyder, and Britta Hochkirchen. 2020. " *Vergleichen und Erzählen. Zur Verflechtung zweier Kulturtechniken. Working Paper des SFB 1288, No. 4*". https://doi.org/10.4119/unibi/2946608.

Landis, J. Richard and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33: 159–174. https://doi.org/10.2307/2529310.

Lei, Lei and Dilin Liu. 2021. *Conducting Sentiment Analysis*. Cambridge: Cambridge University Press.

Ligthart, Alexander, Cagatay Catal, and Bedir Tekinerdogan. 2021. „Systematic Reviews in Sentiment Analysis: A Tertiary Study“. *Artificial Intelligence Review* 54 (7): 4997–5053. https://doi.org/10.1007/s10462-021-09973-3 .

Mooser, Josef. 1997. „Die ‚Geistige Landesverteidigung‘ in den 1930er Jahren: Profile und Kontexte eines vielschichtigen Phänomens der schweizerischen politischen Kultur in der Zwischenkriegszeit“. *Schweizerische Zeitschrift für Geschichte* 47 (4). 685–708.

Mörikofer, Johann Caspar. 1861. D *ie Schweizerische Literatur des achtzehnten Jahrhunderts*. Leipzig: Hirzel.

Salgaro, Massimo. 2023. Stylistics, Stylometry and Sentiment Analysis in German Studies. Göttingen: V&R Unipress.

Sandberg, Beatrice. 2007. „Geistige Landesverteidigung (1933–1945)“. *Schweizer Literaturgeschichte, ed, Peter Rusterholz et al.*, 208–40. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05243-8_5 .

Vlachos, Evgenios and Kamilla Jensen Husen. 2023. Revisiting Pontoppidan: Sentiment analysis and topic modelling on ‘Eagle's Flight’. *Orbis Litterarum* 78: 441–463. https://doi.org/10.1111/oli.12406.

Weber, Robert. 1866. *Die poetische Nationalliteratur der deutschen Schweiz; Musterstücke aus den Dichtungen der besten schweizerischen Schriftsteller von Haller bis auf die Gegenwart, mit biographischen und kritischen Einleitungen.* Glarus: Vogel.

Towards Linked Stage Graph 2.0 - a Knowledge Graph based Research Resource for the Performing Arts

Tietz, Tabea
tabea.tietz@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Deutschland
ORCID: 0000-0002-1648-1684

Sack, Harald
harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Deutschland
ORCID: 0000-0001-7069-9804

Introduction

Theatre as an art form and as a social institution has been deeply embedded into our societies. Theatre is a realm that effortlessly merges the historic and the incredibly modern, continuously evolving and reinventing itself time and time again. Historical plays, which persist to this day, remain remarkably relevant, finding new life through reinterpretation on modern stages. Theatre never exists in isolation and is always embedded in the context of the respective culture and society. These contexts can be shaped by societal and political events or by technological advancements and these contexts also have shown to influence the creative possibilities and limitations on stage. Scholars from various disciplines have been examining the history and development of theatre in all its facets, including stage design, characters, costumes, and language.

A number of tools have been created with the intention to provide access to performing arts data collections. Linked Stage Graph (LSG) as one of these efforts enables the exploration of photographs and metadata of the Stuttgart State Theatre's from the 1880s until the 1940s (Tietz, 2019). The knowledge graph (KG) is accessible on the Web by means of a public SPARQL Endpoint[1] and an exploration interface[2]. While the cultural heritage community has been receiving LSG generally well, there have also been challenges which still have to be resolved for the resource to be even more useful for the cultural heritage community. For instance, the underlying ontology requires an update to provide more meaningful relations and to improve interoperability. This includes: A more accurate distinction between

the original work and the performance event and an improvement of the representation of individual roles and functions of persons active. Furthermore, Linked Stage Graph currently contains textual descriptions of the archival objects. This is difficult to integrate into exploration environments in a useful and meaningful way. An improved extraction of notable entities and their integration into the KG enables more efficient querying.

This contribution reports on the ongoing process of creating a new version of LSG as a technically improved and even more useful research resource for the culture community. Goal of this paper is to provide a strategy towards Linked Stage Graph 2.0 (Section 4.2) based on systematically extracted requirements (Section 3) and by taking into account the challenges (Section 4.1) of the data set. These contributions are not only limited to LSG and concern performing arts data in general. Therefore, this paper is furthermore an invitation to discuss current systems and data models as well as strategies to more forward in the performing arts sector from a more technical point of view.

Related Work

There are a number of systems, platforms and data models which provide access to (historical) performing arts data. Often, their entry point is focused on either a biographical, a regional or an institutional approach. LSG is based on historical data created at Stuttgart State Theatres and provided by the Baden-Württemberg National Archives.

Biographical approaches include Ipsen Stage[3], the Pina Bausch Archive (Thull, Diwisch and Marz, 2015) and Staging Beckett (McMullan et al., 2014). Furthermore, Re-Collecting Theatre History (Probst and Pinto, 2020) provides a data capturing tool and an exploration environment for theatre sciences. Approaches with a more institutional focus include the Abby Theater Platform (Bradley and Keane, 2015) and the Specialised Information Service for Performing Arts (Beck et al., 2017). Important regional approaches include the Dutch project ONSTAGE (Blom, Nijboer and Zalm, 2020), AusStage (Bollen, 2016), the Swiss Performing Arts Platform (Estermann and Julien, 2019) and OperaSampo (Ahola et al., 2023).

The above listed projects and platforms are valuable and relevant research resources in the performing arts domain. However, to the best of our current knowledge, there are no public SPARQL endpoints available for any of them. Even in cases where ontologies were designed to represent the data, they were not made publicly available for reuse. The SPA project is relevant for LSG, as it provides a comprehensive and well-documented published data model. Linked Stage Graph is accessible by means of a SPARQL endpoint and all data and development progress is open and documented on GitHub[4].

Requirements

A requirement analysis has been carried out to improve LSG systematically. The requirements were extracted from scientific literature in the domain of the performing arts. All requirements can be traced down to the literature they were extracted from and the full overview is available on GitHub. This analysis is an ongoing process and will be extended as part of iterative ontology development.

REQ1: Context. Provide as much context as possible. No entity in the performing arts exists on its own (e.g. archival object, performance, person, prop) and should be viewed in its context to other entities.

REQ2: Perspective. Provide various perspectives on performing arts data for exploration to enable a holistic view.

REQ3: Interoperability. Enable the interconnection between disciplines, data sets, archives, performing arts institutes as well as regional and international efforts.

REQ4: Persons. All persons on, behind and in front of the stage of a performance and their roles and functions are relevant for research.

REQ5: Change. Performing arts are dynamic and the change over time should be represented in terms of persons, occupations, stage design, etc.

REQ6: Events. Performing arts data is often event-based. It should be distinguished between an original work, a production and the performance as an event.

REQ7: Stage Elements. Objects on stage should be captured. If possible, the meaning of an object on stage should be represented as well.

REQ8: Querying. A data model that represents performing arts data should be as lightweight as possible to enable intuitive querying.

REQ9: Provenance. It has to be possible to verify and track research results, e.g. biographical data has to be linked to their data source.

REQ10: Data Quality. The quality of the data used in performing arts research has to be clear and should be quantifiable.

These requirements are being carefully taken into account throughout the development process. However, not all requirements can be fully met due to the sparsity of the available metadata.

Towards Linked Stage Graph 2.0

This chapter discusses additional challenges within LSG and provides a strategy with concrete tools and standards to overcome them and meet the requirements.

Current Challenges

In the following, the most relevant challenges within the LSG data are described.

CH1: Archival Structure. The original data structure is based on the folders in the archive before digitization and not intuitive for web-based exploration. These folders reveal information about the carrier type of the photographs in the data set (e.g. glass plate), but also the performance types (e.g. ballet).

CH2: Semi-structured Data. The titles and descriptions of the archival objects contain many unstructured information about the performances, dates, and persons involved.

CH3: Performance Categories. Performances are categorized by genre (e.g. opera, ballet) and by type (e.g. premiere, repertoire) in a semi-structured way. Between these concepts relations exist (opera, romantic opera) but they are not reflected in the original data set. So far, no reference lists could be determined for a unique identification.

CH4: Heterogeneity. Not every archival object in the data set is part of a performance, e.g. outdoor photographs of theatre buildings.

CH5: Metadata Sparsity. Performance data often lack metadata, e.g. actors are often not listed and performance dates are sometimes missing.

Even though these challenges were observed within LSG, they are not limited to this data set only and are generalizable to historical German archival data.

Strategy

According to the requirements, performing arts research data should be interconnected, and interoperable with as much context provided as possible, and should represent events and change meaningfully. To fulfil these needs, KGs present a state-of-the-art solution. The following sections discuss the intended strategy in order to fulfil the listed requirements as part of a KG-based solution.

Data Model

To the best of our knowledge, the Swiss Performing Arts (SPA) data model is the most semantically expressive and best documented model that is available in the performing arts domain (Estermann and Julien, 2019). It is event based and emphasizes the importance of differentiating the original work and the performance event (REQ6) by reusing CIDOC, FRBR and FRBRoo. SPA uses RiC-O to represent the archival structure, which will be utilized to cope with CH1 and REQ9. However, the SPA is incredibly complex (REQ8) and assumes rich metadata, which is not present within LSG. For instance, there is no property between the work[5] and the performance event[6] itself and all persons and their functions (e.g stage designers) are connected with the production. However, in Linked Stage Graph it is not clear which production a performance event belongs to. Therefore, all persons are linked to the respective performance event. The model further connects the Performing Arts Production with a Performance Plan[7] and Performance Work[8] which are not available in the Linked Stage Graph metadata. Therefore, the model cannot be reused entirely, but will partially be utilized. Modeling challenges and strategies are furthermore discussed in Tietz et al. (2023).

Semi-structured Data

To cope with CH2, Named Entity Recognition and Linking will be conducted to extract mentions of persons and their functions, works, and dates from semi-structured descriptions of the archival objects. A mapping with existing data sources like GND an Wikidata will improve interoperability (REQ3) and enables to provide more context to the data by means of enrichment (REQ1). Furthermore, NIF2 (Hellmann et al., 2013) can be used to provide confidence levels (REQ10).

Perspective

As mentioned in CH5 for many archival objects within Linked Stage Graph no rich metadata are available. However, to meet REQ2 and provide a holistic data exploration environment, sufficient metadata is crucial. Therefore, the photographs within the collections are analyzed by means of state-of-the-art object detection and caption generation (REQ7). Previous experiments (Tietz et al., 2020) have revealed the challenges within this task. To integrate the image analysis results into the KG the Web Annotation Ontology[9] can be applied. As a consequence, means of explorations can exceed the current timeline-based approach and also allow to search within the photographs.

Categorizing Performances

The categorization of performances by type and genre as mentioned in CH3 is crucial for data exploration. However, to the best of our knowledge no reference lists exists which are widely accepted by the performing arts community to relate individual concepts (e.g. opera, comic opera, comedy, romantic opera) to each other. However, the Art and Architecture Thesaurus[10] can be reused partially for this task. Furthermore, a mapping with Wikidata entities is utilized.

Context and Interoperability

Publishing performing arts data by means of a KG and connecting individual entities, e.g. persons, organizations, events to existing sources like the GND or Wikidata also provides the opportunity to enrich the existing data set with further context (REQ1). As mentioned, the performing arts are also embedded into societal changes and more context enables to answer questions like: Which theatres performed plays by Bertolt Brecht during the 1930s?. Furthermore, the KG based approach allows to easily connect the resources with further research data, e.g. via NFDI4Culture. Having

registered LSG in the Culture Information Portal[11] it is possible to find useful connections to other data sets by means of federation.

This chapter outlined current challenges and means to progress and create a more scientifically and technically correct, interconnected, and open resource by reusing open standards, allowing connections with further performing arts related resources and thus, increasing its value.

Conclusion

This paper contributes a road map towards Linked Stage Graph 2.0 as a KG-based performing arts resource, including a requirement analysis as well as a discussion of ongoing challenges and a future strategy. This road map is not intended to be the sole truth towards an open and interoperable resource in the performing arts domain. Instead, it serves as a means to progress in the field by leveraging state-of-the-art technologies and standards. This contribution is furthermore an invitation to discuss use cases, data models and exploration environments with the cultural heritage community. The development process of Linked Stage Graph is transparent, with ongoing reporting of lessons learned. In addition, all data, ontologies, requirements created along the way are being published on GitHub.

Fußnoten

1. https://slod.fiz-karlsruhe.de/sparql
2. https://slod.fiz-karlsruhe.de/vikus
3. https://ibsenstage.hf.uio.no/
4. https://github.com/ISE-FIZKarlsruhe/LinkedStage-Graph
5. SPA-E2: https://www.iflastandards.info/fr/frbr/frbroo#F1
6. SPA-E9/10: https://www.iflastandards.info/fr/frbr/frbroo#F31
7. SPA-E8: https://www.iflastandards.info/fr/frbr/frbroo#F25
8. SPA-E7: https://www.iflastandards.info/fr/frbr/frbroo#F20
9. https://www.w3.org/TR/annotation-vocab/
10. https://www.getty.edu/research/tools/vocabularies/aat/
11. https://nfdi4culture.de/resource/E3590/about.html

Bibliographie

Ahola, Annastiina, Eero Hyvönen, Heikki Rantala, and Anne Kauppala. 2023. *Publishing and Studying Historical Opera and Music Theatre Performances on the Semantic Web: Case OperaSampo 1830-1960.*" In Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH), co-located with ISWC 2023. CEUR WS Vol-3540.

Beck, Julia, Michael Büchner, Stephan Bartholmei, and Marko Knepper. 2017. *Performing Entity Facts: The Specialised Information Service Performing Arts*. Datenbank Spektrum 17 (1): 47–52.

Blom, Frans R. E., Harm Nijboer, and Rob van der Zalm. 2020. *ONSTAGE, the Online Data System of Theatre in Amsterdam from the Golden Age to Today*. Research Data Journal for the Humanities and Social Sciences 5 (2): 27–40.

Bollen, Jonathan. 2016. *Data Models for Theatre Research: People, Places, and Performance*. Theatre Journal, 615–632.

Bradley, Martin, and Aisling Keane. 2015. *The Abbey Theatre Digitization Project in NUI Galway.* New Review of Information Networking 20 (1-2): 35–47.

Estermann, Beat, and Frédéric Julien. 2019. *A Linked Digital Future for the Performing Arts: Leveraging Synergies along the Value Chain*. In Canadian Arts Presenting Association (CAPACOA) in cooperation with the Bern University of Applied Sciences.

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. *Integrating NLP using linked data*. In 12th International Semantic Web Conference, Sydney, NSW, Australia, Proceedings, Part II 12, 98–113. Springer.

McMullan, Anna, Trish McTighe, David Pattie, and David Tucker. 2014. *Staging Beckett: constructing histories of performance*. Journal of Beckett Studies 23 (1): 11–33.

Probst, Nora, and Vito Pinto. 2020. *Re-Collecting Theatre History, Theaterhistoriografische Nachlassforschung mit Verfahren der Digital Humanities*. In Neue Methoden der Theaterwissenschaft, edited by Benjamin Wihstutz and Benjamin Hoesch, 157–179. transcript.

Thull, Bernhard, Kerstin Diwisch, and Vera Marz. 2015. Linked Data im digitalen Tanzarchiv der Pina Bausch Foundation. In Corporate Semantic Web: Wie semantische Anwendungen in Unternehmen Nutzen stiften, 259–275.

Tietz, Tabea, Oleksandra Bruns, and Harald Sack. 2023. *A Data Model for Linked Stage Graph and the Historical Performing Arts Domain*. In Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH), co-located with ISWC 2023. CEUR WS Vol-3540.

Tietz, Tabea, Jörg Waitelonis, Mehwish Alam, and Harald Sack. 2020. *Knowledge Graph based Analysis and Exploration of Historical Theatre Photographs*. In Qurator 2020.

Tietz, Tabea, Jörg Waitelonis, Kanran Zhou, Paul Felgentreff, Nils Meyer, Andreas Weber, and Harald Sack. 2019. *Linked Stage Graph.* In SEMANTICS Posters&Demos.

Verknüpfungen und Kontextualisierung durch Annotationen - Forschen mit multimodalen Daten.

Kröber, Cindy
cindy.kroeber@uni-jena.de
FSU Jena, Deutschland

Bruschke, Jonas
jonas.bruschke@uni-wuerzburg.de
Uni Würzburg, Deutschland

Utescher, Ronja
ronja.utescher@uni-jena.de
FSU Jena, Deutschland

Maiwald, Ferdinand
ferdinand.maiwald@uni-jena.de
FSU Jena, Deutschland

Pattee, Aaron
aaron.pattee@lmu.de
LMU München, Deutschland

Abstract

In der kunsthistorischen Forschung ist die eingehende und umfassende Auseinandersetzung mit Quellenmaterial verschiedenster Natur ein aufwendiger und anspruchsvoller Schritt, um ein ganzheitliches Verständnis des Forschungsgegenstandes zu erlangen. Dabei ist neben dem Inhalt der Quellen auch ihre Verlässlichkeit und ursprüngliche Intention relevant. Annotationen können dabei behilflich sein, verhältnismäßig schnell einen Überblick über den Inhalt einer Quelle zu bekommen, sowie Widersprüche und Zusammenhänge hervorzuheben.

Dieser Beitrag zeigt auf, welches Potential die Verknüpfung von multimodalen Daten mittels Annotationen bietet. Dazu wird skizziert welche Arbeitsschritte für die Annotierung von 3D-Modellen, Bildern und Texten nötig sind und welche Möglichkeiten der Automatisierung sich abzeichnen. Verschiedene Anwendungsszenarien für die Forschungsplattform 4D Browser verdeutlichen Einsatzmöglichkeiten für die Forschung in der Kunst- und Architekturgeschichte.

Einleitung

Texte wie auch Fotografien und Bilder sind wichtiges Quellenmaterial für die geschichtswissenschaftliche Forschung und bilden die Grundlage für themen- und theorieorientierte Untersuchungen u.a. in der Architekturwissenschaft, der Kunstgeschichte und den Kulturwissenschaften. Die Quellen ermöglichen eine (digitale) Rekonstruktion oder Untersuchung verschiedener Gebäude, ihrer Baugeschichte und möglicher Effekte auf die Stadtentwicklung. Um ein Gebäude umfassend zu analysieren, ist sehr viel Quellenmaterial erforderlich. Die Suche nach Bildern und Texten, aber auch deren Kontextualisierung und Auswertung ist herausfordernd (Beaudoin and Brady 2011). Einerseits gibt es eine riesige Menge an online verfügbaren Daten, andererseits ist eine Filterung der Ergebnisse in der Regel nicht zufriedenstellend und wird teilweise sogar vermieden.

Im BMBF-geförderten Projekt HistKI wird an einer Lösung gearbeitet, die unterschiedliche Daten in einer Forschungsplattform namens 4D Browser (https://4d-browser.urbanhistory4d.org/) verknüpft und somit die Verarbeitung von multimodalen Daten für die Forschung unterstützt. Die prototypische Webanwendung bietet eine Vielzahl an Funktionalitäten und wird stetig weiterentwickelt (Dewitz et al. 2019). Historische Bild- und Textquellen sind in einem virtuellen 3D Stadtmodell von Dresden verortet (siehe Abbildung 1). Eine Zeitschiene kann neben einer Suchleiste für die Filterung der Quellen genutzt werden. Die Daten können neben der altbekannten Metadatenabfrage auch durch die Projektion auf das 3D-Modell gefiltert werden. Somit können Bilder von Gebäuden gefunden werden, ohne dass der Objektname Bestandteil der Metadaten ist. Zusätzlich stehen Visualisierungswerkzeuge für die Analyse von Kamerastandpunkten zur Verfügung, die Forschung im Bereich der Stadtentwicklung und -wahrnehmung unterstützen.

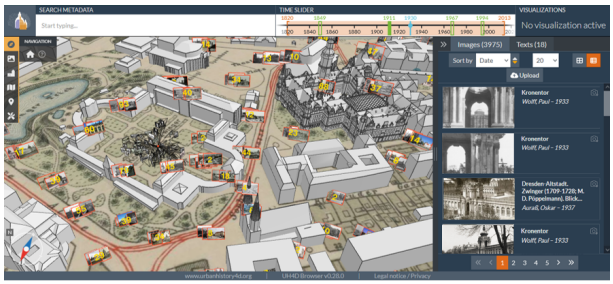


Abbildung 1. Benutzeroberfläche des 4D Browsers mit räumlich und zeitlichen verorteten Fotos in einem 3D-Stadtmodell.

Annotieren der Daten

Nachdem in einem ersten Schritt in der Forschungsplattform 4D Browser Bilder, 3D Modelle und Texte zur Ver-

fügung gestellt wurden, sollen diese Daten nun im Sinn der Multimodalität miteinander verknüpft werden. Zur Verknüpfung der drei Datentypen werden Wörter bzw. Begriffe wie architektonische Elemente und Bezeichnungen aus der Stadt- und Architekturgeschichte als Label bzw. Annotation verwendet. Im Bereich der digitalen Geisteswissenschaften ist der Getty Art & Architecture Thesaurus (AAT) etabliert und bietet eine Hierarchie für architektonische Elemente (Baca and Gill 2015). Des Weiteren wird auch Wikidata mit einer Vielzahl an Entitäten und Klassen sowie entsprechenden semantischen Beziehungen untereinander für u.a. Kunst und architektonische Elemente zunehmend im Bereich des kulturellen Erbes verwendet (Schmidt, Thiery, and Trognitz 2022). Relevante Begriffe wurden mit Hilfe einer Textanalyse von Beiträgen zum Dresdner Zwinger bzw. barocker Architektur, Baugeschichte und Bauforschung identifiziert (Utescher et al. 2022). Schließlich konnte eine Liste mit ca. 400 Einträgen - darunter 140 Architekturelemente - zusammengestellt werden, die sowohl AAT- wie auch Wikidata IDs enthält.

Texte

Bisher stehen mehrere Wikipedia-Artikel, wissenschaftliche Publikationen und populärwissenschaftliche Beiträge im 4D Browser zur Verfügung. Teilweise können Dokumente mit Hilfe eines PDF-Viewers betrachtet werden, der auch weitere Werkzeuge bereitstellt. Einige Artikel wurden bereits annotiert. Für diese steht zusätzlich ein ‚Plain Text' bereit, in dem die Annotationen eingeblendet werden können (siehe Abbildung 2). Die Annotierung von Text wurde mit Hilfe von KI-Ansätzen realisiert. Für semantische Textannotationen basierend auf der Identifikation von Wortähnlichkeiten kommt fastText zu Einsatz (Bruschke et al. 2023). In den Texten werden Wörter oder Wortgruppen aus der zusammengestellten Begriffsliste annotiert sowie Begrifflichkeiten mit einer hohen semantischen Ähnlichkeit. Named Entity Recognition (spaCy's NER) als Transformer-basiertes Model kommt zur Eigennamenerkennung zum Einsatz. So können Personen und Orte im Text automatisch annotiert werden.

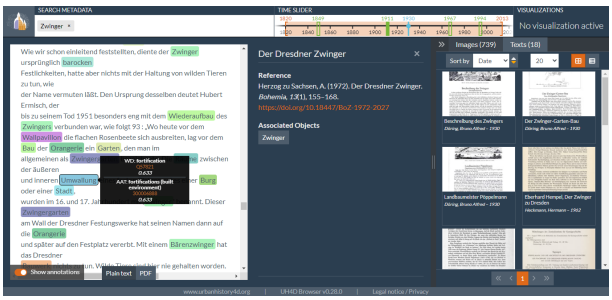


Abbildung 2. Beispiel eines annotierten Textes im 4D Browser.

3D Modelle

Im Projekt wurde für Testzwecke ein detailliertes 3D-Modell des Kronentors vom Dresdner Zwinger in einer Modellierungsumgebung manuell annotiert. In diesem Modell sind die meisten architektonischen Elemente separate Objekte, die neben der Bezeichnung des Elements noch die entsprechende AAT- oder Wikidata-ID tragen. Einige Objekte sind gruppiert und bilden eine Einheit im Sinn einer Hierarchie von Elementen.

Fotografien

Bisher wurden im 4D Browser ca. 4000 meist historische Fotografien von Dresden manuell oder mittels eines semiautomatischen Ansatzes verortet (Maiwald 2022, Maiwald et al. 2023). Neu hinzugekommen ist ein kleiner, manuell segmentierter Bilddatensatz (siehe Abbildung 3). Die Segmentierung und Annotierung erfolgte mit Label Studio (https://labelstud.io/), einer Anwendung zur Beschriftung und Annotation von Daten, welches ein standardisiertes Ausgabeformat bietet. Jede Benennung folgt dem gleichen Schema wie bei den 3D-Modelle und enthält sowohl AAT- als auch Wikidata-IDs. Eine automatische Übertragung der Annotationen vom 3D-Modell auf die verorteten Fotografien ist geplant, allerdings erreichten bisherige Ansätze noch nicht den angestrebten Grad an Zuverlässigkeit bzw. Genauigkeit.

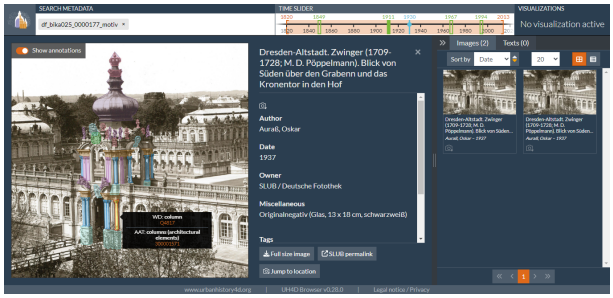


Abbildung 3. Beispiel eines segmentierten und annotierten Fotos vom Dresdner Zwinger im 4D Browser. Foto: Oskar Auraß, https://www.deutschefotothek.de/documents/obj/72018745

Mit den neuen Entwicklungen in Bereich der KI können viele der bisher manuell bearbeiteten Schritte automatisiert werden. Jedoch existierte bisher kein Trainingsdatensatz, der die domänenspezifischen architektonischen Elemente enthält, die für die Forschungsszenarien relevant sind. Daher sind unsere annotierten Daten Teil eines öffentlich-zugänglichen Datensatzes und sollen perspektivisch als Trainingsdatensatz genutzt werden (Bruschke et al. 2023).

Verknüpfung der verschiedenen Daten

Neben einer Visualisierung der Annotationen in Text, 3D und Bild bietet der 4D Browser noch eine Übersicht aller Annotationen eines Objektes in einer Wordcloud (siehe Abbildung 4). Wordclouds gibt es für alle drei Datentypen.

Die Zahl nach dem Begriff gibt an, wie häufig eine Annotation im Dokument bzw. 3D-Modell vorkommt. Somit bietet die Wordcloud indirekt einen Überblick über Themen eines Dokuments und erlaubt eine schnelle Einschätzung der Relevanz eines Dokuments für die eigene Arbeit. Die Begriffe in der Wordcloud wie auch in den einzelnen Dokumenten sind klickbar und bieten eine externe Verlinkung zu den AAT- und Wikidata-Einträgen sowie eine Interne Verlinkung, um weitere passende Inhalte im 4D Browser zu finden.

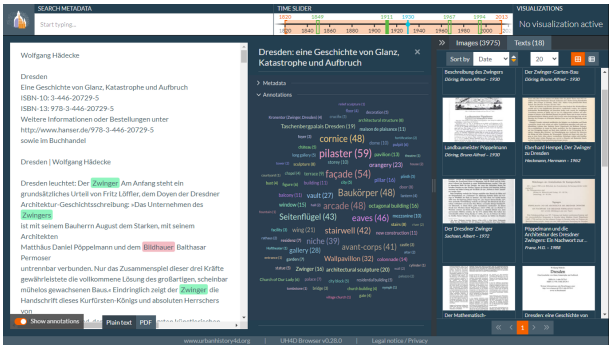


Abbildung 4. Beispiel einer Wordcloud für einen Beispieltext. Die Zahl in Klammern gibt an, wie häufig eine Annotation im Text vorkommt.

Neuerungen des 4D Browsers

Der 4D Browser ist eine Anwendung, die seit 2016 stetig weiterentwickelt und angepasst wird. Neben dem Einbringen von Annotationen wird derzeit an einem Werkzeug zur automatischen Detektion von baulichen Veränderungen in Bildern gearbeitet. Diese stetige Weiterentwicklung ist eng mit Nutzertests und Nutzerstudien verknüpft (Kröber et al. 2021). Zwar sind auch regelmäßig Design und Interface Teil der Tests, aktuell liegt der Fokus der Nutzertests jedoch auf verschiedenen Forschungsszenarien. Auch wenn der 4D Browser mit dem limitierten Datenumfang keine komplette Recherche ermöglicht, unterstützt er das Erlangen eines ersten Eindrucks mit relevanten Themen sowie die Planung nächster Schritte mit Hilfe weiterer Schlagwörter, Autorennamen oder Quellen. Insbesondere die Nutzung von Annotationen als eine Art Inhaltsübersicht unterstützt eine schnelle Evaluation einer Quelle. Auch haben bestimmte Ansätze aus der Computerlinguistik das Potential Quellenmaterial auf Verlässlichkeit zu überprüfen, da eine Verarbeitung und Analyse vieler Texte Diskrepanzen zwischen Quellen aufdecken kann oder die Einschätzung der Glaubwürdigkeit eines Autors zulässt.

Forschungsszenarien

Der 4D Browser bietet eine Vielzahl an Funktionen für die Erforschung und Visualisierung mit Fotografien und Texten. Die folgenden Szenarien sollen das Potential der Werkzeuge verdeutlichen.

Bildüberlagerung mit dem 3D-Modell

Der 4D Browser kann sehr gut für die Identifikation von Gebäuden auf einem Foto genutzt werden. Dazu steht die Funktion ‚Jump to Location' zur Verfügung und ermöglicht es an den Kamerastandpunkt eines Fotos im 3D-Modell zu springen. Durch die Überlagerung mit dem 3D-Modell können mittels Mouse-Over alle Objektnamen angezeigt werden. In der überlagerten Ansicht kann man sich mit Hilfe von Pfeilen am Bildrand (rechts, links, oben, unten) zum nächstgelegenen Foto bewegen und so genauer wahrnehmen, was in einem Bereich abgebildet wurde und sich schnell z.B. einen 360°-Überblick von einem Gebäude oder Platz verschaffen.

Die Überlagerung von 3D-Modell und Foto bietet weiter die Möglichkeit, Diskrepanzen zwischen Foto und Objekt oder Raum zu erkennen. So können retuschierte Bilder wie z.B. Postkarten oder auch Malerei analysiert werden, in denen die Perspektive verändert wurde oder bestimmte Elemente verändert wurden.

Zeitleiste und Bauliche Veränderungen

Mittels der Zeitleisten-Funktionalität können Fotos eines Objektes von verschiedenen Zeitpunkten recherchiert werden. Verknüpfte Texte bieten Hinweise zu baulichen Veränderungen durch Restaurierung oder Umbau und können somit der Verifikation dienen. Punktwolken, generiert aus den historischen Fotografien, können auch automatisch verglichen werden und somit Umgestaltungen offenlegen. So kann chronologisch zu einem Gebäude recherchiert werden, um z.B. die Biografie des Objektes zu rekonstruieren.

Visualisierungen

Der 4D Browser bietet verschiedene Visualisierungen, um die Verteilung der Fotografien in einem Gebiet analysieren zu können. In einer Heatmap ist dies am simpelsten gelöst und zeigt farbkodiert an, an welchem Stellen die meisten Fotos aufgenommen wurden. Weitere Visualisierungen wie ein ‚Radial Fan' oder ‚Particles' ähnlich einer Strömungsdarstellung beziehen auch die Aufnahmerichtung in die Visualisierung mit ein. So kann man erkennen, welche Fassade eines Gebäudes am häufigsten fotografiert wurde bzw. ob ein Gebäude eine ‚Schokoladenseite' besessen hat. Aber auch Wirkungsgebiete eines Fotografen können so visualisiert und analysiert werden.

Recherche mit verknüpften multimodalen Daten

Während der Entwicklung des 4D Browsers und seiner Werkzeuge wurde vermehrt auf die Untersuchung von Ar-

chitektur und architektonischen Elementen eingegangen. Bei der Untersuchung eines bestimmten architektonischen Elementes wie einer Skulptur stellen sich folgende Fragen: Wo kann man die Skulptur finden? Wer hat sie geschaffen und wann? Was oder wer ist abgebildet? Warum wurden Motiv und Standort gewählt? Wie wurde oder wird die Skulptur wahrgenommen und wie wirkt sie im Kontext des Gebäudes und der Umgebung? Gibt es ähnliche oder andere Skulpturen, die es wert sind, untersucht zu werden, z.B. innerhalb des Gebäudes, mit einem ähnlichen Motiv, oder von demselben Schöpfer? Ist die aktuelle Skulptur das Original oder eine Kopie? Stand sie schon immer an diesem Ort oder wurde sie versetzt? Ausgangspunkt für die Untersuchung ist das 3D-Modell an der Position der Skulptur. Verortete Fotos geben Hinweise zum Aussehen der Figur. Relevante Texte automatisch verknüpft durch Annotationen können helfen weitere Informationen zu finden. Bestimmte Textpassagen erlauben eine direkte Zuordnung zu Bildern, da sie detaillierte Beschreibungen zu architektonischen Elementen enthalten und zusätzliche Angaben zur Position wie nördlich, rechts oder unten. Mit Hilfe von textlichen Quellen lassen sich meist weitere Schlagwörter oder Quellen finden.

Der Vergleich ist eine unglaublich wichtige Methode in der kunsthistorischen Forschung. Ein Ziel kann sein, die Entwicklung eines Künstlers anhand seiner Werke zu untersuchen oder die Inspiration für ein bestimmtes Werk zu erforschen. In vielen Fällen liefert die Biographie oder das Werksverzeichnis eines Künstlers oder Architekten die notwendigen Namen und Orte, welche in den Texten im 4D Browser annotiert sind und somit schnell erschlossen und zeitlich eingeordnet werden können. Eine Recherche nach relevanten Informationen zu einer Person kommt mit dem 4D Browser einer ganzheitlichen Erschließung der Thematik nahe, da die Quellen miteinander verknüpft sind und man sich von Information zu Information weiterbewegt. Orte und Objekte regen zu weiterer Erkundung an oder bieten neue Schlüsselwörter. Texte enthalten Beschreibungen, die über Annotationen mit Bildern verknüpft sind, welche wiederum visuell verglichen werden können. Da multimodale Daten im 4D-Browser verknüpft sind, wäre die Recherche schneller und umfassender im Sinne einer ganzheitlichen Erschließung.

Fazit

Der 4D Browser wird seit 2016 stetig weiterentwickelt. Die Verknüpfung der verschiedenen Datentypen durch Annotationen eröffnet neue Forschungs- und Analyseansätze und ist ein großer Gewinn für die Forschungsplattform. Um einen richtigen Mehrwert für Wissenschaftler darzustellen, bedarf es aber eines sehr großen Datensatzes mit automatisch annotierten Daten von hoher Qualität und Zuverlässigkeit. Dann können neue Filter- und Analysewerkzeuge helfen, den Daten weitere Erkenntnisse zu entlocken. Dieser Werkstattbericht bietet dafür einen Einblick in die aktuellen Möglichkeiten und Entwicklungen.

Bibliographie

Baca, Murtha, and Melissa Gill. 2015. "Encoding multilingual knowledge systems in the digital age: the getty vocabularies." *NASKO*:41-63.

Beaudoin, Joan E., and Jessica Evans Brady. 2011. "Finding Visual Information: A Study of Image Resources Used by Archaeologists, Architects, Art Historians, and Artists." *Art Documentation: Journal of the Art Libraries Society of North America* 30 (2):24-36.

Bruschke, J, C Kröber, F Maiwald, R Utescher, and A Pattee. 2023. "Introducing a Multimodal Dataset for The Research of Architectural Elements." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 48:325-331.

Dewitz, Leyla, Cindy Kröber, Heike Messemer, Ferdinand Maiwald, Sander Münster, Jonas Bruschke, and Florian niebling. 2019. "Historical Photos and Visualizations: Potential for Research." CIPA 2019, Ávila.

Kröber, Cindy, Katharina Hammel, Cornelia Schade, Nicole Filz, and Leyla Dewitz. 2021. "User Involvement for Application Development: Methods, Opportunities and Experiences from Three Different Academic Projects." Workshop on Research and Education in Urban History in the Age of Digital Libraries.

Maiwald, Ferdinand. 2022. "A window to the past through modern urban environments." Phd, Technische Universität Dresden.

Maiwald, Ferdinand, Jonas Bruschke, Danilo Schneider, Markus Wacker, and Florian Niebling. 2023. "Giving Historical Photographs a New Perspective: Introducing Camera Orientation Parameters as New Metadata in a Large-Scale 4D Application." *Remote Sensing* 15 (7):1879.

Schmidt, Sophie C, Florian Thiery, and Martina Trognitz. 2022. "Practices of linked open data in archaeology and their realisation in Wikidata." *Digital* 2 (3):333-364.

Utescher, Ronja, Aaron Patee, Ferdinand Maiwald, Jonas Bruschke, Stephan Hoppe, Sander Münster, Florian Niebling, and Sina Zarrieß. 2022. "Exploring Naming Inventories for Architectural Elements for Use in Multi-modal Machine Learning Applications." Workshop on Computational Methods in the Humanities 2022.

Vertrauen in die Wirklichkeit AI, Trust und Reliability in den Digital Humanities

Kurz, Susanne
susanne.kurz@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-2824-1485

Eide, Øyvind
oeide@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0002-7766-6287

Digitale Objekte

Eine der wesentlichen Aufgaben der geisteswissenschaftlichen Forschung ist die Bewahrung des Kulturerbes und damit die Beschreibung, Interpretation und Kontextualisierung von Objekten des Kulturerbes. Kulturerbe wird hier im Sinne der UNESCO als Summe von materiellen und immateriellen Kulturgütern verstanden, die Zeugnisse der menschlichen Schaffens- und Schöpfungsfähigkeit zu einer bestimmten Zeit an einem bestimmten Ort darstellen. Objekte des Kulturerbes bilden somit einen wesentlichen Teil des Fundaments der Forschung über die unterschiedlichen methodologischen Ansätze, Verfahrensweisen und Fragestellungen in den verschiedenen Disziplinen der Geisteswissenschaften hinweg.

Moderne Objekte des Kulturerbes liegen häufig originär in digitaler Form vor (born digital objects)[1], während historische Kulturerbe-Objekte als physische Objekte und zunehmend zusätzlich als deren digitalisiertes Abbild (Digitalisat) verfügbar sind. Ist die Digitalisierung der Original-Objekte nach geeigneten Qualitätskriterien durchgeführt worden, können, je nach Forschungsfrage, auch die Digitalisate die Basis für Forschung sein. Zudem stehen Original-Objekte Forschenden häufig aus verschiedensten Gründen nicht zur Verfügung und in der Regel weisen die digitalen Objekte bei tiefer Erschließung auch einen Mehrwert auf.

Dieses Paper konzentriert sich auf die digitalen Objekte, die ein digitales Abbild eines ehemals oder aktuell vorhandenen Original-Objekts[2] darstellen.

Stellen solche Objekte die Grundlage für Forschung dar, vertrauen die Forschenden den für die Digitalisierung verantwortlichen Institutionen, dass das digitale Objekt ein unverfälschtes Abbild auf der Grundlage der aktuellen technischen Möglichkeiten eines tatsächlich existierenden Original-Objektes darstellt. Durch den Einsatz verschiedener technischer und workflowbasierter Verfahren kann eine Institution in der Regel die für unterschiedliche Forschungsziele erforderlichen Qualitäten erzeugen und gleichzeitig besteht im Allgemeinen ein vorbehaltloses Vertrauen seitens der Forschenden in die Authentizität des digitalen Objektes.

Nichtsdestotrotz muss und wird eine gewisse Fehlerrate akzeptiert. Keine Institution kann zu 100% sicherstellen, dass es bei den angewendeten Transformationsverfahren zum digitalen Objekt zu keinen relevanten Fehlern und ggf. problematischen Veränderungen gekommen ist.

Neben solchen technischen Fehlern stellen bewusst manipulierte digitale Objekte eine weitere mögliche Fehlerquelle dar. Manipulationen können von so hoher Güte sein, dass diese nur mit aufwendigen Verfahren zu erkennen sind. Das Erstellen von solchen hochwertigen Manipulationen ist aber gleichermaßen aufwendig und somit ist die Anzahl der zu erwartenden manipulierten digitalen Objekte eher gering und fällt in den Bereich der akzeptierten Fehlerrate.

Artifizielle digitale Objekte

Mit der breiten Verfügbarkeit von unterschiedlichen Softwaresystemen, die auf der Basis von KI-Algorithmen digitale Objekte mit scheinbar authentischen, tatsächlich aber fiktiven[3] Inhalten generieren können, entsteht ein neuartiges Problem für die moderne geisteswissenschaftliche Forschung. Forschende müssen die Möglichkeit in Erwägung ziehen, dass vorliegende digitale Objekte nicht aus vertrauenswürdigen Digitalisierungsprozessen hervorgegangen sind, sondern von entsprechenden Softwaresystemen mit artifiziellen Inhalten generiert wurden. Diese generierten digitalen Objekte weisen keinerlei Manipulationen auf, da es sich nicht um ein manipuliertes Abbild eines realen Original-Objektes handelt, sondern um ein generiertes Abbild eines „fiktiven Objektes“[4]. Da es keinen Transformationsprozess analog-digital gab, sind sie als Ganzes Fälschungen und können alle derzeit anerkannten Sicherheitskriterien erfüllen. Einzig die Zuordnung der digitalen Objekte zu Realien ist nicht möglich. Diese Objekte werden im Folgenden als artifizielle digitale Objekte bezeichnet[5].

Problematisch wird dies durch die hohe Skalierbarkeit der generierenden Verfahren. Die geschickte Nutzung von entsprechenden Softwaresystemen ermöglicht die kurzfristige Produktion einer großen Anzahl von unterschiedlichen, dabei aber inhaltlich zusammenhängenden artifiziellen Objekten (Urkunden, Briefe, Bilder, Ton-/Videoobjekte, ...). Solchermaßen kontextbezogene und untereinander abgestimmte Objekte erzeugen trotz ihrer Fiktion eine hohe Plausibilität.

Wird eine große Anzahl von in sich abgestimmten artifiziellen Objekten erzeugt, erscheinen die ggf. wenigen ordnungsgemäßen Objekte als manipuliert. So sind Kontextanalysen und weitere bisher zielführenden Verfahren zur Offenbarung von Falsifikaten (Barata, 2004) nur bedingt

geeignet, um konstruierte, alternative Realitätsdarstellungen aufzudecken.

Deepfake[6] kann im Kontext der digitalen Bewahrung von Kulturgütern insbesondere zur Kontextualisierung von artifiziellen Objekten eingesetzt werden. Nicht nur in gefälschten generierten Texten kann plausibler Kontext geschaffen werden, sondern Personen oder Institutionen, denen wir Vertrauen schenken, belegen, erläutern oder bekräftigen angeblich in Videos oder Podcasts Sachverhalte, die erst durch artifizielle Objekte geschaffen wurden. Dies stärkt einerseits die Glaubwürdigkeit artifizieller Objekte und ist andererseits als Fälschung nicht identifizierbar, da es sich, genau wie bei jedem artifiziellen Objekt, um ein generiertes Objekt handelt, das keine technischen Fälschungsparameter aufweist.

Existenzfrage

Es kann die Frage gestellt werden, ob artifizielle digitale Objekte bereits in unseren Sammlungen und Archiven vorhanden sind. Grundsätzlich existiert die Option artifizielle Objekte in unsere Speichersysteme einzuschleusen, weil nicht alle Softwaresysteme über entsprechende Sicherheitsmaßnahmen zum Schutz vor Injection-Angriffe[7] verfügen.

Letztlich kann aber keine Antwort auf die Existenzfrage gegeben werden. Auch wenn die angenommene Wahrscheinlichkeit tendenziell eher niedrig ist, muss bedacht werden, dass nur die Angriffe auf Softwaresystem wirklich erfolgreich sind, die nicht bemerkt werden. Die Realisation und Offenlegung von erfolgten Angriffen bestimmt jedoch unsere Einschätzung für deren Auftrittswahrscheinlichkeit. Erschwerend kommt hinzu, dass der Erfolg von Angriffen daran zu bemessen ist, dass niemand darauf aufmerksam wird.

Dies, gepaart mit der fehlenden Detektionsoptionen für artifizielle Objekte, führt zu der Feststellung, dass es sich unserer Kenntnis entzieht, ob und wenn ja wie viele artifizielle digitale Objekte bereits verbreitet sind.

Vertrauen in die Institutionen

Das Fälschen von Berichten, Bildern und Objekten ist kein neues Phänomen einer digitalen Gesellschaft. Zu allen Zeiten wurden Informationen absichtlich oder aus Versehen falsch weitergegeben. Das Vertrauen in verantwortliche Institutionen war und ist eine der wesentlichen Komponenten für die Entscheidung, ob Menschen Inhalte als wahrheitsgemäß oder falsch bewerten. Durch die Etablierung der digitalen Informationsverbreitung und -bewahrung hat sich nur die Art des Veränderns von Inhalten geändert, nicht aber die Tatsache selbst.

Generell konnten in der Prä-KI-Zeit die meisten Fälschungen durch Kontext- und Plausibilitätsprüfungen aufgespürt werden. Erst mit Einführung der generativen KI ist es möglich geworden, eine so große Anzahl an untereinander abgestimmten Fälschungen zu erstellen, dass diese in sich geschlossen und plausibel sind und die wahrhaft authentischen Objekte als Fälschungen erscheinen zu lassen.

Das Ergebnis jeder Forschungstätigkeit[8] unabhängig von der Forschungsdisziplin wäre wertlos, wenn diese auf artifiziellen digitalen Objekten mit fiktiven Inhalten beruhen würde. Aus diesem Grund benötigen Forschenden zukünftig eine neue Vertrauenskomponente in digitalen Objekten, die sicherstellt, dass es sich nicht um artifizielle, sondern um sorgfältig digitalisierte substanzielle Objekte handelt, deren Gegenstand das authentische digitale Abbild eines zum Zeitpunkt der Digitalisierung tatsächlich vorhandenen realen Original-Objektes ist.

Das wichtige Urvertrauen, das von Forschenden und der Gesellschaft in die Kulturerbe-Institutionen gesetzt wird, wird allein nicht mehr ausreichend sein und muss durch technische Komponenten unterstützt und gerechtfertigt werden.

Security Objectives

Lt. Simon (2020) ermöglicht Vertrauen auch in Situationen der Ungewissheit die Entwicklung eines Sicherheitsgefühls, wenn es möglich ist, sich auf das Handeln anderer zu verlassen.

Die Umsetzung von präventiven Schutzmaßnahmen ist ein bewährtes Mittel zur Vertrauensbildung. Elementare Kernpunkte für den Schutz von digitalen Objekten des Kulturerbes sind in den Schutzzielen der Informationssicherheit formuliert (NIST, 2004). Institutionen können das Vertrauen von Forschenden in die von ihnen zur Verfügung gestellten digitalisierten Objekte durch Beachten der international anerkannten Schutzziele für Informationssicherheit[9] rechtfertigen.

CIA-Triade

1. Confidentiality-Vertraulichkeit
2. Integrity-Integrität
3. Availability-Verfügbarkeit

Zwei Schutzziele des Identitätsmanagements ergänzen diese Triade:

1. Authenticity - Echtheit im Sinne von Ursprünglichkeit
2. Non-Repudiation - überprüfbare Beweise werden erstellt, dass das Erstellen des Objektes nicht in Abrede gestellt werden kann.

Vertraulichkeit bedeutet hier, dass nur berechtigte Personen digitale Objekte einsehen können (Autorisierung/Verschlüsselung).

Integrität hingegen zielt auf unbemerkte Modifikationen/Manipulationen und stellt die Korrektheit und Vollständigkeit sicher (elektronische Signatur/Siegel[10] oder Blockchain (Lo Duca et al., 2020)).

Die *Verfügbarkeit* beschreibt die Sicherstellung des Zugriffs auf die Daten und Vermeidung von Datenverlusten (BackUp-Strategien und gegebene Abrufbarkeit).

Ist ein Objekt mit geeigneten Verfahren und kryptografischen Algorithmen signiert, gesiegelt oder in einer Blockchain gesichert, kann sichergestellt werden, von wem und wann es erstellt und dass es anschließend nicht verändert wurde (BSI, 2020).

Bedacht werden muss aber unbedingt, dass elektronische Signaturen und Siegel sowie Blockchain Verfahren und digitale Wasserzeichen immer auf bestehende digitale Objekte angewendet werden und sie finden grundsätzlich erst nach der Transformation eines realen in ein digitales Objekt statt.

Für die Umsetzung der oben genannten Schutzziele gibt es konkrete Handlungsempfehlungen für die Institutionen des Kulturerbes, die auch im Forschungsdatenmanagement der DH zu finden sind. Jedoch bieten all diese keine unmittelbare Möglichkeit die oben beschriebene Gefährdung durch artifizielle Objekte zu verhindern, da sie erst dann ansetzen, wenn die Transformation von einem real existierenden in ein digitales Objekt erfolgt ist.

Es handelt sich um Schutzziele für alle digitale Objekte, zu denen auch die artifiziellen Objekte gehören. Sie stellen prinzipbedingt kein geeignetes Mittel zur Identifikation von artifiziellen digitalen Objekten dar.

Jedoch besteht die Möglichkeit, zertifikatsbasierte Signaturen/Siegel oder Blockchain-Lösungen sowie digitale Wasserzeichen von unterschiedlicher Ausprägung in digitalen Objekten zu verwenden, um sicherzustellen, dass eine bestimmte Institution (Zertifikatsinhaber) ein Objekt erzeugt hat und sich damit für die korrekte Transformation verantwortlich zeigt. Auf solchen Objekten basierende Forschungsergebnisse beruhen dann mit hoher Wahrscheinlichkeit nicht auf artifiziellen Objekten.

Digitales Vertrauen

Ein ungelöstes Kernproblem in der digitalen Welt ist das Abbilden von Vertrauen. Unterschieden werden muss unbedingt zwischen Proof und Trust. Nachweise stellen ein der Grundlagen für Vertrauen dar, können selbst aber nur bedingt Vertrauen schaffen. Vertrauen entsteht jenseits des Zweifels und ist wie bei Luhmann beschrieben eine soziale Kulturtechnik. Es stellt sich die Frage,

1. wann kann ein digitaler Inhalt als vertrauenswürdig gelten?
2. kann Vertrauen hergestellt werden, indem eine Vertrauenskomponente in digitale Objekte verankert wird?
3. was bedeutet der Verlust von Vertrauen in jegliche digitalen Objekte/Inhalte?

Content Authenticity Initiative

Die Idee der *Content Authenticity Initiative* "CAI"[11] ist, durch Anwendung kryptografischer Verfahren Inhalte und Metadaten gegen unbemerkte Manipulationen mit zertifizierten digitalen Signaturen zu schützen und Transparenz zu erzeugen, so dass Nutzende entscheiden können, ob sie den Inhalten Vertrauen schenken oder nicht. Basierend auf den C2PA Spezifikationen[12] werden Herkunfts-, Veränderungs- und Urhebernachweise festgehalten, die über ein Info-Icon in der Repräsentation des Objekts aufgerufen werden kann.

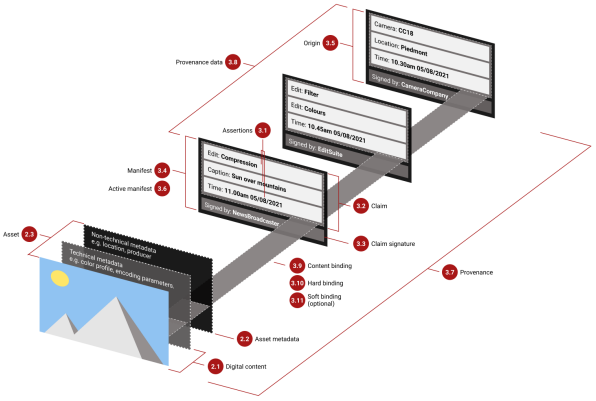


Abbildung 1: Bildnachweis: Coalition for Content Provenance and Authenticity, CC BY-SA 4.0, via Wikimedia Commons

So entsteht die Möglichkeit, selbsterzeugte Objekte CAI-konform zu kennzeichnen und festzuhalten, ob und wenn ja in welchem Umfang KI-Methoden Anwendung gefunden haben. Es erfolgt aber an keiner Stelle eine Überprüfung auf Wahrheitsgehalt. Es liegt ausschließlich bei den Usern zu entscheiden, ob sie dem System vertrauen oder nicht.

Bei dieser Vorgehensweise kann, genau wie bei allen anderen freiwilligen Selbstverpflichtungs-Maßnahmen bezüglich einer KI-Verwendungskennzeichnung[13], nicht sichergestellt werden, dass ein Objekt ohne Kennzeichnung, trotzdem KI generierte Inhalte enthält. Jede Art der freiwilligen Kennzeichnung darf nicht zu dem Umkehrschluss führen, dass nicht gekennzeichnete Objekte garantiert ohne KI entstanden sind.

Implikationen

Es kann durch unentdeckte artifizielle Objekte eine Bedrohungslage für eine authentische wissenschaftliche Forschung entstehen und es wird möglich, gezielt die Realität und die Historie verzerrt darzustellen. Die Manipulationsstrategie bei Verwendung artifizieller Objekte setzt nicht wie bisher auf der Ebene einzelner Objekte, sondern auf der Ebene des Gesamtverständnisses eines Sachverhaltes durch den Menschen an. Dies stellt eine deutlich umfangreichere Bedrohungslage dar.

Insbesondere eine Kombination aus Verbreitung artifizieller Objekte über soziale Netzwerke und deren Verfügbarkeit bei Portalen vertrauenswürdiger Institutionen stellt eine große Gefahr dar. Über die sozialen Netzwerke wird eine sehr hohe Reichweite erzielt und über Portale vertrauenswürdiger Institutionen großes Vertrauen in die Objekte hergestellt.

Die Kontrolle der Vergangenheit kann als eine Methode der Machtausübung eingesetzt werden, wie bereits George Orwell in seinem Roman 1984[14] schreibt: " Wer die Vergangenheit kontrolliert, der kontrolliert die Zukunft. ... Wer die Gegenwart kontrolliert, der kontrolliert die Vergangenheit." Falsche Aussagen von Machtausübenden können durch manipulierte historische Dokumente gestützt werden. Artifizielle Objekte unterstützen derartige Prozesse optimal.

Trustmanagement

Eine eindeutige Identifizierung von KI-generierten Material und ein berechtigtes Vertrauen in digitale nicht automatisch generierte Inhalte wird nicht nur in den Digital Humanities, sondern in vielen Disziplinen aus verschiedensten Gründen dringend benötigt.[15]

Zur Absicherung des prinzipiell vorhandenen Urvertrauens, das den Institutionen von Forschenden entgegengebracht wird, sollten Institutionen und Einrichtungen des Kulturerbes Maßnahmen (vor allem technischer Natur) ergreifen, um dieses Vertrauen in ihre digitalen Inhalte zu rechtfertigen.

Vertrauen ist auch im Forschungsprozess ein zentraler Aspekt.

Leider wird der Begriff Vertrauen oder Trust häufig nicht für das verwendet, was hier unter Vertrauen verstanden wird. Golbecks, Parsias, und Hendlers ‚Web of Trust' (2003) beschreibt ein Netz gegenseitiger Bestätigungen für digitale Schlüssel und Rahimzadeh Holagh und Mohebbi (2019) beschreiben für ihre im Semantic Web of Things enthaltenen Trust-Layer einen Blockchainansatz als Nachweiskette. Beides wird hier als Proof, nicht als Trust verstanden.

Ein interessanter Ansatz ist das ‚TrustNet' aus der angewandten KI-Forschung von Schillo und Funk (2000), das Agenten in einem Multiagenten System in die Lage versetzt, das Vertrauen in andere zu bewerten.

Im Kontext des Modells des Semantic Webs wurde 2001 die Idee eines Trust-Layer vorgestellt[16], aber es finden sich keine technischen Realisationen dazu.

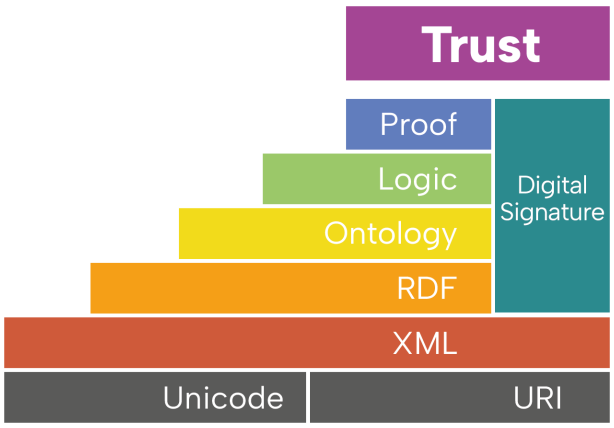


Abbildung 2: Semantic Web Layer Cake

Die Frage, die sich stellt, ist, wie ein erweitertes Modell der nicht neuen Idee des Trustmanagements aussehen könnte, bei dem eine Vertrauenskomponente explizit auf die Authentizität zur Sicherstellung von Non-Artifizialität im Objekt angewendet wird und somit eine Zero-Trust Lösung darstellt, die im Objekt selbst verankert ist und nicht auf einer Trust Chain beruht.

Fußnoten

1. Born digital data bringen für Kulturerbe-Institutionen viele Probleme mit sich und es besteht ein großer Forschungsbedarf in diesem Sektor. Dies ist ausdrücklich nicht Thema dieses Papers.
2. Konkrete Beispiele für digitale Objekte ohne verfügbares, zugehöriges Original-Objekt sind Abbilder von mit der Zeit unbrauchbar gewordenen Materialien wie Zeitungen, bei denen die zunächst vorhandenen Original-Objekte aufgrund der Papierbeschaffenheit mit der Zeit zerfallen oder digitale Objekte der Ruinenstadt Palmyra, wo bewusste Zerstörung des Baal-Tempels 2015 dazu führte, dass kein Original-Objekt verfügbar ist.
3. Fiktiv wird hier nicht in literaturwissenschaftlichen Sinn verwendet, sondern im Sinn von angenommen, erdacht oder erdichtet.
4. „Fiktive Objekte" werden hier als imaginäre Objekte gesehen, die das Potential haben, dass es sie tatsächlich geben könnte. Tatsächlich existieren sie aber nur in der Vorstellungskraft.
5. Born digital Objects haben ebenfalls keine Existenz außerhalb ihrer digitalen Form. Dennoch sind diese keine artifiziellen Objekte in diesem Sinn, da das Konzept dieser Objekte keine Idee eines digitalen Abbildes aufweist.
6. Deepfake bezeichnet eine Technologie, die die Erstellung von täuschend echt gefälschten Bildern, Videos oder Audiodateien ermöglicht. Diese Inhalte werden mithilfe von KI erzeugt, um Personen in bestimmten Situationen oder Handlungen darzustellen, die in Wirklichkeit nicht stattgefunden haben. Deepfakes können verwendet werden, um Gesichter in Videos zu manipulieren, Stimmen zu

fälschen oder Szenarien zu erstellen, die authentisch wirken, obwohl sie in Wahrheit konstruiert sind. Die Technologie hat weitreichende Auswirkungen auf die Medienlandschaft, die Privatsphäre und die Glaubwürdigkeit von Inhalten, da sie die Grenzen zwischen Realität und Fiktion verschwimmen lässt.
7. Als Injection-Angriff wird das Ausnutzen von Sicherheitslücken im Zusammenhang mit Datenbanken verstanden, wobei Angreifende Datenbankbefehle einschleusen, Daten einfügen, auslesen, verändern oder löschen oder die Kontrolle über den Datenbankserver erlangen.
8. Gemeint ist hier jede Forschung, die auf digitalen Objekten beruht und Forschende davon ausgehen, dass diese ein ordnungsgemäß transformiertes Abbild eines realen Originals darstellen. Denkbar ist eine zukünftige Forschung, die artifizielle Objekte zum Gegenstand hat. Für diese gilt diese Aussage nicht.
9. Beschrieben in ISO/IEC 27001, http://www.itref.ir/uploads/editor/42890b.pdf 12.07.2023 und NIST (2004)
10. „Technisch sind diese vergleichbar mit den einer juristischen anstatt einer natürlichen Person.“. https://www.bsi.bund.de/DE/Themen/Oeffentliche-Verwaltung/eIDAS-Verordnung/Elektronische-Signaturen-Siegel-und-Zeitstempel/elektronische-signaturen-siegel-und-zeitstempel_node.html 13.07.2023
11. https://contentauthenticity.org/ 20.11.2023
12. https://c2pa.org/ 20.11.2023
13. Zum Beispiel 21.07.2023 Selbstverpflichtung der große Techunternehmen, 09.2023 Kennzeichnungsoption bei TikTok, 11.2023 Konzept AI Safty Institute und Bletchley Declaration und vieles mehr.
14. https://politik.brunner-architekt.ch/wp-content/uploads/orwell_george_1984.pdf S. 366; 11.07.2023
15. Wie reagieren Systeme auf die Verwendung von KI-generierten Texten für das Trainieren von KI-Softwaresystemen? Shumailov et al. (2023) in‚The curse of recursion: Training on generated data makes models forget‘ festgestellt, das dies zu einem Modellkollaps führt. Außerdem werden bei Data Poisoning Attacken die Vorhersagemodelle korrumpiert und die gesellschaftlichen Auswirkungen eines Vertrauensverlustes in die Kulturerbeinstitutionen sind weitreichend.
16. https://www.w3.org/2001/12/semweb-fin/w3csw 21.11.2023

Bibliographie

Barata, Kimberly. 2004. Archives in the Digital Age. Journal of the Society of Archivists 25 (1): 63–70. https://doi.org/10.1080/0037981042000199151

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2020. Leitlinie für digitale Signatur-/ Siegel-, Zeitstempelformate sowie technische Beweisdaten (Evidence Record). https://www.bundesnetzagentur.de/EVD/DE/SharedDocuments/Downloads/Anbieter_Infothek/BSI_TR_03125.pdf?__blob=publicationFile&v=1 13.07.2023

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2021. Leitlinie für die beweiswerterhaltende Aufbewahrung gemäß BSI TR-03125 TR-ESOR. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/TechnischeRichtlinien/TR03125/BSI_TR-ESOR-LEIT.pdf 13.07.2023

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2023. IT-Grundschutz-Kompendium. 6. Edition 2023, Reguvis Fachmedien GmbH, Köln. https://d-nb.info/1282075888 und https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/IT-Grundschutz-Kompendium/it-grundschutz-kompendium_node.html 12.07.2023

Coester, Ulla, Norbert Pohlmann. 2021. Vertrauen – ein elementarer Aspekt der digitalen Zukunft. Datenschutz und Datensicherheit – DuD. Springer Nature. https://doi.org/10.1007/s11623-021-1401-x

Dittmann, Jana. 2000. Digitale Wasserzeichen: Grundlagen, Verfahren, Anwendungsgebiete., Berlin: Springer.

Fliehe, Marc, Brummel, Elisa. 2021. Eckpunkte eines sicheren Ökosystems für KI-Anwendungen. Datenschutz Datensicherheit - DuD 45, 444–447. https://doi.org/10.1007/s11623-021-1468-4

Golbeck, Jennifer, Bijan Parsia und James Hendler. 2003. Trust Networks on the Semantic Web. In: Klusch, Matthias, Andrea Omicini, Sascha Ossowski und Heimo Laamanen. 2003. Cooperative Information Agents VII. Lecture Notes in Computer Science, vol 2782. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-45217-1_18

Hornig, Anna, Christian Grünwald, Daniel Bonin, Jan Reichert, Marie-Kristin Komendzinski, Julian Sachs, Holger Glockner, Michael Astor. 2022. Studie: Die Zukunft des Vertrauens in digitalen Welten. [PDF]. Abgerufen von https://www.vorausschau.de/SharedDocs/Downloads/vorausschau/de/Foresight_Vertraueensstudie_Langfassung.pdf?__blob=publicationFile&v=1 (24.11.2023)

Liang, Xueping, Sachin Shetty, Deepack Tosh, Charles Kamhoua, Kevin Kwiat und Laurent Njilla. 2017. ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, Spain, 2017, pp. 468-477. https://doi.org/10.1109/CCGRID.2017.8

Lo Duca, Angelica, Clara Bacciu und Andrea Marchetti. 2020. The Use of Blockchain for Digital Archives: a comparison between Ethereum and Hyperledger (AIUCD 2019). Umanistica Digitale, 4(8). https://doi.org/10.6092/issn.2532-8816/9959

NIST National Institute of Standards and Technology. 2004. Standards for Security Categorization of Federal Information and Information Systems. Federal Information Processing Standards Publications FIPS PUB 199. Gaithersburg USA. https://doi.org/10.6028/NIST.FIPS.199

Rahimzadeh Holagh, Sam und Keyvan Mohebbi. 2019. A glimpse of Semantic Web trust. SN Appl. Sci. 1, 1732. https://doi.org/10.1007/s42452-019-1598-6

Simon, Judith. 2020. The Routledge Handbook of Trust and Philosophy. Routledge New York https://doi.org/10.4324/9781315542294

Schillo, Michael, Petra Funk, und Michael Rovatsos 2000. Using trust for detecting deceitful agents in artificial societies. Applied Artificial Intelligence, 14:8, 825-848, https://doi.org/10.1080/08839510050127579

Shan, Shawn, Ding, Wenxin, Passananti, Josephine, Zheng Haitao, Zhao, Ben. Oktober 2023. Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. arXiv:2310.13828 [cs.CR]. https://doi.org/10.48550/arXiv.2310.13828

Shumailov, Ilia, Shumaylov, Zakhar, Zhao, Yiren, Gal, Yarin, Papernot, Nicolas, Anderson, Ross. Mai 2023. The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493v2 [cs.LG]. https://doi.org/10.48550/arXiv.2305.17493

Stančić, Hrvoje. 2020. Trust and Records in an Open Digital Environment (1st ed.). Routledge. https://doi.org/10.4324/9781003005117

Von Menschen und Maschinen: Transdisziplinäre Workflows im Münsteraner Editionsprojekt Heinrich Scholz

Dietz, Katharina
kdietz@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0001-7405-3656

Dinger, Patrick
patrick.dinger@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0002-2649-4737

Horstmann, Jan
jan.horstmann@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0001-8047-2232

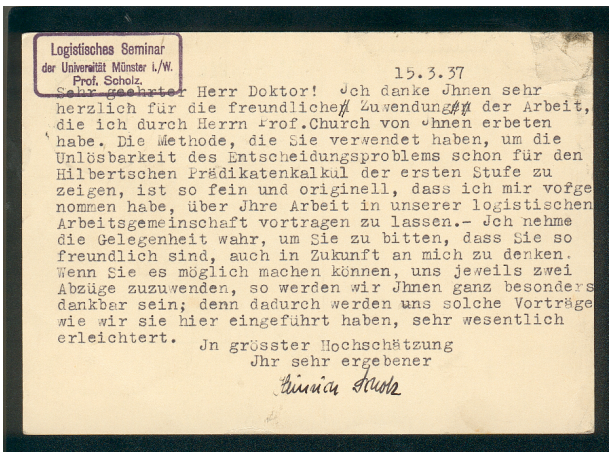
Normann, Immanuel
immanuel.normann@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0003-4702-1282

Schäfftlein, Vitus
vitus.schaefftlein@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0009-0003-3620-5884

Das Akademieprojekt Heinrich Scholz: Schnittstellen zwischen Menschen und Maschinen

Das Projekt „Heinrich Scholz und die Schule von Münster – Mathematische Logik und Grundlagenforschung“ ist ein 2023 gestartetes und auf 13 Jahre angelegtes Akademieprojekt, das den umfangreichen Nachlass des Theologen, Philosophiehistorikers, Metaphysikers, Mathematikers und Logikers Heinrich Scholz (1884–1956) den FAIR-Prinzipien entsprechend erforschbar machen wird.[1] Unter der Leitung von Prof. Dr. Niko Strobach (Philosophisches Seminar, Universität Münster) ist auch die ULB Münster mit mehreren Teilprojekten involviert. Ziel ist eine vollständige digitale Ausgabe zu Heinrich Scholz einschließlich seiner nicht veröffentlichten Manuskripte, eine digitale Edition seiner Korrespondenz sowie die semantische Verknüpfung, Annotation und dynamisch-interaktive Bereitstellung mittels Semantic-Web-Technologie in der Linked Open Data Cloud (vgl. etwa Wettlaufer, 2018). Zum Einsatz kommen u. a. Kalliope, Visual Library, verschiedene OCR/HTR-Tools, oXygen zur TEI-Edition sowie die Modellierung in RDF zur Publikation und Visualisierung des Nachlasses als Wissensgraph.

Noch im Jahr von Alan Turings einschlägiger Arbeit „On Computable Numbers, with an Application to the Entscheidungsproblem“ (Turing, 1937), die den Grundstein für die heutige Informatik legte, organisierte Heinrich Scholz das erste Seminar über das Turingmaschinenmodell. Er schrieb an Turing: „Die Methode, die Sie verwendet haben, um die Unlösbarkeit des Entscheidungsproblems schon für den Hilbertschen Prädikatenkalkül der ersten Stufe zu zeigen, ist so fein und originell, dass ich mir vorgenommen habe, über Ihre Arbeit in unserer logistischen Arbeitsgemeinschaft vortragen zu lassen.“


Logistisches Seminar
der Universität Münster i./W.
Prof. Scholz.

15.3.37

Sehr geehrter Herr Doktor! Ich danke Ihnen sehr herzlich für die freundliche Zuwendung der Arbeit, die ich durch Herrn Prof.Church von Ihnen erbeten habe. Die Methode, die Sie verwendet haben, um die Unlösbarkeit des Entscheidungsproblems schon für den Hilbertschen Prädikatenkalkul der ersten Stufe zu zeigen, ist so fein und originell, dass ich mir vorgenommen habe, über Ihre Arbeit in unserer logistischen Arbeitsgemeinschaft vortragen zu lassen.- Ich nehme die Gelegenheit wahr, um Sie zu bitten, dass Sie so freundlich sind, auch in Zukunft an mich zu denken. Wenn Sie es möglich machen können, uns jeweils zwei Abzüge zuzuwenden, so werden wir Ihnen ganz besonders dankbar sein; denn dadurch werden uns solche Vorträge wie wir sie hier eingeführt haben, sehr wesentlich erleichtert.

In grösster Hochschätzung
Ihr sehr ergebener
Heinrich Scholz

Abbildung 1: Postkarte von Heinrich Scholz an Alan Turing vom 15.03.1937, The Turing Digital Archive, Signatur d5.17a (https://turingarchive.kings.cam.ac.uk/correspondence-amtd/amt-d-5 04.07.2023)

Die Mitglieder des von Scholz ins Leben gerufenen Zentrums für mathematische Logik und Grundlagenforschung wurden auch bekannt als „Schule von Münster“. 1943 entstand in Münster der erste Lehrstuhl in Deutschland für mathematische Logik und Grundlagenforschung. Bis heute gibt es keine Gesamtausgabe der Werke von Scholz. Seine umfangreiche Rezensionstätigkeit und Briefkorrespondenz wurden nicht systematisch erschlossen (vgl. Molendijk, 2022).

Bereits zu Beginn des Projekts waren hinsichtlich notwendiger Absprachen der beteiligten Personen, Abteilungen und Systeme zahlreiche Hürden zu nehmen. Es galt, sowohl eine gemeinsame Sprache trotz stark divergierenden fachlichen Hintergründen (Digitalisierung, Katalogisierung, Digital Humanities, Sammlungsforschung und Philosophie) zu finden, als auch Workflows so zu strukturieren, dass die Anforderungen der jeweiligen Teilschritte bereits in den anderen Prozessschritten antizipiert werden und für Reibungslosigkeit an den Schnittstellen gesorgt ist. Eine Unterteilung von Schnittstellen, die auch unseren Vortrag strukturieren wird, kann bezüglich der Akteure *Mensch* und *Maschine* vorgenommen werden, wodurch sich die folgenden Typen von Schnittstellen ergeben:

1. Mensch-Mensch
2. Mensch-Maschine
3. Maschine-Maschine

1) Aus einer systemtheoretischen Perspektive zerfällt der Akteur *Mensch* in verschiedene Abteilungen und Rollen, die jeweils ihre eigenen Praktiken und Sprachen entwickeln, sodass es bei dieser Schnittstelle vor allem um Kommunikation geht: Als Dokumentationsmedium spielt die wikiartige Software *Confluence* im Projekt eine zentrale Rolle, auf die alle Projektbeteiligten Zugriff haben und die eine sorgfältige Planung und Pflege der Informationsstruktur erfordert. Regelmäßige, abteilungsübergreifende Arbeitstreffen organisieren Arbeitsabläufe. In diesen Arbeitstreffen und ihrer Dokumentation treffen verschiedene Fachjargons aufeinander, die zuvor selten den Weg über Abteilungen hinweg fanden. Etwa die Worte „Ausheben“, „Entmetallisieren“, „Sigel“, „Signatur“, „URN“, „DOI“, „Faksimile“, „Scan“ oder „Digitalisat“ können je nach Äußerungskontext in verschiedenen Abteilungen etwas anderes bedeuten.

2) *Mensch-Maschine-Schnittstellen* führen dann zu Reibungen, wenn der Mensch mit unvertrauter Software oder kontraintuitiven grafischen Interfaces konfrontiert wird, deren Benutzung u. U. erst ermöglicht oder erlernt werden muss. Analog zu abteilungsspezifischen Fachjargons ist die Expertise abteilungsspezifischer Formate und Datenmodelle zur (Meta-)Datenerfassung: Während bei der Katalogisierung von Archivalien das EAD-Format[2] bekannt ist, wird bei der Digitalisierung durch Bibliotheken METS/MODS (vgl. Altenhöner u.a., 2023) verwendet, bei der Transkription TEI und bei der semantischen Anreicherung schließlich RDF. Da man die Arbeiten der im Workflow jeweils vorhergehenden Abteilung nachnutzen will, muss jeweils voneinander gelernt werden, was in einem bestimmten Schritt von wem erfasst wurde. Ein gemeinsames Verständnis muss hier an den Schnittstellen der Abteilungen aufgebaut werden. Die Formate dienen am Ende dazu, von Menschen geschaffene Information in eine maschinenlesbare Form zu gießen.

3) Eine ebenso wichtige Rolle spielen Formate für die Interoperabilität von Softwaresystemen, womit sie als eine wichtige *Maschine-Maschine-Schnittstelle* aufgefasst werden können. Auf basaler Ebene stehen sich XML und JSON als Formatparadigmen gegenüber: Während XML-basierte Formate zu dokumentarischen Zwecken bevorzugt eingesetzt werden, überwiegt JSON als Datenaustauschformat für Webschnittstellen. Prominent sind METS/MODS und TEI für das XML-Paradigma und IIIF[3] oder die Formate von Suchmaschinen-APIs für das JSON-Paradigma. Eine Aufgabe besteht darin, Daten zwischen den Paradigmen automatisch zu konvertieren. So müssen etwa die bei der Digitalisierung in METS/MODS enthaltenen Bildmetadaten in ein IIIF-konformes JSON-Manifest konvertiert werden, um die Digitalisate für die Edition über einen IIIF-Server bereitzustellen. Auch innerhalb eines Paradigmas stehen verschiedene Konvertierungen (etwa von TEI-XML nach HTML).

Im Folgenden wollen wir beispielhaft zwei große Aspekte des Projekts anhand der in ihnen vorkommenden unterschiedlichen Schnittstellen thematisieren: zum einen die Katalogisierung/Digitalisierung des Scholz-Nachlasses, zum anderen die Texterkennung verschiedener Nachlassmaterialien.

Katalogisierung und Digitalisierung

Die Katalogisierung und Digitalisierung des umfangreichen Scholz-Nachlasses stellt eine Herausforderung dar. Einerseits wurden neue Werkzeuge eingeführt und bestehende Workflows der Bereitstellung digitalisierter Kulturgüter erweitert, andererseits mussten die Kommunikationswege für den interdisziplinären Austausch des

Projektteams gefunden werden. Da der Nachlass zu Projektbeginn nur grob strukturiert war, wurde eine im Vorfeld erstellte Findliste (vgl. Heitfeld-Rydzik u.a., 2022) herangezogen. In regelmäßigen Abstimmungstreffen konnte so ein gemeinsamer Startpunkt für die Feinsortierung und Umsystematisierung von Dokumenten und Objekten einerseits, die Festlegung eines Startpunkts für die zeitgleiche Katalogisierung und Digitalisierung andererseits festgelegt werden. In der Folge wurde der Nachlass in Abschnitte geteilt und zyklisch feinsortiert. Damit die erschlossenen Materialien mit DH-Methoden prozessiert werden können, wurde zunächst eine Mensch-Mensch-Schnittstelle geschaffen. Damit war es möglich, Arbeitsschritte in den Bereichen Aushebung, Katalogisierung und Digitalisierung zu parallelisieren, was eine frühe Bereitstellung erster digitalisierter Dokumente ermöglichte.

Der weitere Erschließungsprozess in der ULB ist durch zwei voneinander abhängige Mensch-Maschine-Interaktionen geprägt: die Katalogisierung und die Digitalisierung. Durch die Katalogisierung in der Kalliope-Verbunddatenbank[4] wird das Wissen der Aushebung und Feinsortierung konsolidiert und standardisiert in Katalogaufnahmen und Strukturbäumen festgeschrieben. Aufbauend auf den erzeugten Metadaten der Katalogisierung werden die Nachlassdokumente digitalisiert (vgl. Altenhöner u.a., 2023). Während die Digitalisierung selbst als Interaktion der Scanoperator*innen mit physischen Nachlassobjekten und dem Digitalisierungssystem geprägt ist (Mensch-Maschine-Schnittstelle), findet im Hintergrund ein Informationsaustausch der Systeme über verschiedene maschinelle Schnittstellen statt. Granulare Identifier[5] werden vergeben, Metadaten und Strukturinformationen ausgetauscht. Diese Maschine-Maschine-Interaktion ermöglicht, das Wissen über die physikalischen Dokumente (Katalogaufnahmen), die festgelegte Nachlass-Tektonik und die Digitalisate in einem System, hier der Visual Library[6], zusammenzuführen. Die Visual Library ist dabei jedoch mehr als das Zielsystem des Teilworkflows. Über die definierten Maschine-Maschine-Schnittstellen wie OAI-PMH für die METS/MODS-Daten oder IIIF wird das Portal *Kulturgut Digital*[7] der ULB Münster zur Datenmanagementplattform für die automatisierte Bereitstellung der erschlossenen Nachlassdokumente der digitalen Edition.

Erzeugte Digitalisate und Metadaten durchlaufen von dem Zeitpunkt ihrer Entstehung bis zu ihrer Repräsentation als Linked Open Data verschiedene Systeme und Anwendungen. In jedem Bearbeitungs- und Anreicherungsschritt entstehen unterschiedliche Nutzformate, die die vorhandenen Informationen aufgreifen, erweitern oder zusammenführen. Um diese Abhängigkeiten des Workflows von spezifischen Eingangs- und Ausgangsformaten einerseits, die Flexibilität des wissenschaftlichen Erkenntnisgewinns andererseits aufrecht zu halten, werden Mechanismen der Versionierung von Anfang an mitgedacht und verankert (s.u.). Mithilfe gezielter Maschine-Maschine-Interaktionen, aber auch über definierte Schnittstellen (SRU, OAI-PMH, IIIF), Standardisierung und dem Zusammenführen zentraler Informationen aus verschiedenen Nutzformaten wird eine einheitliche und reproduzierbaren Datengrundlage für den Editionsprozess geschaffen, die im Folgenden näher erläutert wird.

Texterkennung

Texterkennung im Scholzprojekt geschieht sowohl in der Erschließung der Werke Heinrich Scholz', die bereits in gedruckter Form vorliegen, als auch bei der Erschließung des knapp 100.000 Einzelseiten umfassenden handschriftlich wie maschinenschriftlich verfassten Nachlasses. Bei den gedruckten Werken konnte sofort mit dem Scannen begonnen werden. Zur Distribution der Scans innerhalb des Projekts wird die Datenaustauschplattform Sciebo genutzt, das eine Cloud-Speicherung auch größerer Datenmengen sowie kollaborative Bearbeitung ermöglicht. Auch das Scannen des Nachlasses konnte nach umfangreichen Vorbereitungen mittlerweile beginnen.

Die Scans der veröffentlichten Scholz-Rezensionen wurden zur Qualitätsverbesserung zunächst mit dem Tool Scan Tailor Advanced[8] nachbearbeitet und schließlich mit der OCR-Software tesseract[9] weiterverarbeitet. In der OCR-Erkennung und den dafür notwendigen Workflows zeigen sich entscheidende Mensch-Maschine-Schnittstellen: ScanTailor Advanced und tesseract wurden beide als Kommandozeilenprogramme entwickelt. Um die Zugänglichkeit für studentische Hilfskräfte zu verbessern, haben wir zwei Skripte geschrieben, die eine Bedienung via grafischer Nutzeroberfläche ermöglichen.[10]

Das Skript für ScanTailor fordert die User zunächst zur Auswahl einer TIF-Datei auf, falls diese nicht bereits als Kommandozeilenargument vergeben wurde. Das Skript erstellt einen Ordner mit einer Kopie der zu bearbeitenden Datei und öffnet ScanTailor Advanced. Der ScanTailor-Output (mehrere einseitige TIF-Dateien) wird in eine einzelne PDF-Datei umgewandelt. Eine vorkonfigurierte Maschine-Maschine-Schnittstelle im Post-Processing ermöglicht eine anschließende Ausführung der PDF-Datei im zweiten Skript: der Texterkennung. Eine grafische Benutzeroberfläche (vgl. Abb. 2) erlaubt die Angabe von Trainingsdatensätzen, die je nach Schriftart und Sprache variieren, und die Auswahl von dateispezifischen Optionen wie automatisches Entzerren oder Überschreiben eines bereits vorhandenen OCR-Layers. Ein Klick auf den Startbutton erzeugt aus den Angaben den entsprechenden Kommandozeilenbefehl für ocrmypdf und führt ihn im Hintergrund aus.

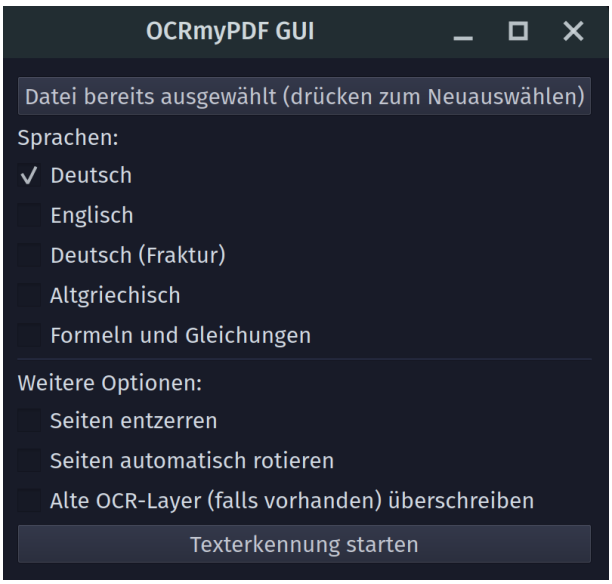


Abbildung 2: GUI für OCRmyPDF

Auch der Ablauf der Nachlassaufbereitung hält herausfordernde Schnittstellen bereit, etwa zwischen Maschine und Maschine. Da OCR4all (vgl. Reul et al. 2019) vorab Informationen benötigt, um welchen Schrifttyp es sich bei dem jeweils zu erkennenden Scan handelt, sind zusätzliche Vorverarbeitungsschritte notwendig. Insgesamt gestaltet sich der schriftliche Nachlass in Bezug auf Art, Größe und Schrifttyp ausgesprochen heterogen. Angesichts des Nachlassumfangs ist eine automatisierte Lösung zu finden. Da OCR und Handschriftenerkennung an den Schrifttyp angepasste Modelle erfordern, muss eine entsprechende Vorsortierung erfolgen. Diese Aufgabe umfasst ggf. auch das Zergliedern einer einzelnen Seite in mehrere Einheiten (vgl. Abb. 3).

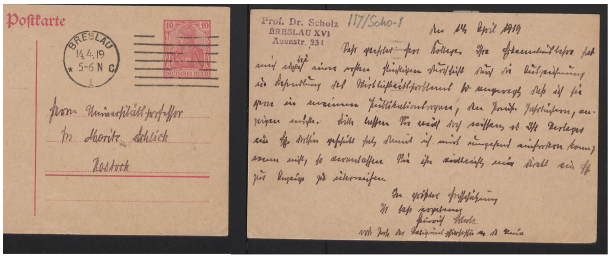


Abbildung 3: Postkarte vom 14.04.1919, Heinrich Scholz an Moritz Schlick, Testscan aus dem Nachlass Scholz, Universitäts- und Landesbibliothek Münster

Zur Segmentierung der Seiten bietet OCR4all zwei Möglichkeiten: Zum einen die Einbindung der selbstlernenden Layout-Erkennungssoftware LaReX, zum anderen ein Interface zur händischen Segmentierung nach Layout. Diese händisch markierten Seiten dienen außerdem als Trainingsmaterial für die nachfolgende maschinelle Layout-Erkennung. Während die grafische Benutzeroberfläche die Ausführung der händischen Segmentierung erleichtert und beschleunigt, waren Vorschläge zur automatischen Selektion unzureichend. Ein Teil wird daher manuell von Hilfskräften als Goldstandard und Trainingsgrundlage bearbeitet, auf dessen Grundlage geprüft werden kann, ob sich die Vorschläge von LaReX im Laufe der Arbeit verbessern. Die erkannten Texte sowie Metadaten werden anschließend in TEI-XML codiert und mit dem dazugehörigen Scan verknüpft. Eine statistische Bewertung der OCR-Qualität wird für verschiedene Schrift- und Dokumenttypen vorgenommen (vgl. Neudecker et al., 2021, 138–165).

Eine besondere Herausforderung besteht in der Erkennung der zahlreichen im Material befindlichen mathematischen Formeln. Die genaue Art und der Detailgrad der zu erfassenden Informationen muss bestimmt werden. Ob Formeln in der Textfassung nur als solche abgebildet oder auch weitere semantische Informationen hinterlegt werden sollen, hat im weiteren Forschungsverlauf Konsequenzen für sämtliche Typen von Schnittstellen.

Im Bereich der Maschine-Maschine-Interaktion stellt die Datenhaltung einen besonderen Aufgabenbereich dar, der im Hintergrund an jeder Stelle des Projekts über den gesamten Projektzeitraum von essentieller Bedeutung ist. Zu Beginn der Projektarbeiten mussten bereits die zentralen Pfeiler dieser Architektur bestimmt und implementiert werden. Die Daten, die im Projekt geschaffen werden, werden von zahlreichen verschiedenen Zugangspunkten eingespeist und später zusammengeführt. Alle IT-Subsysteme im Projekt sind durch die gemeinsame Versionierung verbunden. Neben der punktgenauen Zusammenführung der verschiedenen Dateien gibt es einige weitere spezifische Anforderungen, die an ein Versionierungssystem zu stellen sind, um den Anforderungen des Projekts gerecht zu werden. Zum einen ist eine Datenmenge von mehreren Terabyte zu erwarten, zum anderen verteilt sich diese Datenmenge auf viele einzelne Dateien. Eine Langzeitsicherung muss von Anfang an mitbedacht werden, da die Projektlaufzeit mit 13 Jahren den Zeitraum, den Projektdaten allgemein verfügbar vorgehalten werden, bereits überschreitet. Aus diesen Gründen fiel die Entscheidung auf das auf git-annex[11] basierende Versionierungs-Tool datalad[12], in dem der datenbasierte Projektworkflow vollständig abgebildet werden soll.

Diskussion und Ausblick: Workflows, Absprachen und Interoperabilität

Wenn man eine digitale Edition nicht auf dem leeren „Papier“, sondern innerhalb vorhandener Infrastrukturen und ihrer jeweiligen Vorgaben beginnt, entstehen zahlreiche potenzielle Reibungsflächen, die wir in unserem Beitrag als vielfältige Schnittstellen zwischen Menschen und Maschinen beschreiben. Es ist für den Projekterfolg relevant, sich bereits in der Konzeption von Arbeitsschritten und Workflows bewusst zu machen, dass Menschen mit unterschiedlichen disziplinären Hintergrün-

den, auf verschiedenen Ausbildungsstufen und mit divergierenden Zugangsrechten nicht nur untereinander eine gemeinsame Sprache finden müssen (Mensch-Mensch-Schnittstelle), sondern auch unterschiedliche Voraussetzungen haben, um mit digitalen Tools umgehen zu können (Mensch-Maschine-Schnittstelle): Einige Nutzer*innen können etwa mit Kommondozeilentools umgehen, andere benötigen eine grafische Oberfläche. Diese Punkte einzukalkulieren und mit einer fundierten Planung insbesondere der Übergabepunkte zu begegnen, ist ebenso wichtig wie auf technischer Ebene über Schnittstellen eine Interoperabilität der eingesetzten (ggf. proprietären) Tools zu erreichen (Maschine-Maschine-Schnittstelle).

Beim vorgestellten Akademieprojekt zum Nachlass Scholz zeigten sich diese Herausforderungen von Beginn an. Auch nach der Digitalisierung, Katalogisierung und Texterkennung werden mit steigender Komplexität der erzeugten Daten und Aufgaben potenzielle Reibungsflächen einzuplanen sein: Die erkannten Drucke und Handschriften sollen in TEI-XML editiert (Schnittstelle: Page-XML zu TEI-XML) und schließlich als Linked-open-Data in RDF modelliert werden (vgl. Wettlaufer 2018). Während Austauschformate für die Interoperabilität auf Datenebene sorgen, wird bei einer API über Protokolle die Kommunikation der Daten geregelt. Wir gehen davon aus, dass in den meisten Fällen das im Web omnipräsente HTTP-Protokoll auch bei uns die größte Rolle spielen wird. Aber ebenso wie TEI auf XML aufsetzt, so setzt jede Web-API zwar auf HTTP auf, jedoch in jeweils spezieller Ausprägung.

Vorhaben und Dauer des Projekts drängen eine genauere Auseinandersetzung mit den zahlreichen vergleichbaren DH-Editionsprojekten bezüglich angewandter Technologien, Standards und Tools geradezu auf. Im einschlägigen Akademien-Vorhaben „Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung“ (vgl. Kraft und Dumont 2020) an der Berlin-Brandenburgischen Akademie der Wissenschaften entsteht die Edition *humboldt digital* (Laufzeit 2015–2032).[13] Mit Fokus auf X-Technologien werden u.a. Reisejournale, Tagebücher und auch Korrespondenzen in TEI-XML ediert. Als Editionsumgebung kommt hierbei *ediarum*[14] zum Einsatz, in der *oXygen* als XML-Editor und die XML-Datenbank *existdb* als zentrales Repositorium fungiert. Das Münsteraner Scholz-Projekt setzt ebenfalls *oXygen* ein, aber als zentrales Repositorium die Versionsverwaltungssoftware Git, da die Versionskontrolle im Editionsprozess von größerer Relevanz ist als die Suchmöglichkeiten, wie sie primär von XML-Datenbanken bereitgestellt werden. Anders als bei *humboldt digital* werden der Semantic-Web-Technologie für die Erfassung von Daten in RDF sowie ihre Bereitstellung über einen SPARQL-Endpoint eine zentrale Rolle zugemessen.

Im ersten Teilprojekt – der Edition der Briefe aus dem Nachlass Heinrich Scholz – lassen sich in Materialität, Ziel und Methodik beispielsweise auch Parallelen zum an der Universität Hamburg durchgeführten Projekt *Dehmel Digital* ziehen (vgl. Nantke et al., 2022).[15] Von der Nachnutzung bibliothekarischer Katalogdaten (HANS/Kalliope) über die automatische Auszeichnung (NER) bis hin zur Modellierung als Netzwerk setzt das Münsteraner Projekt ähnliche Akzente.

Möchte man gegenwärtig eine digitale Edition erstellen, die auch noch in 20 Jahren und länger funktioniert, lohnt sich ein Blick auf Projekte, die bereits sehr lange laufen und vor längerer Zeit zukunftsfähige Software- und Format-Entscheidungen treffen mussten. Das bereits seit 1956 in Bearbeitung befindliche Akademieprojekt zur Leibniz-Edition (Hannover, Münster, Potsdam und Berlin)[16] entspricht hinsichtlich der Datenmodellierung zwar nicht den heutigen Standards, hat mit seiner Wahl der Software *TUSTEP*[17] aber dennoch Weitsicht bewiesen. Neben der Tool-Wahl ist hinsichtlich der Langlebigkeit eines Datenprojektes wie der digitalen Edition immer auch die Beachtung von langzeitarchivierbaren Datenformaten entscheidend. Die HBZ gibt hier wertvolle Einschätzungen.[18]

Fußnoten

1. Vgl. www.heinrich-scholz.de (22.06.2023).
2. Encoded Archival Description; vgl. https://www.loc.gov/ead/ (30.06.2023).
3. Vgl. in diesem Zusammenhang etwa die Diskussion von Mertens (2021).
4. Vgl. https://kalliope-verbund.info/ (21.06.2023), vgl. Grothe 2006.
5. Für Digitalisate und Metadaten werden granulare URNs vergeben, Kalliope definiert für die Katalogaufnahmen eigene persistente Identifier (Kalliope-IDs). Vgl. Sommer u.a. 2008 sowie https://kalliope-verbund.info/de/support/sru.html (21.06.2023).
6. Vgl. https://www.semantics.de/visual_library/ (21.06.2023).
7. Vgl. https://sammlungen.ulb.uni-muenster.de/ (29.06.2023).
8. Vgl. https://github.com/4lex4/scantailor-advanced (30.06.2023).
9. Vgl. https://github.com/tesseract-ocr/tesseract und https://ocrmypdf.readthedocs.io (30.06.2023).
10. Vgl. https://codeberg.org/opensource-philosophy/PDFPostProcessingGUI (30.06.2023).
11. https://git-annex.branchable.com/ (29.11.2023).
12. https://www.datalad.org/ (29.11.2023).
13. Vgl. https://edition-humboldt.de/ (14.11.2023).
14. Vgl. https://www.ediarum.org/ (14.11.2023).
15. Vgl. https://dehmel-digital.de/ (14.11.2023).
16. Vgl. https://www.uni-muenster.de/Leibniz (14.11.2023).
17. Vgl. http://www.tustep.uni-tuebingen.de/ (14.11.2023).
18. Vgl. https://www.hbz-nrw.de/produkte/langzeitverfuegbarkeit/langzeitverfuegbarkeit-fuer-hochschulen-v1/lzv-dateiformatliste (14.11.2023).

Bibliographie

Altenhöner, Reinhard, Andreas Berger, Christian Bracht, Paul Klimpel, Sebastian Meyer, Andreas Neuburger, Thomas Stäcker und Regine Stein. 2023. *DFG-Praxisregeln 'Digitalisierung'*. Aktualisierte Fassung 2022. Zenodo. DOI: 10.5281/zenodo.7435724.

Grothe, Ewald. 2006. „Die kooperative Erschließung von Autographen und Nachlässen im digitalen Zeitalter. Probleme und Perspektiven". In *Bibliothek. Forschung und Praxis* 30.3: 283–289. DOI: 10.1515/BFUP.2006.283.

Heitfeld-Rydzik, Birgit, Ingeburg Abdul Wahed und Jens Brumann. 2022. *Nachlass Scholz / Sammlung Frege: Findlisten*. [Electronic ed.]. URN: urn:nbn:de:hbz:6-93009597846.

Kraft, Tobias und Stefan Dumont. 2020. „The Humboldt Code". In *Wiener Digitale Revue* 1: Tagebuch. DOI: 10.25365/WDR-01-03-02.

Mertens, Ina. 2021. „Zwei Seiten einer Medaille – IIIF und die Arbeit mit digitalen Bildbeständen". In *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel. DOI: 10.17175/2021_002.

Molendijk, Arie L. 2022. „The troubled Life of Heinrich Scholz". In *Journal for the History of Modern Theology / Zeitschrift für Neuere Theologiegeschichte* 29.2: 316–349. DOI: 10.1515/znth-2022-0016.

Nantke, Julia, Sandra Bläß und Marie Flüh. 2022. „Literatur als Praxis: Neue Perspektiven auf Brief-Korrespondenzen durch digitale Verfahren". In *Digitale Verfahren in der Literaturwissenschaft*, hg. von Jan Horstmann und Frank Fischer. Sonderausgabe #6 von Textpraxis. Digitales Journal für Philologie. DOI: 10.17879/64059432335.

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". In *Applied Sciences* 9.22. DOI: 10.3390/app9224853.

Sommer, Dorothea, Christa Schöning-Walter und Kay Heiligenhaus. 2008. „URN Granular: Persistente Identifizierung und Adressierung von Einzelseiten digitalisierter Drucke. Ein Projekt der Deutschen Nationalbibliothek und der Universitäts- und Landesbibliothek Sachsen-Anhalt". In *ABI Technik* 28.2: 106–114. DOI: 10.1515/ABITECH.2008.28.2.106.

Turing, Alan M. 1937. „On Computable Numbers, with an Application to the Entscheidungsproblem". In *Proceedings of the London Mathematical Society* s2-42.1: 230–265. DOI: 10.1112/plms/s2-42.1.230.

Wettlaufer, Jörg. 2018. „Der nächste Schritt? Semantic Web und digitale Editionen". In *Digitale Metamorphose: Digital Humanities und Editionswissenschaft*, hg. von Roland S. Kamzelak und Timo Steyer. DOI: 10.17175/sb002_007.

War das nicht schon immer ein Denkmal? Herausforderungen des Sammelns und Visualisierens von Denkmallisten in Zeit und Raum

Klemstein, Franziska
f.klemstein@gmail.com
Akademie der Wissenschaften und der Literatur, Mainz / Hochschule Mainz, Deutschland
ORCID: 0000-0003-3137-6732

1. Einleitung

Denkmalinventare und Denkmal-Topografien gibt es seit dem 19. Jahrhundert (Noell, 2020).

Bei den Topografien kann man jedoch kaum von einer umfassenden Erfassung sprechen, da die einzelnen Bände nur den Zeitpunkt des Denkmalinventars zu einem bestimmten Zeitpunkt abbilden können. Dennoch dienen gerade die Topografien der Vermittlung des Denkmalbestandes in Wort, Bild und Karte an eine breite Öffentlichkeit (Euler-Rolle, 2008; Klemstein, 2022). Sie sind maßgeblich für die Vermittlung und Erforschung des Denkmalbestandes verantwortlich. Während einige Landesämter für Denkmalpflege in Deutschland ihre aktuellen Bestände bereits in Form von Listen oder Kartenviewern zur Verfügung stellen (z.B. die Landesämter in Berlin[1] und Niedersachsen[2]), sind insbesondere frühere Denkmalinventare im deutschsprachigen Raum noch nicht als digitale Ressource für die Öffentlichkeit zugänglich gemacht worden (siehe Tabelle 1).

Vor diesem Hintergrund möchte ich untersuchen, wie insbesondere Kunsthistoriker:innen und Denkmalpfleger:innen digitale Ressourcen für Forschungszwecke nutzen können und welchen zusätzlichen Nutzen die Abbildung von Informationen in Raum und Zeit in diesem Zusammenhang haben könnte. Konkret geht es im folgenden Beitrag zunächst um die Frage, wie verschiedene analoge und digitale Quellen zum Denkmalbestand einer bestimmten Region über einen bestimmten Zeitraum zugänglich gemacht werden können, um sie mit digitalen Methoden zu analysieren, auszuwerten und zu visualisieren. Zu diesem Zweck wird die zunächst die Qualität der bereits digital vorliegenden Quellen untersucht und bewertet, um die Herausforderungen zu identifizieren und die Auswirkungen der Nutzung

dieser heterogenen Quellen und Datensätze im Forschungsprozess zu analysieren. Schließlich wird die Abbildung der Informationen selbst untersucht. Dieser letzte Abschnitt konzentriert sich auf die Frage: Welchen Einfluss hat die Visualisierung oder das geografische Mapping von Daten mit Python auf Forschungsprozesse? Dies soll dann anhand der Zusammenführung der Denkmalbestände aus Ost- und West-Berlin im Jahr 1995 exemplarisch erörtert werden.

2. Digitale Datenbestände und analoge Quellen

2.1 Datensammlungen

Tabelle 1. Öffentlich zugängliche Denkmallisten der Landesämter für Denkmalpflege in Deutschland (Stand: November 2022)

Bundesland	Öffentlich zugängliche Denkmalliste	Datenformat
Baden- Württemberg	-	
Bayern	+	PDF
Berlin	+	PDF, CSV, TXT, KML, RDF (Metadaten)
Brandenburg	+	PDF
Bremen	+	PDF
Hamburg	+	PDF
Hessen	-	
Mecklenburg-Vorpommern	-/+ (abhängig vo m Kreis bzw. Gemeinde)	PDF
Niedersachsen	-	
Nordrhein-Westfalen	-	
Saarland	+	PDF
Sachsen	-	
Sachsen-Anhalt	-	
Schleswig-Holstein	+	JSON, CSV
Thüringen	-	
Rheinland-Pfalz	-/+ (abhängig vom Kreis bzw. Gemeinde)	PDF

Die Tabelle zeigt, dass es nicht möglich ist, alle notwendigen Informationen zum Denkmalbestand in Deutschland über die Landesbehörden zu erhalten. Gleichzeitig wird deutlich, dass die Veröffentlichung der Liste als PDF-Datei ein gängiges Verfahren ist. Lediglich Berlin und Schleswig-Holstein veröffentlichen andere Formate und scheinen auf diese Weise die weitere Nutzung der Daten auch im Hinblick auf die Interoperabilität des Datenbestandes sicherstellen zu wollen. Bei den PDF-Dateien scheint es kein einheitliches Schema zu geben, so dass auch hier die Varianz hinsichtlich der Informationen zum Denkmalbestand sehr groß ist. Alle Bundesländer verzichten auf die Angabe von Geokoordinaten in ihren Listen. Allerdings existieren für den Berliner Denkmalbestand fünf verschiedene KML-Dateien, die in Baudenkmale, Denkmalbereiche, Ensembleteile, Gartendenkmale und Bodendenkmale unterteilt sind, wodurch keine direkte Vergleichbarkeit oder gemeinsame Analyse ermöglicht wird. In den Dateien werden die jeweiligen Denkmale angezeigt und besitzen, sozusagen als Metadaten-Angabe, die Denkmal-Objektnummer als Referenz. Allerdings existieren einige dieser Marker bzw. Objektnummern mehrfach, wodurch Uneindeutigkeiten entstehen (siehe Abb. 1). Zu vermuten ist, dass dies dann entsteht, wenn Denkmalbereiche bzw. Polygone zur Kartierung des Denkmalbestandes eingesetzt wurden.

Abb. 1. KML-Datei zu den Gartendenkmalen der Stadt Berlin. Sichtbar werden im unteren linken Bereich zwei Pins, die die Objektnummer 09046284 tragen.

Schleswig-Holstein ist nicht nur das einzige Bundesland, das seine Denkmaldaten auch im JSON-Format zur Verfügung stellt, sondern auch das einzige Bundesland, das auch Vorgängerlisten (ab April 2021) anbietet und damit Änderungen an den Denkmaldaten seit etwas mehr als einem Jahr dokumentiert.[3]

Auch wenn nicht alle Bundesländer ihre Denkmallisten veröffentlichen, gibt es in vielen Bundesländern Kartenviewer, in denen die Denkmäler eingesehen werden können. Hier sind die Daten jedoch meist weder exportierbar noch weiterverwendbar, so dass keine weitere Recherche mit diesen Daten möglich ist.

Neben den Landesämtern für Denkmalpflege in Deutschland gibt es eine unvollständige Sammlung von Denkmallisten der DDR, die online zugänglich ist. Diese werden als PDF-Dateien über ein Portal des Leibniz-Instituts für Raumbezogene Sozialforschung in Erkner zur Verfügung gestellt.[4]

Sowohl die aktuellen Denkmallisten als auch die Listen des Denkmalbestandes der DDR werden oft mit nur wenigen Metadaten zur Verfügung gestellt, was das Auffinden und Nachschlagen der entsprechenden Listen zusätzlich erschwert.

2.2 Daten sammeln und aufbereiten: Das Auffinden von weiteren Informationen und Quellen zum Denkmalbestand

Weitere Listen der Denkmalinventare in Ost- und Westdeutschland sind in verschiedenen Archiven (z.B. Staatsarchive, Archive der Denkmalbehörden) zu finden. Für das Denkmalinventar der DDR besitzen viele Stadt- und Staatsarchive die Regional- und Bezirkslisten. Diese sind jedoch zumeist noch nicht digitalisiert worden. Darüber hinaus gibt es auch Denkmaltopographien zum Denkmalbestand der DDR und der BRD, die den Bestand für bestimmte Regionen abbilden. Auch hierbei handelt es sich ausschließlich um analoge Quellen.

Wie die Pipeline für die Aufbereitung dieser Quellen und Informationen aussieht, soll im Rahmen des Vortrags konkret dargelegt werden.

3. Datenqualität und Methodologie

Wenn man sich die Daten und Quellen ansieht, fallen sofort drei Probleme auf:

1. veraltete Straßen- und Objektbezeichnungen, die eine eindeutige Zuordnung erschweren,
2. der allgemeine Mangel an Geodaten/Georeferenzierung und
3. die Herausforderung der Erfassung analoger Informationen sowie der Transparenz hinsichtlich der Provenienz der Daten.

Darüber hinaus stellen sich Fragen nach der Art und Weise, wie Geodaten modelliert und aufbereitet werden. Dazu gehört auch die Normierung oder Transformation von Koordinaten. In dem hier beschriebenen Fall wurde die Entscheidung für das dezimale Gradsystem getroffen. Bei fehlenden Koordinatendaten wurden die Raumbezugsinformationen durch Geokodierung mittels Python zugewiesen. Im Hinblick auf die GIS-Datenqualität wurde versucht, die entsprechende ISO-Norm 19157 einzuhalten.

Weitere Informationen zu Objekten, die den Topographien entnommen werden konnten, wurden zusätzlich in einer OpenDocument-Datenbank gespeichert.

Im Hinblick auf den methodischen Ansatz mag es überraschen, dass keine etablierte GIS-Software verwendet wird. Diese Entscheidung wurde jedoch ganz bewusst getroffen, da das Arbeiten in einer Python-Umgebung zum einen keine Softwareinstallation erfordert und somit ressourcenschonend und flexibel (auch im Hinblick auf verschiedene Betriebssysteme) einsetzbar ist. Zum anderen liegt der Schwerpunkt der hier vorgestellten Projektstudie auf der Visualisierung und Analyse der Darstellung von Geodaten und nicht auf der Erfassung und Verwaltung umfangreicher Datenbestände. Für die hier vorgestellten Bedürfnisse erweisen sich GIS-Systeme als zu komplex, da sie nicht nur die Erfassung, Verarbeitung, Analyse und Visualisierung, sondern auch die Verwaltung dieser Daten umfassen. Das Arbeiten mit Python ermöglicht zudem einfache Exportmöglichkeiten für die Visualisierungen, die auch eine unkomplizierte Einbettung in Webanwendungen erlauben. Darüber hinaus können verschiedene Formen der Visualisierung schnell implementiert und je nach Fragestellung leicht angepasst und neu visualisiert werden. Die Programmbibliothek folium baut auf den Stärken des Python-Ökosystems in der Datenverarbeitung und den Stärken der Leaflet.js-Bibliothek im Mapping auf und macht es einfach, Daten in einer Leaflet-Map zu manipulieren. Neben der Darstellung von Daten in Form von Punkten (Markern) können auch Polygone erstellt werden.

Das wichtigste Argument für den Einsatz einer Python-Entwicklungsumgebung bzw. die Verwendung der Programmbibliothek folium gegenüber dem Einsatz einer Softwarelösung ist jedoch die damit verbundene stärkere (erzwungene) Einbindung in den technischen Prozess der Kartierung. Bei der Erstellung der Karte mit ihren jeweiligen Inhalten und Informationen müssen alle Einzelschritte im Skript entsprechend spezifiziert werden, wodurch eine ständige Hinterfragung der Kartenerstellung möglich und notwendig wird. Dies führt unweigerlich zu einer theoretischen Reflexion der Kartenerstellung sowie der Datenerhebung.

Abschließend sollen hierbei auch generalisierbare Aspekte des Gesamtprojektes thematisiert werden, um die Anschlussfähigkeit aufzuzeigen, die insbesondere für Denkmalbehörden einen Mehrwert bilden könnte.

4. Visualisierung anhand des Fallbeispiels „Berlin“

Die Erstellung von Visualisierungen und das Mapping von Objekten ist mit zahlreichen Überlegungen und Entscheidungen verbunden. In diesem Zusammenhang erscheint es sinnvoll, nicht unbedingt auf leistungsfähige Softwarelösungen zurückzugreifen, sondern - je nach Problemstellung und Datenlage - mit Python-Programmbibliotheken wie folium zu arbeiten. Dies bietet den Vorteil, dass sowohl die Modellierung der Daten als auch die Darstellungs- und Auswertungsmöglichkeiten Schritt für Schritt aktiv geplant werden können.

Die folgenden Abbildungen (siehe Abb. 2) und (siehe Abb. 3) sind das Ergebnis der schrittweisen Modellierung und Visualisierung des Denkmalbestandes der DDR. Ausgangspunkt hierfür waren die Zentrale Denkmalliste der DDR sowie die Bezirksliste für Berlin (hier ein Ausschnitt von Friedrichshain). Während die Zentrale Denkmalliste der DDR als PDF-Datei über das Portal des Leibniz-Instituts für Gesellschafts- und Raumforschung in Erkner zugänglich ist. Während die Zentrale Denkmalliste der DDR als PDF-Datei über das Portal des Leibniz-Instituts für Gesellschafts- und Raumforschung in Erkner zugänglich ist, basieren die Daten zu den Denkmälern der Bezirksliste auf der Denkmaltopographie "Bau- und Kunstdenkmale der DDR" und mussten zunächst von der analogen Quelle in ein digitales Format übertragen werden.

In einem weiteren Schritt werden diese Karten mit dem Denkmalbestand Berlins aus dem Jahr 1995 (dem Jahr der Zusammenführung der Denkmalbestände aus Ost- und West-Berlin) sowie dem aktuellen Denkmalbestand Berlins (Stand: 05.05.2021) zusammengeführt und durch verschiedene Layer in unterschiedlichen Kombinationen visualisierbar gemacht (siehe Abb. 4).

Dabei fanden unterschiedliche Kartenstile und Marker Verwendung, um verschiedene Visualisierungsmethoden zu testen und zu hinterfragen sowie weitere Informationen durch Tooltips, Popups oder unterschiedliche Farben der Marker und die Integration von Legenden zu liefern.

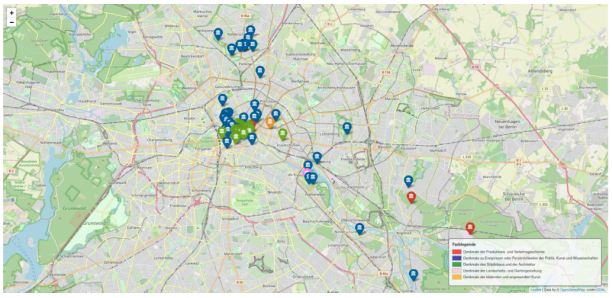


Abb. 2. Denkmalbestand von Ost-Berlin, Zentrale Denkmalliste der DDR.

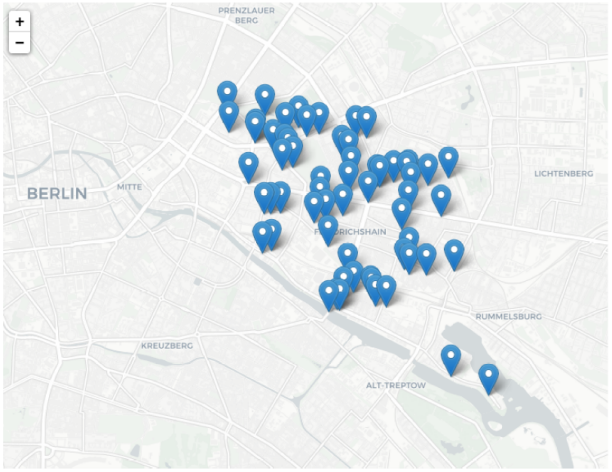


Abb. 3. Denkmalbestand der Kreisliste Friedrichshain (DDR).

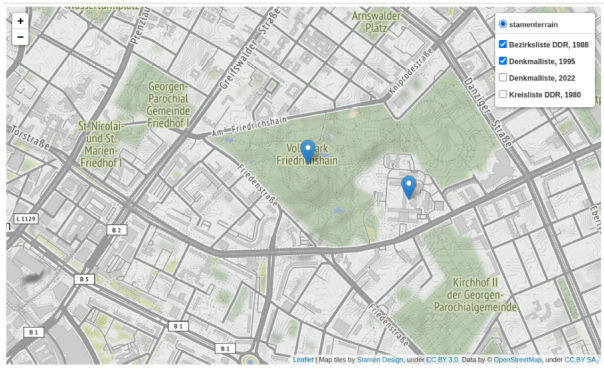


Abb. 4. Denkmalbestand Berlin mit verschiedenen Layermöglichkeiten, die verschiedene Denkmallisten zu unterschiedlichen Zeitpunkten visualisieren.

5. Zusammenfassung & Ausblick

Darstellungen des aktuellen Erbes auf Karten finden sich inzwischen in großer Zahl. Bislang fehlt jedoch die zeitliche Komponente, so dass der Eindruck entstehen kann, Denkmalkataster seien ein statisches oder unveränderliches Produkt.

Die hier vorgestellten Visualisierungen des Denkmalbestandes ermöglichen nicht nur einen Überblick über den Denkmalbestand zu verschiedenen Zeiten, sondern zeigen auch, welche Denkmäler der DDR bei der Überarbeitung der Denkmalliste nach dem Zusammenbruch des DDR-Systems übernommen wurden. Darüber hinaus lassen sich durch solche Visualisierungen und die Aufbereitung von Datensätzen zu Denkmalbeständen zu verschiedenen Zeiten auch generelle Veränderungen im Denkmalbestand (z.B. in Bezug auf bestimmte Zeiträume oder Bautypen) erfassen, analysieren und hinterfragen. Sie ermöglichen es, Wandlungen, Umbrüche oder auch Verschiebungen des "Zeitgeistes" in der Praxis der Denkmalinventarisierung sichtbar zu machen, die in der bisherigen Forschung kaum berücksichtigt wurden, und können so die Grundlage für weitere Forschungen bilden.

In Bezug auf das Fallbeispiel Berlin kann durch diese Aufbereitung und Analyse der Daten aufgezeigt werden, dass nicht nur immer wieder neue Objekte in den Status eines Denkmals erhoben werden, sondern auch, dass insbesondere in gesellschaftlichen Transformationszeiten neue Aushandlungsprozesse aktiviert werden, die den Denkmalbestand gravierend verändern können. Wie wichtig diese Art der Datenaufbereitung, Modellierung und Visualisierung ist, zeigen im internationalen Kontext Projekte wie SurveyLA[5] und HistoricPlacesLA[6] (Avrami, 2019).

Fußnoten

1. https://www.berlin.de/landesdenkmalamt/denkmale/liste-karte-datenbank/denkmalliste/, letzter Zugriff: 17.07.2023.
2. https://denkmalatlas.niedersachsen.de/viewer/liste/, letzter Zugriff: 19.07.2023.
3. https://opendata.schleswig-holstein.de/dataset/denkmalliste, letzter Zugriff: 17.07.2023.
4. Denkmallisten der DDR, zur Verfügung gestellt durch das Leibniz Institut für Raumbezogene Sozialforschung in Erkner, http://ddr-planungsgeschichte.de/denkmallisten/, letzter Zugriff: 15.12.2022.
5. SurveyLA, https://planning.lacity.org/preservation-design/historic-resources-survey, letzter Zugriff: 14.12.2022.
6. HistoricPlacesLA, http://www.historicplacesla.org/, letzter Zugriff: 14.12.2022.

Bibliographie

Noell, Matthias. 2020. Wider das Verschwinden der Dinge: Die Erfindung des Denkmalinventars, Wasmuth & Zohlen Verlag, Berlin.

Euler-Rolle, Bernd. 2008. „Inventarisation und öffentliches Interesse“ In Sozialer Raum und Denkmalinventar, 10-15. Sandstein, Dresden.

Klemstein, Franziska. 2022. „Diversität und Denkmalpflege: Zwischen analog und digital“ In Avantgarde oder uncool? Denkmalpflege in

der Transformationsgesellschaft, 128-135. Mitzkat, Holzminden.

Avrami, Erica. 2019. „Preservation and the New Data Landscape“, Columbia University Press, New York.

Zusammenführung audiovisueller Ressourcen für tanz- und theaterwissenschaftliche Forschung Mediatheken der Darstellenden Kunst digital vernetzen

Beck, Julia
J.Beck@ub.uni-frankfurt.de
Universitätsbibliothek Johann Christian Senckenberg, Deutschland

Henniger, Christine
c.henniger@iti-germany.de
Internationales Theaterinstitut Deutschland / Mediathek für Tanz und Theater, Deutschland

Illmayer, Klaus
klaus.illmayer@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich
ORCID: 0000-0001-7253-996X

Tiefenbacher, Sara
S.Tiefenbacher@ub.uni-frankfurt.de
Universitätsbibliothek Johann Christian Senckenberg, Deutschland
ORCID: 0000-0002-6300-4805

Voß, Franziska
F.Voss@ub.uni-frankfurt.de
Universitätsbibliothek Johann Christian Senckenberg, Deutschland

Wittenbecher, Maxim
m.wittenbecher@iti-germany.de
Internationales Theaterinstitut Deutschland / Mediathek für Tanz und Theater, Deutschland

Digitale Verfahren in der Theaterwissenschaft bzw. in der Forschung zu darstellenden Künsten (“performing arts”) gewinnen in den letzten Jahren an Fahrt (Varela 2021). Trotzdem tut sich das Fach sichtlich schwer, einen eigenen Zugang zu entwickeln, der anschlussfähig ist an laufende Digital Humanities-Debatten. Obwohl seit der Gründungsphase der wissenschaftlichen Disziplin in den 1920er Jahren im deutschsprachigen Raum viel Material angehäuft wurde - das oft noch einer Digitalisierung harrt - und eine große Breite an methodischen Zugängen entwickelt wurde, fehlen gemeinsame Vereinbarungen zu Standards, Metadatenformaten und Datenmodellen (Probst et al. 2020). Zwar gibt es mit dem Fachinformationsdienst für Darstellende Kunst (FID DK, Voß 2017) eine Institution, die Metadaten zu theaterwissenschaftlichen (Archiv-)Material aus einer Vielzahl an Sammlungen zusammenführt und in einem zentralen Suchportal zur Verfügung stellt, aber es fehlen größere Initiativen aus der Forschung, die digitale theaterwissenschaftliche Verfahren ent wickeln und auf dieses Material anwenden. Was überrascht, da die Auseinandersetzung mit den digitalen Umbrüchen in den Untersuchungsgebieten des Faches - Theater, Tanz, Zirkus und alle weiteren künstlerischen performativen Ausdrucksformen - durchaus vorhanden sind und reflektiert werden (beispielhaft Dixon 2007, Blake 2014). Auch die künstlerischen Institu tionen selbst setzen sich zuweilen intensiv mit neuen oder veränderten Ausdrucksformen auseinander, in denen digitale Aspekte eingebunden werden, z.B. VR-/AR-Experimente (Heinrich-Böll-Stiftung et al. 2020). Nichtsdestotrotz zeigt eine Auswertung der DHd-Abstracts der letzten Jahre, dass Beiträge zu theaterwissenschaftlichen Fragestellungen - abseits der mehr den Literaturwissenschaften zuzuordnenden Dramenforschung - spärlich gesät sind. Eine mögliche Erklärung wäre das Fehlen von spezifischen, digital verfügbaren Untersuchungsmaterialien wie audiovisuellen (AV) Ressourcen, welche aktuelle Forschungs - trends im Fach adäquat bedienen können. Tatsächlich zeigt sich hier eine Lücke, da es zwar viel historisches Archivmaterial gibt, aber zu aktuellen Produktionen der darstellenden Künste schon aus Gründen der Verwertungsrechte nur wenige frei verfügbare Aufzeich nungen vorliegen. Nachdem eines der wichtigsten Elemente in der theaterwissenschaftlichen Forschung die Aufführung ist, die in Ko-Präsenz mitverfolgt und über die in weiterer Folge reflektiert wird (Fischer-Lichte 2004), ist die Nachvollziehbarkeit bei Nicht-Teilnahme auf den passiven Konsum von Aufzeichnungen reduziert. Weil große Theater und Festivals meist Aufnahmen verbieten, ist die Forschung auf das Verfügbarmachen durch diese Institutionen angewiesen. Zugleich gibt es keine gute Möglichkeit, zu suchen, wo Aufzeichnungen von Aufführungen zu welchen Bedingungen vorhanden wären. Im Vergleich zur Filmwissenschaft ist dies eine unzufriedenstellende Situation, wo diese doch über umfangreiche Datenbanken mit detaillierten Informationen zu Filmen (z.B. IMDb) und einem gut funktionierenden Distri butionsnetzwerk für den Bezug des Anschauungsmaterial verfügt, sei es durch Kauf, Leihe oder der eigenen Fernsehaufzeichnung. Ähnliches lässt sich für die Verfügbar-

keit von AV-Ressourcen zu Aufführungen der darstellenden Künste nicht behaupten.

Hier setzt das im Vortrag präsentierte Projekt “Mediatheken der darstellenden Kunst vernetzen” (mv:dk) [1] an, das sich zum Ziel gesetzt hat, ein Portal für die Identifikation und Auffindbarkeit von Aufzeichnungen zu Aufführungen prototypisch zu entwickeln. Wobei eine große Bandbreite digitaler Techniken aufgeboten wird, um dies umzusetzen: von der Datenaggregation und -anreicherung, über die Ausarbeitung eines umfangreichen Daten modells bis zur Möglichkeit für Forscher:innen, Analysen und Auswertungen zu den gesam melten Metadaten durchzuführen. Es soll damit nicht nur der Zugang zu Aufzeichnungen von Aufführungen erleichtert werden - insbesondere auch im Sinne der FAIR-Prinzipien -, sondern auch eine Datensammlung bereitgestellt werden, an der Verfahren der Digital Humanities angewandt werden können. In unserem Vortrag werden wir nicht nur die Genese des Projektes beschreiben und die abschließende Webplattform vorstellen, sowie verschiedene Szenarien zur Auswertung skizzieren, sondern auch die dabei aufgetretenen Probleme und Schwierigkeiten aufzeigen. Zusätzlich wollen wir im Sinne des Tagungsthemas “Quo Vadis” anhand des Projekts mv:dk beispielhaft auf den aktuellen Stand einer digitalen Tanz- und Theaterwissenschaft eingehen und einige Herausforderungen für die Digital Humanities aus dem Fach aufzeigen.

Das vom DFG im Bereich "Wissenschaftliche Literaturversorgungs- und Informationssysteme" (LIS) geförderte mv:dk-Projekt wurde im April 2021 begonnen und endet Anfang nächsten Jahres. Es ist eine Kooperation zwischen dem FID DK und dem Internationalen Theaterinstitut Deutschland, Berlin/Mediathek für Tanz und Theater. In einem Posterbeitrag für die DHd2022 wurden das Projekt und die angestrebten Ziele bereits vorgestellt (Illmayer et al. 2022). Nun sollen abschließend die Resultate und Erfahrungen präsentiert werden, um im Projekt erfolgte Standardisierungsbestrebungen aufzuzeigen sowie Erkenntnisse aus der Zusammenführung von Daten von den datengebenden Institutionen und Potentiale für Datenauswertungen durch Forscher:innen aufzuzeigen.

Bereits seit vielen Jahrzehnten werden an tanz- und theaterwissenschaftlichen Institutionen - neben Theatermuseen, Theaterhäusern sowie Künstler:innen selbst - audiovisuelle Aufzeichnungen von Aufführungen gesammelt (bspw. Fuxjäger 2020). Oft sind diese Sammlungen als Mediatheken innerhalb eines Instituts organisiert und dienen den Zwecken von Forschenden, Lehrenden und Studierenden. Eine intensive Auseinandersetzung und Analyse eines performativen Ereignisses ist ohne solche Hilfsmittel schwer möglich, auch wenn diese nicht das jeweilige Live-Erlebnis vollständig ersetzen können. Diese AV-Aufzeichnungen sind zugleich historisches Material, das auch für andere Fächer von Bedeutung sein kann, an dem nicht nur ästhetische Fragestellungen herangetragen, sondern auch prägende Diskurse einer zeitlichen Epoche abgelesen werden können. Für das mv:dk-Projekt konnten eine Reihe solcher tanz- und theaterwissenschaftlicher Mediatheken als Partner:innen gewonnen werden, die sich bereit erklärten, ihre Metadaten in ein zentrales Suchportal einzuspielen. Wobei sich auch zeigte, dass die Sammlungen unterschiedliche und sich verändernde Erfassungsstrategien und -systematiken verfolgten. So liegt der Fokus mancher Mediatheken inzwischen stärker auf der Sammlung von Filmen und Fernsehserien, insbesondere wenn sich dort das Fachgebiet Theaterwissenschaft auf Film-, Fernseh- und Medienwissenschaft ausgeweitet hat. Schließlich ist die Verfügbarkeit für solche Materialien deutlich besser gegeben, als es bei Tanz- und Theateraufzeichnungen der Fall ist. Aufbauend auf eine Umfrage vor Projektstart, war bereits bekannt, dass es zwar ein Bewusstsein über das Vorhandensein solcher Mediatheken bei den verschiedenen theaterwissenschaftlichen Instituten gab, aber welche Materialien jeweils dort liegen oder wie der Zugriff für Forscher:innen darauf ermöglicht werden könne, war und ist oft unklar. Ganz zu schweigen von der Möglichkeit, das Videomaterial selbst betrachten zu können, ohne jeweils vor Ort reisen zu müssen. Rechtliche Unsicherheiten bedingen und verschärfen diese für Forscher:innen unzufriedenstellende Situation.[2]

Für das Projekt mv:dk galt es somit, zunächst ein konsolidiertes Datenmodell zu erstellen, wobei zunächst in den vorzufindenden heterogenen - oft selbst entwickelten - Datenbanken die Gemeinsamkeiten identifiziert wurde, bevor weitere Felder hinzugefügt wurden, die für den Anreicherungsprozess sinnvoll sind sowie eine Verbindung zu weithin anerkannten Datenschemata herzustellen erlauben. Ein besonderes Augenmerk wurde dabei auf das Europeana Data Model (EDM) gelegt, da dieses am FID DK für das Suchportal Verwendung findet und im Sinne eines nachhaltigen Datenmanagement die Integration einer Teilmenge der gesammelten Daten aus mv:dk in dieses Suchportal vorgesehen ist. Zugleich war das Projektteam bemüht, darüber hinaus eine weitreichende Verlinkung zwischen den importierten Datensätzen herzustellen. Damit soll besser erkennbar werden, in welchen Mediatheken verschiedene Fassungen von Aufführungen einer Inszenierung zu finden sind. Gerade dieser Vergleich ist für die theaterwissenschaftliche Forschung von großem Interesse. Die Möglichkeit, eine Aufführung aus verschiedenen Perspektiven - weil an verschiedenen Orten zu unterschiedlichen Zeiten im Rahmen einer Tournee Aufzeichnungen vorliegen - zu analysieren, ergibt sich sonst nur durch einen großen Mehraufwand. Da solche Fassungen an jeweils anderen Mediatheken liegen können, ohne dass diese darüber Bescheid wissen, war von Beginn des Projekts an ein Erkennen solcher Überschneidungen ein dezidiertes Ziel. Damit ergibt sich auch ein komplexeres Bild auf das Datenmodell, da nicht nur die einzelnen Aufführungen sondern auch die übergeordnete Kategorie der Inszenierung und - so vorhanden - der Bezug auf Werke hergestellt werden soll, um in weiterer Folge Fassungen zu identifizieren aber auch unterschiedliche Einstiegspunkte für einen Sucheinstieg zu geben. So ist vorstellbar, dass Literaturwissenschaftler:innen mehr daran interessiert sind, Aufführungen zu einem Werk zu analysieren, während Tanzwissen schaftler:innen die ästhetische Entwicklung einer Company anhand der verfügbaren Aufzeichnungen nachvollziehen möchten. Aus diesem Grund

fiel die Entscheidung, im Portal von mv:dk nativ mit Linked Data zu arbeiten. Da in einigen theaterwissenschaftlichen Projek ten zudem mit CIDOC CRM, FRBRoo und WikiData modelliert wird (z.B. Lee 2018, Estermann 2020, Weiberg 2020), wurde zum Zwecke der Anreicherung und Vernetzung eine Ontologie entwickelt, die entsprechende Anknüpfungspunkte - sei es durch Mapping oder durch direkte Integration in diese mv:dk-Ontologie - ermöglicht. Die mv:dk-Ontologie ist ein zentrales Ergebnis des Projekts und wird bis spätestens Ende 2023 in der neuesten Fassung auf dem GitHub-Repositorium veröffentlicht.

Die Erstellung der Ontologie und damit verknüpft die Ausarbeitung projektspezifischer kontrollierter Vokabularien war zudem die Voraussetzung, die digitale Infrastruktur des Projektes aufzubauen. Dabei muss zusätzlich berücksichtigt werden, dass es verschiedene Zugriffsstufen auf die Suchplattform geben muss, da manche Materialien nur für Forscher:innen einzusehen sind. Die Infrastruktur umfasst inzwischen eine Ingest-Pipeline, mit der Daten auf unterschiedlichsten Wegen importiert, auf das Datenmodell von mv:dk gemapped und anschließend angereichert werden, bspw. mit Identifikatoren zur GND. Des weiteren dient als zentrale Datenbank ein RDF-Triplestore (GraphDB), womit auch SPARQL-Abfragen ermöglicht werden. Dem zur Seite gestellt ist eine Backend-Komponente, die eine API implementiert (Python/Django) um den Zugriff auf den Triplestore zu ermöglichen und mittels eines AAI-Layers (Authentication and Authorization Infrastructure) die Zugriffsstufen berücksichtigt. Schließlich noch das Frontend, welches in Vue3/Typescript entwickelt wurde und das Suchen sowie Navigieren in den vorhandenen Metadaten ermöglicht. Im ersten Teil des Vortrags werden alle diese Komponenten vorgestellt und wir erläutern, wie sie zusammenwirken, welche Herausforderungen zu bewältigen waren und welche Überlegungen zu dieser Infrastruktur führten. Wichtige weitere Themen dazu sind Skalierbarkeit, Nachhaltigkeit und Interoperabilität dieser Lösung. Zugleich werden einige Workflows erläutert, wie das Onboarding neuer Datensammlungen, die Einbindung der Zugriffsstufen und die Schritte, um von den in der Plattform ersichtlichen Metadaten zu den AV-Aufzeichnungen zu gelangen, die nicht Teil der mv:dk-Plattform sind und bei den Datenpartner:innen vor Ort liegen.

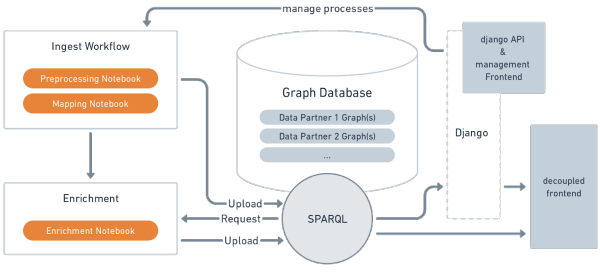


Illustration 1: Infrastruktur und Workflows der mv:dk-Plattform

Für den zweiten Teil des Vortrags möchten wir auf Potentiale für Forscher:innen eingehen, abseits des Servicecharakters der Plattform. Wir zeigen dabei auf, wie die Datenqualität von Interaktionen mit Forscher:innen profitieren kann. Danach gehen wir darauf ein, wie die Zusammenführung von Daten aus unterschiedlichen Institutionen interessante Einblicke in Sammlungsstrategien und somit auch in Schwerpunktsetzungen von Instituten zu unterschiedlichen Zeiten offenbaren. Zugleich können damit Kanonisierungstendenzen in der Tanz- und Theaterwissenschaft überprüft werden, die auch durch das (Nicht-)Vorhandensein von Aufzeichnungen begründet sein können. Besonders die Lücken in den Sammlungen, die durch die Zusammenführung noch deutlicher akzentuiert werden, können zu Reflexionen über die Themenvielfalt in der Forschung zu darstellenden Künsten anregen. Wobei hier auch von Interesse ist, welche weiteren Informationen von den Datenpartner:innen geeignet wären, um noch bessere Ergebnisse zu erzielen. Schließlich ist es auch ein Anliegen des mv:dk-Projektes, Forscher:innen dazu einzuladen, die Datensammlung auf vielfältige weitere Aspekte zu untersuchen, wozu wir einige Vorschläge geben werden. Die Möglichkeit, mit SPARQL-Anfragen gezielt nach Mustern zu suchen oder die vorhandenen Metadaten mit weiteren Datensammlungen zu verknüpfen - insbesondere wären hier Annotationen von AV-Aufzeichnungen zu erwähnen (de-Lahunta et al. 2021) -, sind auch für DH-Forscher:innen abseits der Theater- und Tanzwissenschaft von Interesse. Abschließend soll noch auf den Beitrag solcher digitalen Zusammenführungen von heterogenen Sammlungen und darauf angewandte digitale Verfahren für die Gedächtnispflege in den darstellenden Künsten eingegangen werden.

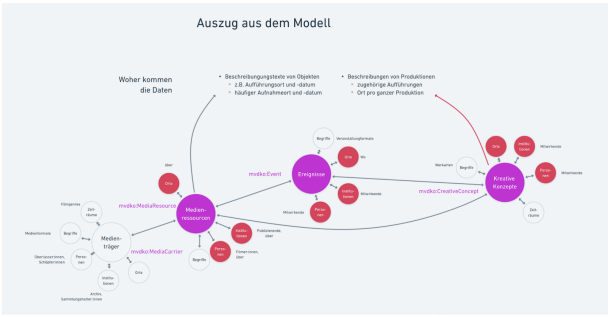


Illustration 2: Datenmodellauszug mit möglichen Recherchepfaden

Fußnoten

1. Das Projektteam umfasst neben den Autor:innen dieser Einreichung noch die studentischen Hilfskräfte Yannis Illies, Annika Müller und Marius Pahl. Ein Projektblog mit weiteren Informationen findet sich unter .
2. Für die Fragestellungen, die sich im Projektverlauf von Seiten der Projektpartner:innen ergeben haben, wurde eine “Handreichung ‘Mediatheken der Darstellenden Kunst und Recht’” von Prof. Dr. Paul Klimpel eingeholt,

siehe dazu den Blogeintrag und die Handreichung selbst als PDF-Datei:

Bibliographie

Blake, Bill . 2014. Theatre & the Digital. Houndmills . Basingstoke: Palgrave Macmillan.

deLahunta, Scott, David Rittershaus und Rebecca Stancliffe (Hg.). 2021. Digital Annotation and the Understanding of Bodily Practices. International Journal of Performance Arts and Digital Media 17/1. https://www.tandfonline.com/toc/rpdm20/17/1 (zugegriffen: 19. Juli 2023).

Dixon, Steve . 2007. Digital Performance. A History of New Media in Theater, Dance, Perfor mance Art, and Installation . Cambridge, MA: MIT Press.

Estermann, Beat . 2020. “Creating a Linked Open Data Ecosystem for the Performing Arts (LODEPA).” In Arti dello Spettacolo/Performing Arts Special , 31–49.

Fischer-Lichte, Erika . 2004. Ästhetik des Performativen . Frankfurt/Main: Suhrkamp.

Fuxjäger, Anton . 2020. "Die wissenschaftliche Videothek des Instituts für Theater-, Film- und Medienwissenschaft an der Universität Wien: Geschichte, Organisation, Technik". https://tfm.univie.ac.at/sammlungen-einrichtungen/videothek/hintergrundinformationen/ (zu gegriffen: 19. Juli 2023).

Heinrich-Böll-Stiftung, nachtkritik.de (Hg.). 2020. Netztheater. Positionen, Praxis, Produk tionen , Berlin. https://www.boell.de/de/netztheater (zugegriffen: 19. Juli 2023).

Illmayer, Klaus , Sara Tiefenbacher, Franziska Voß, Julia Beck, Christine Henninger und Maxim Wittenbecher. 2022. “Mediatheken der Darstellenden Kunst digital vernetzen.” In DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts , hg. von Michaela Geierhos, 325-326 10.5281/zenodo.6304590 .

Lee, Deborah . 2018. “Documenting Performance and Contemporary Data Models. Posi tioning Performance within FRBR and LRM”. In Proceedings from the Document Academy 5/1. https://ideaexchange.uakron.edu/docam/vol5/iss1/2/ (zugegriffen: 19. Juli 2023).

Probst, Nora und Vito Pinto . 2020. “Re-Collecting Theatre History. Theaterhistoriografische Nachlassforschung mit Verfahren der Digital Humanities”. In Neue Methoden der Theaterwis senschaft , hg. von Benjamin Wihstutz und Benjamin Hoesch. Bielefeld: Transcript, 157-179.

Varela, Miguel Escobar . 2021. Theater as Data. Computational Journeys into Theater Research . Ann Arbor, Michigan: University of Michigan Press 10.3998/mpub.11667458 .

Voß, Franziska . 2017. “Der Fachinformationsdienst Darstellende Kunst.” In Die vierte Wand. Organ der Initiative TheaterMuseum Berlin e.V. , Berlin: Initiative TheaterMuseum Berlin e.V., Heft 7, 2017, S. 62-63.

Weiberg, Birk . 2020. “Modeling Performing Arts. On the Representation of Agency” In Arti dello Spettacolo/Performing Arts Special , 50–56.

Doctoral Consortium

Automatische Erkennung von Bezügen zwischen Epistolographie und Literatur

Göggelmann, Michael
michael.goeggelmann@smail.uni-koeln.de
Universität zu Köln / Universität Tübingen, Deutschland

Einleitung und Fragestellung

Die Epistolographie von Schriftstellerinnen und Schriftstellern tritt vor dem eigentlichen literarischen Werk naturgemäß eher in den Hintergrund. Dabei können einige Briefsammlungen sowohl quantitativ als auch hinsichtlich ihrer ästhetischen Tragweite als „Werk neben dem Werk“[1] geltend gemacht werden. Die hohe inhaltliche Varianz und das Auftreten in bisweilen großen, formal konstanten Datenmengen prädestiniert das Briefwerk als Untersuchungsgegenstand für quantitativ-computergestützte Zugriffe, die den methodischen Rahmen dieses Projekts bilden sollen. Im Arbeitsvorhaben soll möglichen Bezügen zwischen epistolaren und literarischen Werken nachgespürt werden, die sich - je in Briefen reflektiert – beispielsweise im Kontext von Schaffens- und Publikationsprozessen, der literarischen Verarbeitung alltäglicher Erfahrungen oder der Reaktion auf externes Feedback offenbaren können. Mit dem Ziel einer stärkeren Vernetzung von literarischen und epistolaren Werken eines Autors, ergeben sich die folgenden Forschungsfragen:

1. Lassen sich Bezüge mit Hilfe maschineller Lernverfahren automatisiert erfassen?
2. Welche Methoden eignen sich dazu; welche hingegen nicht?
3. Wie lässt sich ein Modell zur automatischen Erkennung so generalisieren, dass es auch autorenübergreifend angewendet werden kann?

Das Projekt verspricht in zweifacher Hinsicht Innovationspotenzial: neben einem Beitrag zur Entwicklung quantitativer Methoden der Textanalyse soll die Beantwortung solcher literaturwissenschaftlicher Forschungsfragen vereinfacht oder ermöglicht werden, die eine stärkere Verknüpfung von Briefen und fiktionalen Werken voraussetzen.

Positionierung innerhalb der DH

Die Digitalisierung und Edition von Briefkorpora rückte schon früh in das Blickfeld computergestützter Geisteswissenschaft (Cheney, 1983). Auch aktuelle Projekte zur digitalen Briefedition knüpfen häufig an ältere Editionsprojekte an, die nach Jahrzehnten analogen Arbeitens um ein Digitalisierungsvorhaben ergänzt wurden.[2] Auf Basis nachträglich digitalisierter Korpora, insbesondere aber im Rahmen neuerer Projekte zu digitalen Briefeditionen werden Tools und Methoden (weiter-)entwickelt, denen es - primär auf Briefmetadaten (d.h. mitunter Sender, Empfänger, Ziel- und Ursprungsort sowie Datumsangaben) zurückgreifend – gelingt, neue Perspektiven auf bisweilen bereits erforschte Korpora zu schaffen.[3]

Computergestützte Projekte zu Briefsammlungen verbindet somit, dass sie als dezidierte Editionsprojekte zumeist spezifisch-epistolare Fragestellungen zur Aufbereitung eines Datensatzes behandeln und sich computergestützter Analysen vorrangig zur Visualisierung vorhandener, oder der Generierung neuer Metadaten bedienen.[4] Hier sollen mit dem Arbeitsvorhaben daher neue Wege gegangen werden, indem in einer gemeinsamen Analyse epistolarer und literarischer Korpora unter Verwendung quantitativer Methoden primär der *Briefinhalt* fokussiert wird.

Methodische Schnittpunkte hingegen ergeben sich insbesondere mit Projekten, die sich computergestützt mit (Teil-)Aspekten der Intertextualität befassen. Hierzu gab es auch in den vergangenen Jahren immer wieder Beiträge bei der DHd (u.a. Liebl und Burkhardt, 2020).

Daten

Das Arbeitsvorhaben stützt sich zunächst auf das digitale Briefkorpus von Charles Dickens, das etwa 14.000 Briefen umfasst und auf der zwölfbändigen Pilgrim-Edition seiner Briefsammlung basiert (House et al., 2001).

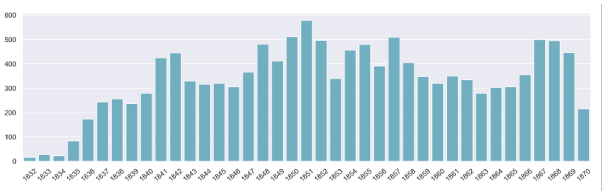


Abb. 1: Anzahl Briefe pro Jahr

Das im Projekt bislang noch nicht weiter untersuchte Teilkorpus der literarischen Werke von Dickens wurde aus dem Gutenberg-Projekt zusammengestellt.[5] Später soll die Methode an weiteren Autoren erprobt werden; autorenübergreifende Korpora könnten auf Basis von Epochen/Genres erstellt werden.

Methode

Als erster Orientierungspunkt für die automatische Erkennung von Bezügen dient die methodische Zweiteilung in eine vorangestellte *Reference Detection*, d.h. eine Erkennung von *spans* in den Briefen , die als Literaturbezug in Frage kommen, und deren anschließende, in einem zweiten Schritt erfolgende Zuordnung zu einem literarischen Werk des Autors.

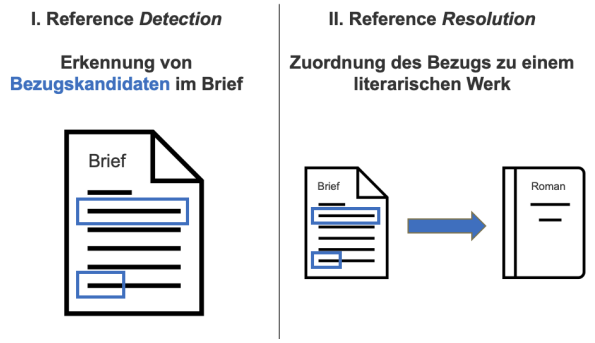


Abb. 2: Verfahren zur automatischen Erkennung von Bezügen

Hinsichtlich der *Reference Detection* soll auf möglichst generische, d.h. autorenunabhängige Eigenschaften der Bezüge zurückgegriffen werden. Hierbei wird vor allem mittels manueller Annotationen ein erster Überblick über Erkennungsmerkmale dieser Bezüge gewonnen. Als Kandidaten zeigen sich bislang

- die Verwendung von Anführungszeichen und Zitate generell
- eine höher frequentierte Verwendung von Eigennamen
- wiederkehrende Textbausteine, sowohl im Abgleich mit dem Literaturkorpus als auch zwischen den einzelnen Bezügen in den Briefen

Daraus ergeben sich für die *Reference Detection* bislang methodische Anleihen aus der *Quotation Detection*, *Named Entity Recognition* sowie *Text Reuse*.

Während nicht auszuschließen ist, dass sich der Schritt der *Reference Resolution* mitunter methodisch mit der oben dargestellten Kandidatenerkennung überschneiden wird, liegt das Augenmerk hier vor allem auf inhaltlichen Faktoren: *Welche* Eigennamen werden verwendet? Können diese mit Hilfe von Methoden des *Entity Linking* über Wissensdatenbanken literarischen Werken zugeordnet werden? Lassen sich über *Embeddings* modellierte Ähnlichkeitswerte zwischen Referenzkandidaten und literarischen Werken zur Zuordnung verwenden? Diese und weitere Fragen dienen als Orientierung für anstehende Versuche und Analysen.

Fußnoten

1. So etwa verorten Irmgard Wirtz und Alexander Honold das umfangreiche Briefwerk Rilkes (Honold und Wirtz, 2019, 7.).
2. Ein prominentes Beispiel ist das jüngst abgeschlossene „Darwin Correspondence Project“ (Burkhardt und Secord, 1985-2023).
3. Aus dem deutschsprachigen Raum s. Trier Center for Digital Humanities, 2023; Nantke et al., seit 2021 u.a. sowie international Hotson und Lewis, seit 2009; Orchard et al., seit 2008.
4. Vorhandene Metadaten werden meist als Netzwerke visualisiert; unter die Generierung neuer Metadaten fällt etwa die automatische Extraktion von Schlagwörtern.
5. Teil des Korpus sind die 14 veröffentlichten Romane sowie weitere Erzählprosa, wie etwa die Weihnachtsgeschichten.

Bibliographie

Cheney, David R. 1983. „Advantages and Problems of Editing Letters on the Computer.“ In *Sixth International Conference on Computers and the Humanities*, 89–93.

Burkhardt, Frederick und James A. Secord (Hg.) unter Mitarbeit der Editoren des Darwin Correspondence Project. 1985–2023. *The Correspondence of Charles Darwin*. https://www.darwinproject.ac.uk/

Honold, Alexander und Irmgard M. Wirtz. 2019. „Rilkes Korrespondenzen: Das Briefwerk als Medium kommunikativer Selbstentwürfe und literarischer Interaktion.“ In *Rilkes Korrespondenzen*, hg. von Alexander Honold und Irmgard M. Wirtz, 7–32. Beide Seiten - Autoren und Wissenschaftler im Gespräch 6. Göttingen, Zürich: Wallstein; Chronos.

House, Madeline, Graham Storey, Kathleen Tillotson (Hg.). 2001. *The Letters of Charles Dickens: 1820-1870. Electronic Edition.* Charlottesville, Virginia: InteLex Corporation.

Hotson, Howard und Miranda Lewis (Hg.). Seit 2009. „Early Modern Letters Online.“ http://emlo.bodleian.ox.ac.uk

Liebl, Bernhard und Manuel Burghardt. “‘The Vectorian’ - Eine parametrisierbare Suchmaschine für intertextuelle Referenzen.” In *DHd2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, 232–235.

Nantke, Julia (Hg.) unter Mitarbeit von Marie Flüh und Sandra Bläß. Seit 2021. *Dehmel digital*. https://dehmel-digital.de/

Orchard, Jack, Mark Rogerson, Megan Gooch, Judith Siefring. Seit 2008. *Electronic Enlightenment*. https://www.e-enlightenment.com

Trier Center for Digital Humanities. 2023. Projekte. https://tcdh.uni-trier.de/de/projekte

„Da schunklet älli mit“ – Sammlung und Erforschung dezentraler Kulturdaten der schwäbisch-alemannischen Kneipenfastnacht

Hein, Pascal
pascal.hein@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID: 0000-0002-2138-3069

Die schwäbisch-alemannischen Fastnacht ist schon seit Beginn ihrer modernen Renaissance Gegenstand ethnologischer oder literatur- und theaterwissenschaftlicher Forschung (Mezger, 1999). Der Schwerpunkt der bisherigen Untersuchungen lag dabei meist auf der Analyse der Kostüme (z.B. Hohl, 1999), der Einordnung öffentlich aufgeführter Bräuche, wie Umzügen oder der symbolischen Verbrennung der Fastnacht in einem großen Stroh- oder Reisigfeuer (z.B. Graf, 2019) oder der Erkundung der historischen Ursprünge und (Dis-)Kontinuitäten (z.B. Mezger, 1991).

Bislang weniger in den Blick genommen wurde dabei die Bedeutung der Kneipenfastnacht für die Kulturpraxis im deutschen Südwesten und in der nordwestlichen Schweiz[1]. Diese ist ein ebenso zentraler Bestandteil der Feste und Bräuche rund um die Fastnacht. Aus Einzelhandelsfilialen, Garagen und Vereinsheimen werden in den fastnächlichen hohen Tagen „Besenwirtschaften“, von Vereinen oder Privatpersonen betrieben. Hier wird gemeinsam getanzt und gesungen. Für die eine Höhepunkt, für den anderen Unterbrechung dieser Feiern sind die Auftritte unterschiedlichster Gruppen. Die Kneipenfastnacht wird so in einigen Orten Schmelztiegel der dezentralen Kulturproduktion.

Mein Dissertationsprojekt setzt sich zum Ziel, einen erforschbaren Datensatz der im Rahmen der Kneipenfastnacht entstandenen Lieder, Gedichte und Stücke zu erstellen [1], diesen nach den FAIR- und CARE-Prinzipien einzurichten und mit dem Einverständnis der Autorinnen und Autoren zur Verfügung zu stellen. Außerdem soll ein genauerer Überblick über die behandelten Themen und die Diskurspositionen gewonnen werden: Wie homogen ist die Themenauswahl? Wie stehen lokale, bundesweite und internationale Themen zueinander quantitativ und qualitativ im Verhältnis? Lassen sich die Positionen klar politischen Ausrichtungen zuweisen? Wie stark ist die politische Homogenität? Ich werde hierfür Ansätze im Sinne des Scalable Reading umsetzen und hermeneutische wie auch textstatistische Analysen des Korpus durchführen, beginnend mit KWIC, Keyword Extraction und most frequent n-grams, was dann um Methoden des Argument Mining erweitert werden soll. Auch die Form soll Teil dieser Untersuchung sein, um so auch in diesem Bereich Innovationen und Kontinuitäten erkennen zu können.

Den Fokus dieser Vorstellung möchte ich auf die Villinger Fastnacht und die dortigen Formen legen, da hierzu schon das meiste Datenmaterial vorliegt. Die Gruppen der Villinger Kneipenfastnacht unterscheiden sich einerseits hinsichtlich ihres Charakters als Zug/Abteilung/Gruppe einer Fastnachtsvereinigung oder als freie Gruppe, die in dieser Form nur für die Kneipenfastnacht zusammenkommt und andererseits bezogen auf die Gattung der jeweiligen Aufführung. Ich konzentriere mich hierbei auf diejenigen Aufführungen, die sich in Text niederschlagen können, also vor allem Lieder, Gedichte und Stücke.

Schon eine erste Exploration des bisher gesammelten Materials macht deutlich, dass neben politischen Themen und Alltagssujets vor allem fastnachtsspezifische Phänomene behandelt werden. Die Werke dienen hierbei nicht nur der Belustigung, sondern stelleneinen Diskussionsraum dar: Wer darf wann an welchem Brauch in welcher Rolle teilnehmen? An welchen Stellen ist Innovation erlaubt? Wo muss das bestehende gegen Veränderung beschützt werden?

Allein die Sammlung der Datengrundlage bringt in diesem Kontext große Herausforderungen mit sich, die zum einen im dezentralen Charakter der Produktion und Datenverfügbarkeit begründet sind, zum andern aber auch in brauchspezifischen Eigenheiten. Die Aufführungen werden in der Regel nicht aufgezeichnet, die Texte sind geschrieben, um im Raum zu verhallen und werden nicht in Schriftform verfügbar gemacht. Jede Gruppe verwaltet selbst die eigene Sammlung der verfassten Texte und viele der Autorinnen und Autoren wollen nicht als solche benannt werden, was nur im Kontext der fastnächtlichen Anonymität zu verstehen ist.

Kulturgüter, die so dezentral und so entfernt vom etablierten Literaturbetrieb generiert werden, blieben häufig unterhalb des Radars vieler Forschender. Die Digital Humanities bieten hier Chancen und es bleibt zu diskutieren, welche Rolle Public Science Ansätze bei der Sammlung, Aufbereitung und Kontextualisierung spielen können.

Fußnoten

1. Lediglich für die Basler Praxis der „Schnitzelbank“, bei der mehrere Figuren mit Sprechmasken in gebundener Rede über das vergangene Jahr zitieren hat hier schon größere Betrachtung gefunden (z.B. Canova, 2006). Besonders hervorzuheben ist hier das digitale Archiv des „Schnitzelbank-Comités“, abrufbar unter: https://www.schnitzelbankbasel.ch/
2. Dabei werden die TEI-Tags verwendet, mit denen Roman Schneider deutsche Pop-Lieder ausgezeichnet hat (2022).

Bibliographie

Canova, Ruth. 2006. *Jo, das isch e Schnitzelbangg!* Basel: Spalentor Verlag.

Graf, Edi. 2019. *Fasnet – Schwäbisch-Alemannische Zünfte und Hochburgen*. Meßkirch: Gmeiner Verlag.

Hohl, Jürgen. 1999. „Gesichtsvermummung in der Fastnacht“. In *Zur Geschichte der organisierten Fastnacht*, hg. von der Vereinigung Schwäbisch-alemannischer Narrenzünfte, 135-144. Vöhrenbach: Dold Verlag.

Mezger, Werner. 1991. *Narrenidee und Fastnachtsbrauch – Studien zum Fortleben des Mittelalters in der europäischen Festkultur*. Konstanz: Universitätsverlag.

Mezger, Werner. 1999. „Vom organischen zum organisierten Brauch“. In *Zur Geschichte der organisierten Fastnacht*, hg. von der Vereinigung Schwäbisch-alemannischer Narrenzünfte, 7-42. Vöhrenbach: Dold Verlag.

Schneider, Roman. 2022. „Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung“. In *Sprachreport 1/2022*. 38-50.

Digitaler Zugang zu vormodernen slavischen Handschriften

Renje, Elena
elena.renje@slavistik.uni-freiburg.de
Universität Freiburg, Deutschland

Der in der Paläoslavistik bislang vorherrschende traditionelle philologisch-qualitative Untersuchungsansatz ermöglicht sorgfältige und besonders präzise Analysen von Handschriften bzw. Handschriftenpassagen, erweist sich jedoch aufgrund der Close-Reading-Vorgehensweise als zeitintensiv. Dadurch sind umfangreiche statistische Analysen aufgrund des begrenzt verwendeten Datenmaterials nur in eingeschränktem Maße durchführbar. Durch den Einsatz neuer digitaler Methoden, wie OCR (Optical Character Recognition) und HTR (Handwritten Text Recognition) eröffnen sich neue Möglichkeiten zur Untersuchung umfangreicher Textkorpora und zur Verarbeitung von Big Data (Muehlberger et al., 2019).

Das Dissertationsprojekt fokussiert sich auf vormoderne slavische Handschriften aus dem 16. Jahrhundert. Die Handschriften entstammen den *Velikie Minei Čet'i* (auch bekannt als *Die Großen Lesemenäen des Metropoliten Makarij*), die in drei Fassungen vorliegen. Sie sind unter der Leitung des Metropoliten Makarij entstanden, der sich zum Ziel setzte, sämtliche religiösen Werke, die im 16. Jh. in der Rus' bekannt waren, in einem Sammelwerk zu vereinen. Die Lesemenäen sind sowohl hinsichtlich ihres Umfangs als auch ihres Inhalts vielfältig und enthalten auf insgesamt knapp 30.000 Folia für jeden Tag des Monats verschiedene Texte unterschiedlicher Art und Genese, an deren Entstehung zahlreiche Schreiber beteiligt waren (Kostjuchina, 2000; Orlov, 1945; Vodovozov, 1972). Aufgrund ihres erheblichen Umfangs wurden sie bisher nur punktuell und primär qualitativ erforscht (s. bspw. Ljachovickij und Šibaev, 2015; Raleva, 2022; Weiher, 2000).

Die zentrale Fragestellung der Untersuchung lautet, ob sich die einzelnen Handschriften in linguistischer Hinsicht systematisch voneinander unterscheiden und nach welchen Parametern sie gruppiert werden können. Dabei ist von Interesse, inwieweit die Textart oder Schreiberabschnitte zu einer Gruppierung beitragen und inwieweit sich daraus Rückschlüsse auf Datierung und Lokalisierung ziehen lassen.

Das methodische Vorgehen ist quantitativ motiviert und besteht aus einem mehrstufigen Workflow. Zu Beginn wird die automatische Transkription der Handschriften mithilfe der Plattform Transkribus (https://readcoop.eu/de/transkribus/) erstellt. Die HTR-Transkriptionen werden anschließend mithilfe des Tools UDPipe tokenisiert (https://github.com/bnosac/udpipe) und für morphologisches sowie PoS-Tagging wird die stand-alone Version des Stanzataggers (https://github.com/yvesscherrer/stanzatagger) verwendet. Darüber hinaus werden in RStudio, neben der deskriptiven und inferentiellen Datenanalyse unter Verwendung des R-Pakets *stylo* (Eder et al., 2016) stilometrische Clusteringmethoden angewandt, um Unterschiede in der linguistischen Variation zwischen Subkorpora festzustellen, oder Schreiberwechsel zu erkennen. Auch der Einsatz des Nearest-Shrunken-Centroids Algorithmus, der bereits in linguistischen Studien zur induktiven Merkmalsgenerierung angewandt wurde (Lahjouji-Seppälä et al., 2022), soll weitere Erkenntnisse über mögliche distinktive linguistische Merkmale zwischen den Subkorpora liefern.

Der vorgestellte Workflow ist nicht frei von Herausforderungen. Bereits der HTR-Output birgt das Problem von Transkriptionsfehlern, wenngleich die CER des Modells unter 4% liegt (Rabus, 2019). Basierend auf bisherigen Forschungsergebnissen ist nicht davon auszugehen, dass diese HTR-Fehlerquote die statistische Analyse einzelner linguistischer Parameter sowie stilometrische Analysen signifikant beeinflusst (Camps et al., 2020; Eder, 2013; Franzini et al., 2018; Rabus und Petrov, 2023). Dementsprechend sind zuverlässige statistische Analysen grundsätzlich möglich. Trotzdem ist bisher unklar, inwieweit die bisherigen Erkenntnisse auf weitere linguistische Parameter ausgeweitet werden können. Auch beim Tagging ist von der Durchführbarkeit zuverlässiger Analysen auszugehen (Besters-Dilger und Rabus, 2021). Allerdings ist nicht absehbar, ob sich die Fehlerquoten des HTR-Outputs und des Tagging-Outputs gegenseitig verstärken, oder ob diese Kombination unproblematisch für die Analyse ist. In stilometrischer Hinsicht muss ebenso erprobt werden, ob diese Methode zur Identifizierung von Schreiberwechseln genutzt werden kann, oder ob das orthographische Signal der Handschriften, deren wesentliches Attribut eine nicht-normierte Orthographie ist,

stärker ist und die stilometrischen Ergebnisse dadurch beeinflusst werden (Büttner et al., 2017).

In Anbetracht der möglichen digitalen Schwierigkeiten ist es notwendig, die Möglichkeiten und Grenzen eines solchen methodischen Vorgehens abzuwägen und gegebenenfalls an geeigneten Stellen zu modifizieren. Gemeint ist damit die Abwägung, basierend auf der Evaluierung von HTR- und Tagging-Fehlern, die dementsprechenden Modelle zu verfeinern. Zugleich sollten jedoch auch die künftigen Potenziale, die sich dadurch erschließen lassen, nicht außer Acht gelassen werden. Der Einsatz digitaler Mittel und quantitativer Ansätze eröffnet tiefergehende Möglichkeiten zur Identifizierung bisher unentdeckter linguistischer Muster und verspricht neue Erkenntnisse für die paläoslavistische Forschung.

Bibliographie

Besters-Dilger, Juliane und Achim Rabus. 2021. “Neural Morphological Tagging for Slavic: Strengths and Weaknesses.” *Scripta & E-Scripta 21* : 79–92.

Büttner, Andreas, Friedrich M. Dimpel, Stefan Evert, Fotis Jannidis, Steffen Pielström, Thomas Proisl, Isabella Reger, Christof Schöch und Thorsten Vitt. 2017. „»Delta« in der stilometrischen Autorschaftsattribution.“ PDF-Format ohne Paginierung. *Zeitschrift für digitale Geisteswissenschaften* . DOI: 10.17175/2017_006 .

Camps, Jean-Baptiste, Thibault Clérice und Ariane Pinche. 2019. „Stylometry for Noisy Medieval Data: Evaluating Paul Meyer’s Hagiographic Hypothesis.“ arXiv preprint, *arXiv* : 2012.03845.

Eder, Maciej. 2013. „Mind your corpus. Systematic errors in authorship attribution.“ *Literary and linguistic computing* 28.4: 603–614.

Eder, Maciej, Jan Rybicki und Mike Kestemont. 2016. “Stylometry with R: a package for computational text analysis.” *R Journal* 8.1: 107–121 https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf. (zugegriffen: 19. Juli 2023).

Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk und Jan Rybicki. 2018. “Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm.” *Frontiers in Digital Humanities 5.4*: 1–15.

Kostjuchina, Ljudmila M. 2000. “Russkie piscy v knižnoj masterskoj mitropolita Makarija (po Uspenskomu Spisku Velikich Minej Chetīi).” In *Abhandlungen zu den grossen Lesemenäen des Metropoliten Makarij. Kodikologische, miszellanologische und textologische Untersuchungen* , Bd. 1, hg. von Christian Voss, Heide Warkentin und Eckhard Weiher, 21–36. Freiburg i. Br.: Weiher.

Lahjouji-Seppälä, M. Zaidan, Achim Rabus und Ruprecht von Waldenfels. 2022. “Ukrainian standard variants in the 20th century: stylometry to the rescue.” *Russian Linguistics 46* : 217–232. https://doi.org/10.1007/s11185-022-09262-9.

Ljachovickij, Evgenij A. und Michail A. Šibaev . 2015. “Zametki o chronologii I porjadke raboty nad Sofijskim komplektom Velikich Minej Čet′ich. ” Trudy istoričeskogo fakul′teta Sankt-Peterburgskogo universiteta 24: 8–13.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen . . . Konstantinos Zagoris. 2019. “Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study.” *Journal of Documentation* , 75.5: 954–976.

Orlov, Aleksandr S. 1945. *Drevnjaja russkaja literatura. XI–XVII vekov* . Leningrad: Izdatel'stvo Akademii Nauk SSSR.

Rabus, Achim. 2019. “Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus.” *Scripta & E-Scripta 19* : 9–32.

Rabus, Achim und Ivan N. Petrov. 2023. “Linguistic Analysis of Church Slavonic Documents: A Mixed-Methods Approach.” *Scando-Slavica* , 69.1: 25–38, DOI: 10.1080/00806765.2023.2189617.

Raleva, Cvetana Ch. 2019. “K voprosu o grafiko-orfografičeskoj norme v Uspenskom komplekte Velikich Minej Velikich Minej Chetīich mitropolita Makarija (na material Mučenija sv. Aleksandra Rimskogo).” *Bolgarskaja rusistika* 1: 28–38.

Vodovozov, Nikolaj V. 1972. *Istorija drevnej russkoj literatury* . izd. 3-e. Moskva: Prosveščenie.

Weiher, Eckhard. 2000. „Zu den Abhängigkeitsverhältnissen der drei Fassungen der Grossen Lesemenäen des Metropoliten Makarij.“ In *Abhandlungen zu den grossen Lesemenäen des Metropoliten Makarij. Kodikologische, miszellanologische und textologische Untersuchungen* , Bd. 1, hg. von Christian Voss, Heide Warkentin und Eckhard Weiher, 139–158. Freiburg i. Br.: Weiher.

Figurenbeschreibungen in deutschsprachigen Romanen (1789–1914)

Hilger, Agnes

agnes.hilger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Das Dissertationsprojekt verortet sich im Bereich Computational Literary Studies (CLS) und erforscht die Ent-

wicklung der Figurenbeschreibung in deutschsprachigen Romanen im ‚langen 19. Jahrhundert'.

Den Ausgangspunkt bildet die für englischsprachige Romane des 19. Jahrhunderts gezeigte Zunahme konkreter Wörter (Heuser und Le-Khac, 2012; Underwood 2019; Piper 2022; Reeve 2023). In Vorarbeiten konnte gezeigt werden, dass in deutschsprachigen Romanen ein ähnlicher Trend existiert: Die relativen Häufigkeiten von Wörtern, die physisch Wahrnehmbares bezeichnen, etwa Kleidung, Möbel oder Köperteile, steigen im untersuchten Korpus über das 19. Jahrhundert hinweg an (Hilger, 2023).

Diese Entwicklung lässt sich mit herkömmlichen literaturwissenschaftlichen Darstellungen nicht erklären.[1]Einmal mehr zeigt sich das Potential von Distant Reading-Verfahren (Moretti, 2013; Jockers, 2013). Allerdings können wir aus einer ‚distanten' Perspektive noch nicht sagen, was auf der Ebene der Textmerkmale passiert.

Das Dissertationsprojekt setzt hier an und untersucht in einer Kombination aus Distant und Close Reading wie sich der Teilbereich der Beschreibung von Figuren verändert. Dafür werden textuelle Phänomene annotiert: Figurenbeschreibungen, deskriptive Elemente, direkte Charakte-risierung und charakterisierende Elemente.

Figurenbeschreibungen und deskriptive Elemente werden als Texteinheiten verstanden, in denen der physisch wahrnehmbaren Außenseite der Figur vergleichsweise stabile Eigenschaften zugeschrieben werden.[2] Figurenbeschreibungen sind längere Sequenzen, die den Fokus auf die Beschreibung legen, während deskriptive Elemente als Satzteile in nicht primär beschreibende Sätze eingelagert sind. Komplementär dazu beziehen sich direkte Charakterisierung und charakterisierende Elemente auf die physisch nicht wahrnehmbare Innenseite der Figur. Tabelle 1 gibt einen Überblick über die Informationen, die nach der aktuellen Version der Guidelines erfasst werden.[3]

Textmerkmal	Figurenbeschreibung und deskriptive Elemente	Direkte Charakterisierung und charakterisierende Elemente	Ersterwähnung
Eigenschaften	- Wer wird beschrieben? - Wer beschreibt? - Was wird beschrieben? - Art der Beschreibung - Polarität - Explizite Verknüpfung mit Eigenschaft des Inneren - Markierung als ‚typisch' für Gruppe	- Wer wird charakterisiert? - Wer charakterisiert - Art der Charakterisierung - Polarität - Markierung als ‚typisch' für Gruppe	- Wer wird erwähnt?

Tabelle 1: Übersicht über die annotierten Kategorien und Properties.

Die Annotation erfolgt zunächst manuell in CATMA (Gius, 2022), wird jedoch später automati-siert und auf alle Texte im Korpus ausgeweitet.

Das Korpus besteht im Moment aus 925 zwischen 1789 und 1914 erschienenen Romanen. Es basiert größtenteils auf den bei TextGrid und im Projekt Gutenberg offen verfügbaren Texten, soll aber noch erweitert und ausbalanciert werden. Neben anderen Metadaten wird die jeweilige ‚Kanonizität' eines Texts erfasst, einerseits um die Zusammensetzung des Korpus transparent zu machen, andererseits, um später bei den Ergebnissen differenzieren zu können. Grundlegend ist dabei Winkos Beschreibung von Kanonisierung als Phänomen der unsichtbaren Hand (2002): Zahlreiche Handlungen auf einer Mikroebene führen gemeinsam auf einer Makroebene zur Kanonisierung eines Autors/einer Autorin, ohne dass dies im Einzelnen beabsichtigt sein muss. Die Rekonstruktion folgt der Logik, dass diese Handlungen zugleich Indikator für die Kanonizität zu einer bestimmten Zeit und in Bezug auf eine bestimmte Gruppe sein können. ‚Kanonizität' wird dementsprechend gemessen über: Nennungen in Literaturgeschichten (Jannidis, 2013; Brottrager u.a., 2021), in der BDSL, in universitären Kurskatalogen und auf Leselisten.

Im Analyse-Teil wird untersucht, wie sich die Figurenbeschreibung über das lange 19. Jahrhundert hin im Korpus entwickelt: Nimmt ihr Auftreten zu, und, wenn ja, in welcher Form? Welche Unterscheide gibt es hinsichtlich der Kanonizität? In welchem Verhältnis stehen direkte Charakterisierung und Figurenbeschreibung? Wie werden männliche, wie weibliche Figuren beschrieben? Wie entwickeln stereotype Zuschreibungen, etwa hinsichtlich Ethnizität über die Zeit hin und welche Eigenschaften werden Gruppen als ‚typisch' markiert? Im Anschluss an die Forschung zu Literatur und Physiognomik interessieren Fragen nach der Verknüpfung von äußeren und inneren Eigenschaften.

Erkenntnisgewinn verspricht sich die Arbeit vor allem von der Kombination quantitativer und qualitativer Verfahren in einem Mixed Methods-Design. Ein solches wird in den CLS seit mehreren Jahren unter verschiedenen Begriffen eingefordert.[4] Indem die Arbeit ein solches Forschungsdesign entwickelt, will sie auch einen Beitrag zur Methodologie leisten und *beiden* Literaturwissenschaften Anschluss bieten, um über die disziplinären Grenzen hinweg zu einem fundierteren Verständnis des gemeinsamen Untersuchungsgegenstandes zu gelangen.

Fußnoten

1. In der Regel wird hinsichtlich der ‚Realitätseffekte' (Barthes) und der physischen Konkretheit vor allem die Sonderrolle des Realismus betont, so etwa von Vedder und Scholz (2018, S. 9).
2. Ich gehe hier von dem bei Jannidis beschriebenen Basistyp der Informationsstruktur der Figur aus (2004).
3. Die Guidelines werden wie in den CLS üblich in einem zyklischen Prozess erarbeitet (vgl. Reiter, 2020).
4. Bereits Jockers fordert 2013 anstelle der alleinigen Konzentration auf die für sein Buch titelgebende Makroanalysis eine Kombination mit der Detailanalyse (2013, S. 26); Thomas Weitin greift den von Martin Mueller geprägten Begriff ‚Scalable Reading' auf (Mueller, o.D.; Weitin, 2017). Einen guten Überblick über die Diskussion um ‚Mixed Methods' in den CLS gibt Rabea Kleymann (2022).

Bibliographie

Brottrager, Judith, Annina Stahl, und Arda Arslan. 2021. „Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features“. *Computational Humanities Research Conference*.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, und Jan Horstmann. 2022. „CATMA“. https://doi.org/10.5281/ZENODO.1470118 (zugegriffen am 19.01.2024).

Jannidis, Fotis. 2004. „Figur und Person. Beitrag zu einer historischen Narratologie“. *Narratologia 3*. Berlin/New York: Walter de Gruyter.

Jannidis, Fotis. 2013. „Kanonbildung in Literaturgeschichten“. In *Handbuch „Kanon und Wertung“. Theorien, Instanzem, Geschichte*, herausgegeben von Gabriele Rippl und Simone Winko, 159–67. Stuttgart, Weimar: Metzler.

Jockers, Matthew L. 2013. “Macroanalysis: Digital Methods and Literary History”. Urbana.

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 4:* 1–66.

Hilger, Agnes. 2023. „Dunkelgrün, blassgrün, fenchelgrün oder: Über die Konkretisierung des Vokabulars im deutschsprachigen Roman (1760–1920)“, *DHd 2023 Conference Abstracts.* https://doi.org/10.5281/zenodo.7711478 (zugegriffen am 19.01.2024).

Kleymann, Rabea. 2022. „Datendiffraktion: Von Mixed zu Entangled Methods in den Digital Humanities“. Herausgegeben von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Trilcke, Niels Walkowski, Joëlle Weis, und Ulrike Wuttke. *Zeitschrift für digitale Geisteswissenschaften. Sonderbände 5.* https://doi.org/10.17175/SB005_008 (zugegriffen am 19.01.2024).

Moretti, Franco. 2013. “Distant reading”. London/New York: Verso.

Mueller, Martin. o.D. “Scalable Reading”. https://scalablereading.northwestern.edu (zugegriffen am 19.01.2024).

Neuroth, Heike, Andrea Rapp, und Sibylle Söring, Hrsg. 2015 „TextGrid. Von der Community - für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften“. Glückstadt.

o.A. „Projekt Gutenberg-DE. Edition 14“. https://www.projekt-gutenberg.org/index.html (zuge-griffen am 19.01.2024).

Piper, Andrew, und Sunyam Bagga. 2022. „A Quantitative Study of Fictional Things“. *Proceedings of the Conference on Computational Humanities Research*, 2022.

Reiter, Nils. 2020. „Anleitung zur Erstellung von Annotationsrichtlinien“. In *Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*, 193–202. De Gruyter. https://doi.org/10.1515/9783110693973-009 (zugegriffen am 19.01.2024).

Reeve, Jonathan. 2023. „The Eye of Modernism“. https://dissertation.jonreeve.com/.

Scholz, Susanne, und Ulrike Vedder. 2018. „Einleitung“. In *Handbuch Literatur und materielle Kultur*, herausgegeben von Susanne Scholz und Ulrike Vedder. Berlin/Boston.

Underwood, Ted. 2019. “Distant horizons. Digital Evidence and Literary Change”. Chicago: The University of Chicago Press.

Weitin, Thomas. 2017. “Scalable reading”. *Zeitschrift für Literaturwissenschaft und Linguistik.* Stuttgart: J.B. Metzler.

Winko, Simone. 2002. „Literatur-Kanon als invisible hand-Phänomen". In *Literarische Kanon-bildung*, herausgegeben von Heinz Ludwig Arnold und Hermann Korte, 9–24. München.

Frauen im frühromantischen Briefnetzwerk Quantitative Einblicke in weibliche Lebenswelten des Bildungsbürgertums um 1800

Suárez Cronauer, Elena
Elena.SuarezCronauer@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Einleitung

Die von Koselleck als „Sattelzeit“ (vgl. Koselleck 1972: XIV) beschriebene Zeitspanne um die Jahrhundertschwelle von 1800 ging mit tiefgehenden Transformationen einher, die das moderne Europa in all seinen Facetten nachhaltig prägten. Im Zuge der französischen Revolution gestalteten sich somit neue Sichtweisen auf die soziale, politische, kulturelle und juristische Ausrichtung der Gesellschaft.

Dies beinhaltete auch Debatten zur Frage der Beziehung der Geschlechter sowie zur Rolle der Frau. Frauen wurden damals gesamtgesellschaftlich nicht als gleichermaßen vernunftbegabte Wesen anerkannt, womit getrennte Wirkungsbereiche von Mann und Frau begründet wurden (vgl. Lange 1991: 411). Gleichwohl gab es auch Gegenstimmen zu dieser rechtlichen, sozialen und bildungspolitischen Un-

gleichbehandlung, wie die Schriften von Theodor Gottlieb von Hippel oder Mary Wollstonecraft belegen (vgl. Weckel 2000, 209-212). Dies verdeutlicht die Diskussionen um den Umgang mit weiblichen Lebenswelten[1] sowie der Rolle von Frauen in einem sich etablierenden Bildungsbürgertum. Als ein Teil dieses Bildungsbürgertums können die Frühromantiker*innen identifiziert werden. Sie verstanden sich als Gruppe, die nicht nur über Literatur und Poesie neu nachdachte, sondern auch die Geschlechterbilder der Zeit kritisch reflektierte und sich an alternativen Lebensformen ausprobierte.

Fragestellung

Innerhalb meiner Dissertation möchte ich einen Beitrag zum Verständnis weiblicher Lebenswelten zur Zeit der Frühromantik unter Einbezug einer historischen Netzwerkanalyse auf Grundlage des frühromantischen Briefwechsels leisten und die Auswertungen dieser Analyse anschließend auf die sich wandelnde Rolle der Frau im Bildungsbürgertum um 1800 rückführen und hinterfragen. Die quantitative Betrachtung von Informationen über die Frauen der Frühromantik und deren Wirkungskreis stellt ein Desiderat in der Forschung dar. Weibliche Lebenswelten sollen als Plural begriffen und somit weder Frauen als eine homogene Gruppe betrachtet noch einzelne Frauen als Ausnahmeerscheinungen fokussiert werden. Vielmehr sollen die heterogenen Ausrichtungen, Möglichkeiten und Grenzen weiblicher Rollen und Funktionen in der bildungsbürgerlichen Gesellschaftsschicht um 1800 untersucht werden. Briefe als Ego-Dokumente und als Schnittstelle zwischen privaten und öffentlichen Raum stellen hierbei einen vielversprechenden, da unmittelbareren Zugang zu diesen Lebenswelten dar.

Quellen

Quellengrundlage sind die Briefe, die innerhalb des DFG-Projekts „Korrespondenzen der Frühromantik. Edition – Annotation – Netzwerkforschung" für den Zeitraum zwischen 1790 und 1802 in einer digitalen Edition erfasst werden.[2] Dieses insgesamt ca. 6.500 Briefe fassende Korpus beinhaltet auch Schriftstücke von Frauen, die nicht direkt der Gruppe der Frühromantiker*innen zugeordnet werden, aber Korrespondenzpartnerinnen dieser Personen waren. Somit ist es möglich, über den Kreis der Frühromantik hinaus auf weitere Akteurinnen innerhalb des (Bildungs-)Bürgertums in die Analysen einzubeziehen. Über das projektinterne Korpus hinaus soll zudem geprüft werden, ob und welche Daten aus anderen Briefeditionen in die Untersuchung einbezogen werden können, z.B. über correspSearch[3] .

Methode

Die Briefe, die innerhalb der Forschungsprojekts „Korrespondenzen der Frühromantik“ erhoben werden, sind in einem Knowledge Graphen mit Daten zu editorischen Informationen, Meta- und Registerdaten sowie Aussagen in den Briefen bezüglich des Wissenstransfers und der Kommunikationsprozesse modelliert. Der Knowledge Graph ist Grundlage für die Netzwerkanalysen, also für Untersuchungen von Akteur*innen und Verbindungen, die diese Akteur*innen verknüpfen und soziale Strukturen formen.

Eine wesentliche Analysekategorie für die Netzwerke stellt das Geschlecht dar. Somit können Untersuchungen auf vier Korpora angewandt werden: Alle Briefe, Frauenbriefe, Männerbriefe sowie Briefe zwischen Frauen und Männern. Netzwerkmethodologisch untersucht werden innerhalb dieser Korpora zunächst Strukturen wie Korrespondenz- oder Kookkurrenznetzwerke erwähnter Personen, Werke, Periodika und Institutionen. Diese ersten Auswertungen bieten Ansatzpunkte für weitere Analysen, z.B. um Phasen intensiven Austausches zwischen Akteur*innen oder Diskussionen über literarische Werke innerhalb von Briefen zu untersuchen. Im Zuge weiterer Analysen wird hierbei die Aktivität in Briefen im Kontext der historischen Gegebenheiten abgeglichen, auch hinsichtlich der Forschung zum Bildungsbürgertum. Zudem bieten sich durch die Annotation von Aussagen innerhalb der Briefe Möglichkeiten, spezifische Themenkomplexe zu identifizieren und deren Entwicklung innerhalb der Netzwerke nachzuvollziehen, bspw. die Mitarbeit von Frauen an bestimmten Werken.

Perspektiven und Anschlüsse

Die Dissertation gliedert sich in das Forschungsfeld der Digital Humanities ein, berücksichtig dabei aber ebenso literaturwissenschaftliche und geschlechtergeschichtliche Untersuchungen sowie interdisziplinäre Einflüsse aus den Genderstudies. Im Kontext dessen muss auch die lückenhafte Quellensituation bezüglich Frauenbriefen angesprochen werden, die lange Zeit nur in Relation mit männlichen Korrespondenzpartnern gedacht wurden sowie von einer männlichen Sammlungspraktik geprägt sind. Daher soll auch geprüft werden, wie und welche Ansätze aus dem data feminism (D'Ignazio/Klein 2020) für die Analyse historischer Daten mit netzwerkanalytischen Methoden adoptiert bzw. übertragen werden können, sodass die Methoden auch auf andere andere Quellenbestände und Forschungsvorhaben, die sich mit Emanzipationsmöglichkeiten marginalisierter Gruppen beschäftigen, angewandt werden können.

Fußnoten

1. ‚Lebenswelten' werden definiert in Anlehnung an Schulz, der diese als „sinnstiftende, konstitutive Leistungen des Subjekts, das die 'objektiv' gegebene Welt deutend in den Alltag einordnet" beschreibt. Dies meint auch kulturelle Praktiken, durch die „selbstverständliches, alltägliches Wissen entsteht, das die Welt ordnet und die Lebensgewohnheiten des Individuums strukturiert" (Vgl. Schulz 2014: 53).
2. Für mehr Informationen zum Projekt vgl. https://www.adwmainz.de/projekte/korrespondenzen-der-fruehromantik-edition-annotation-netzwerkforschung/informationen.html.
3. *correspSearch* stellt Verzeichnisse digitaler und gedruckter Briefeditionen zur Verfügung, vgl. https://correspsearch.net.

Bibliographie

D'Ignazio, Catherine, und Lauren F. Klein. Data Feminism . Cambridge, Massachusetts: The MIT Press, 2020.

Koselleck, Reinhart. „Einleitung". In: Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Bd. 1, S. XIII–XXVII, Stuttgart 1972.

Kremer, Detlef. Romantik: lehrbuch germanistik . Stuttgart: J B Metzler'Sche Verlag, 2015.

Lange, Sigrid, Hrsg. Ob die Weiber Menschen sind. Geschlechterdebatten um 1800. Leipzig: Reclam-Verlag, 1992.

Schulz, Andreas: Lebenswelten und Kultur des Bürgertums im 19. und 20. Jahrhundert. Berlin u.a.: De Gruyter, 2014.

Weckel, Ulrike. „Gleichheit auf dem Prüfstand. Zur zeitgenössischen Rezeption der Streitschriften von Theodor Gottlieb von Hippel und Mary Wollstonecraft in Deutschland". In Tugend, Vernunft und Gefühl. Geschlechterdiskurse der Aufklärung und weibliche Lebenswelten, 209–49, Münster, 2000.

„Mutter, Vater, Kind". Ressourcenarme automatische Metaphernverarbeitung für religionswissenschaftliche Fragestellungen

Rodenhausen, Lina
lina.rodenhausen@rub.de
Ruhr-Universität Bochum, Deutschland
ORCID: 0000-0002-4709-082X

Das zeitgenössische Christentum scheint geprägt von einer Polarisierung zwischen progressiven und konservativen Gruppierungen zu sein, was bisher jedoch kaum systematisch erforscht wurde.[1] Social-Media-Plattformen bilden ein wichtiges Feld, in dem Einblicke in zeitgenössische Religiosität gewonnen werden können. Als Fallstudie wurden die Subreddits r/OpenChristian[2] und r/TrueChristian[3] ausgewählt, bei denen es sich jeweils um eine progressive und eine konservative christliche (englischsprachige) Online-Community handelt. Die folgende Tabelle gibt einen Überblick über den Umfang des untersuchten Korpus.

Subreddit	Threads	Posts	Token	Zeitraum
r/OpenChristian	15.888	158.172	11.720.419	2010–2022
r/TrueChristian	55.986	1.084.214	79.981.720	2012–2022

Um Textmaterial in diesem Umfang ausschöpfend erforschen zu können, ist die Anwendung computergestützter Methoden notwendig. Diese kamen in einem ersten Schritt zum Einsatz, um einen Gesamtüberblick über die vorherrschenden Themen und Einstellungen der Communities zu erhalten. Im Besonderen LDA Topic Modeling mit MALLET (Blei, Ng, und Jordan 2003; McCallum 2002) wurde erfolgreich angewendet. Weitere computergestützte Diskursanalysen (z.B. Stine, Deitrick, und Agarwal 2020) werden ergänzt, auch für Vergleiche mit anderen Communities.

Neben diesen Überblicksanalysen werden einige Aspekte aus der Kommunikation der User:innen gezielt herausgegriffen und genauer analysiert, um sich ihren Überzeugungen und Glaubensvorstellungen nähern zu können. Hierfür stellt Metaphernanalyse ein wichtiges Mittel dar. Im SFB 1475 „Metaphern der Religion", in dem dieses Dissertationsprojekt eingebettet ist, arbeiten wir unter der Annahme, dass Metaphorizität ein zentrales Prinzip religiöser Sinnbildung ist (Krech, Karis, und Elwert 2023). Das Metaphernverständnis orientiert sich dabei an der Conceptual Metapher Theory von Lakoff und Johnson (2003), wonach eine

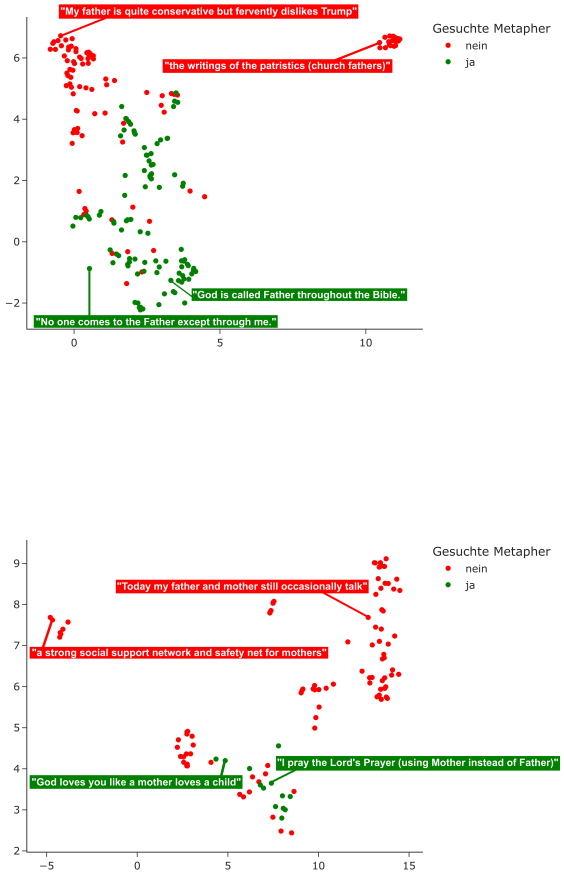
Metapher als ein Mapping von einer Quelldomäne auf eine Zieldomäne zu verstehen ist (Lakoff 1986, 294).

Die Menge an Text führt dazu, dass auch für den Schritt der Metaphernanalyse computergestützte Methoden zum Einsatz kommen müssen. Die automatische Identifikation von Metaphern ist ein wichtiges Thema der Computerlinguistik und wird dort beforscht (Rai und Chakraverty 2021). Auch im Rahmen dieses Projekts findet solche Forschung statt und ergänzt fruchtbar religionswissenschaftliche Perspektiven. Gleichzeitig wird in der Dissertation der Ansatz verfolgt, digitale Methoden in eine geisteswissenschaftliche Forschung zu integrieren. In den Digital Humanities sind viele Ansätze der automatischen Textanalyse aufgrund fehlender annotierter Trainingsdaten, Mathematik- und Programmierkenntnissen, Zeit und weiteren Ressourcen ohne die Unterstützung durch Computerlinguist:innen oft schwer umsetzbar (Suissa, Elmalech, und Zhitomirsky-Geffet 2022). Zudem bestehen in der Regel nicht dieselben Forschungsinteressen. Eine Identifikation aller Metaphern im Text, wie es üblicherweise in der automatischen Metaphernidentifikation angestrebt wird, würde dieser Forschung wenig weiterhelfen, da nicht jede Metapher zur religiösen Sinnbildung beiträgt.

Für diese Untersuchung wurden konkrete Metaphern, in denen die Gott-Menschen-Beziehung als Eltern-Kind-Beziehung konzeptualisiert wird, ausgewählt. Ob Christ:innen Gott als „Vater“ bezeichnen, die Alternative „Mutter“ benutzen, wie sie dies begründen und was sie damit verbinden, sowie die Frage, wer mit den „Kindern“ Gottes gemeint ist, gibt Einblick in ihre Überzeugungen und Glaubensvorstellungen. Metaphern mit den beschriebenen Mappings von Quell- und Zieldomäne müssen so vollständig wie möglich gefunden werden, um quantitative und qualitative Analysen zu ihnen durchführen zu können. Zum Beispiel werden Kookkurrenzen mit dem Quellbegriff betrachtet, um die Metaphern in ihrem Gebrauch verstehen und vergleichen zu können.

Die Aufgabe, zu identifizieren, ob ein Begriff für eine bestimmte Bedeutung – in diesem Fall die Zieldomäne – verwendet wurde, lässt sich auch als ein Disambiguierungsproblem verstehen. In diesem Bereich wurden bereits erfolgreich BERT Embeddings (Devlin et al. 2019), eingesetzt (z.B. Wiedemann et al. 2019). Dabei erhalten die Worte im Korpus eine Vektorrepräsentation abhängig von ihrem Kontext. BERT Embeddings kodieren eine Form von Bedeutung, was sich daran zeigt, dass Vektoren von Instanzen der Quellbegriffe, die mit der Bedeutung der gesuchten Zieldomäne verwendet werden, tendenziell eine geringere Distanz zueinander aufweisen als Instanzen desselben Begriffs, wenn er mit einer anderen Bedeutung verwendet wird. Eine zweidimensionale Visualisierung veranschaulicht das räumliche Clustering von verschiedenen Bedeutung (siehe Abbildungen). Es ist daher möglich, eine geringe Menge der Instanzen des Quellbegriffs manuell zu annotieren und mit Hilfe eines K-Nearest-Neighbor-Algorithmus die restlichen Fälle zu klassifizieren.

Dieses Vorgehen ist durch die zweidimensionale Visualisierung intuitiv und leicht nachvollziehbar und konnte mit begrenzten Programmierkenntissen und Rechenleistung auch ohne Kooperationspartner:innen, wie etwa Computerlinguist:innen, durchgeführt werden. Diese Dissertation erweitert das religionswissenschaftliche Methodenrepertoire um Ansätze aus den Digital Humanities, welche explizit erläutert und reflektiert werden, da ihre Anwendung in der Religionswissenschaft, gerade auch der Forschung zu digitaler Religion, keineswegs starke Verbreitung und Bekanntheit genießt. Die Studie zeigt, wie das Feld von ihnen profitieren kann. Die Dissertation soll dazu beitragen, digitale Methoden zu einem selbstverständlicheren Teil der Forschung zu digitaler Religion sowie zu Metaphern aus geisteswissenschaftlicher Perspektive zu machen.



Abbildungen: 2D-Visualisierungen von Embeddings von annotierten Instanzen der Quellbegriffe „father“ (oben) und „mother“ (unten) im Korpus r/OpenChristian

Fußnoten

1. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1475 – Projektnummer 441126958
2. https://www.reddit.com/r/OpenChristian/
3. https://www.reddit.com/r/TrueChristian/

Bibliographie

Blei, David M., Andrew Y. Ng, und Michael I. Jordan. 2003. „Latent Dirichlet Allocation“ *Journal of Machine Learning Research 3* : 993–1022.

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* : 993–1022. Volume 1 (Long and Short Papers):4171–86. Minneapolis, Minnesota: Association for Computational Linguistics.

Krech, Volkhard, Tim Karis, und Frederik Elwert.2023. "Metaphors of Religion: A Conceptual Framework". Bd. 1.Metaphor Papers. Center for Religious Studies (CERES), Ruhr-Universität Bochum. https://doi.org/10.46586/mp.282.

Lakoff, George. 1986. „The meaning of literal“. *Metaphor and Symbolic Activity*.

Lakoff, George und Mark Johnson. 2003. "Metaphors We Live By". Chicago; London: The University of Chicago Press.

McCallum, Andrew Kachites. 2002. „MALLET: A Machine Learning for Language Toolkit.“ http://mallet.cs.umass.edu.

Rai, Sunny, und Shampa Chakraverty. 2021. „A Survey on Computational Metaphor Processing“. *ACM Computing Surveys 53 (2)* : 1–37. https://doi.org/10.1145/3373265.

Stine, Zachary K, James E Deitrick, und Nitin Agarwal. 2020. „Comparative Religion, Topic Models, and Conceptualization: Towards the Characterization of Structural Relationships between Online Religious Discourses“. In *Proceedings of the Workshop on Computational Humanities Research*, herausgegeben von Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, und Melvin Wevers. 128–48. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2723/long47.pdf.

Suissa, Omri, Avshalom Elmalech, und Maayan Zhitomirsky-Geffet. 2022. „Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science“. *Journal of the Association for Information Science and Technology 73 (2)*: 268–87. https://doi.org/10.1002/asi.24544.

Wiedemann, Gregor, Steffen Remus, Avi Chawla, und Chris Biemann. 2019. „Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings“. arXiv. http://arxiv.org/abs/1909.10430.

Paradigmen einer digitalen Rezeptionswissenschaft Produktiv-literarische Rezeptionsphänomene als Linked Data am Beispiel der deutschsprachigen literarischen Sappho-Rezeption

Untner, Laura
laura.untner@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich
ORCID: 0000-0002-9649-0870

Thema

Die Modellierung von Wissen über literarische Texte und deren Beziehungen zueinander ist eine zentrale Aufgabe der digitalen Literaturwissenschaft. Im Zuge des Dissertationsprojekts werden grundlegende methodologische Überlegungen zu einer digitalen Rezeptionswissenschaft angestellt. Der Fokus liegt auf der Modellierung literaturwissenschaftlichen Wissens über produktiv-literarische Rezeptionszeugnisse als Linked Data, wobei die Sappho-Rezeption im deutschsprachigen Raum als Beispiel dient.

Unter produktiv-literarischen Rezeptionszeugnissen werden literarische Texte von ‚ausführenden‘ Rezipient_innen (Eco, 1973, 29) verstanden, die Elemente eines Rezeptionsgegenstandes enthalten, etwa in der Form von Figuren- und Motivübernahmen, Paraphrasen und Fiktionalisierungen von Autor_innen. Ein Beispiel ist Franz Grillparzers vielfach parodiertes Trauerspiel *Sappho* (UA 1818), worin die titelgebende Dichterin als Figur vorkommt und aus ihrem Werk zitiert wird.

Forschungsstand

Digitale Rezeptionswissenschaft

Bislang sind methodologische Überlegungen für eine digitale Rezeptionswissenschaft nur verstreut zu finden (z. B.

bei Hohl-Trillini und Quassdorf, 2010; Hohl-Trillini, 2020; Barth und Murr, 2017; Women Writers-Project; NEWW Women Writers). Wichtige Anknüpfungspunkte finden sich insbesondere in der digitalen Intertextualitätsforschung, insofern produktiv-literarische Rezeptionsphänomene in der Regel Intertextualitätsphänomene sind (Bluhm und Hölter, 2010, [vii]; Tennis, 2004) und hier wie dort transbiblionome Daten (Wagner et al., 2016) im Vordergrund stehen. Zu nennen sind etwa Editionen, die ihren Fokus auf Intertextualität legen (z. B. Kraus, 2021), und Untersuchungen zur Auffindung (z. B. Coffee und Scheirer, 2018; Manjavacas, 2021) und Modellierung (z. B. Nantke und Schlupkothen, 2018, 2019; Nantke und Sulzbacher, 2020; Oberreither, 2018, 2023a, 2023b; Horstmann et al., 2023) von Intertextualität. Besonders ergiebig sind solche Arbeiten, die sich nicht auf einen lexikalischen Intertextualitätsbegriff beschränken, sondern auch semantische Aspekte miteinbeziehen – was derzeit nur vereinzelt der Fall ist.

Sappho-Rezeption

Als Beispiel dient die produktiv-literarische Sappho-Rezeption im deutschsprachigen Raum vom 15. bis zum 21. Jahrhundert, die ca. 1.000 literarische Texte umfasst. Insbesondere seit dem 20. Jahrhundert ist die Sappho-Rezeption Gegenstand der wissenschaftlichen Forschung, vermehrt noch in den letzten Jahrzehnten im angloamerikanischen und frankophonen Raum (z. B. bei Robinson, 1924; DeJean, 1989; Greene, 1996; Prins, 1999; Andreadis, 2001; Reynolds, 2001). Im deutschsprachigen Raum beschäftigte sich zuerst Horst Rüdiger (1933, 1934) ausführlich mit der literarischen Sappho-Rezeption. Jüngere Arbeiten stammen zum Beispiel von Helmut Saake (1972) und Cornelia Heinsch (2020). Die wichtigsten Erkenntnisse zur deutschsprachigen literarischen Sappho-Rezeption werden aktuell in dem Band *Sappho. Texte zur literarischen Rezeption im deutschsprachigen Raum* (Untner, 2023) versammelt.

Methode

Da es sich bei produktiv-literarischen Rezeptionsphänomenen um netzartige Verhältnisse von ‚Palimpsesten' (Genette, 1993) handelt, eignet sich Linked Data für die Modellierung von Wissen über diese besonders gut. Für Linked Data spricht zudem die Interoperabilität bzw. Kompatibilität von RDF-Datensätzen und dass mit diesen – im Gegensatz zu vielen anderen Formen der Metadatenerfassung – literaturwissenschaftliche Schlussfolgerungen sehr gut kleinteilig nachvollziehbar gemacht und nur implizit vorhandenes Wissen einfach aufgedeckt werden kann.

Detailliert werden produktiv-literarische Rezeptionsphänomene aus rund 100 Texten modelliert, die in *Sappho. Texte zur literarischen Rezeption im deutschsprachigen Raum* zu finden sind. Um das literarische Netzwerk um Sappho in feinkörniger Weise sicht- und analysierbar zu machen, werden etwa Entitäten wie Autor_innen, Werke und Figuren und Relationen wie Zitate und Motivübernahmen modelliert. Damit wird die vorgeschlagene Linked Data-Methode exemplarisch geprüft. Zu den weiteren ca. 900 Rezeptionszeugnissen, die derzeit als CSV erfasst sind, werden nur vereinzelte Metadaten in RDF überführt. Damit wird jedoch zugleich erstmals ein möglichst vollständiger, maschinenlesbarer Katalog mit deutschsprachigen produktiv-literarischen Rezeptionszeugnissen zu Sappho erstellt.

Bei der Modellierung wird primär auf CIDOC CRM, FRBRoo und die OWL-Ontologie INTRO (Intertextual Relationships Ontology for literary studies) zurückgegriffen. Außerdem kommt eine SKOS-Taxonomie zum Einsatz, mit der wiederkehrende Figurennamen, Motive, Themen und Stoffe in der Sappho-Rezeption definiert werden.

Nur am Rande werden zudem Überlegungen zu Methoden bzw. Workflows angestellt, die einer Linked Data-Modellierung vorangehen oder folgen könnten. Dazu zählen etwa Suchabfragen mit regulären Ausdrücken, Annotationen mit XML/TEI, Topic Modeling, Sentimentanalysen, Stilometrie und Netzwerkanalysen bzw. deren möglicher Nutzen für die literaturwissenschaftliche Rezeptionsforschung. Damit soll eine möglichst umfassende sowie anschlussfähige Handreichung garantiert werden.

Ergebnis

Abgeschlossen wird mit einer theoretisch-reflektierenden Doktorarbeit sowie einer Sammlung FAIRer digitaler Daten, die als Open Access zugänglich gemacht wird. Die digitalen Daten werden die verwendete Ontologie, die exemplarische Modellierung der deutschsprachigen literarischen Sappho-Rezeption und die Taxonomie zur literarischen Sappho-Rezeption enthalten. Angedacht wird ein User Interface basierend auf ResearchSpace. Eine Webseite zur Publikation vorläufiger Ergebnisse steht kurz vor der Veröffentlichung.

Bibliographie

Andreadis, Harriette. 2001. *Sappho in Early Modern England. Female Same-Sex Literary Erotics 1550–1714*. Chicago; London: University of Chicago Press.

Barth, Florian und Sandra Murr. 2017. „Digital Analysis of the Literary Reception of J.W. von Goethe's Die Leiden des jungen Werthers." In *Digital Humanities 2017. Conference Abstracts*, hg. von Rhian Lewis, Cecily Raynor, Dominic Foret, Michael Sinatra und Stéfan Sinclair, 540–542. Montreal: McGill University.

Bluhm, Lothar und Achim Hölter. 2010. „Zueignung." In *Produktive Rezeption. Beiträge zur Literatur und Kunst im 19., 20. und 21. Jahrhundert*, hg. von dens. Trier: WVT.

DeJean, Joan. 1989. *Fictions of Sappho 1546–1937*. Chicago; London: University of Chicago Press.

Eco, Umberto. 1973. *Das offene Kunstwerk*, übers. von Günter Memmert. Frankfurt a. M.: Suhrkamp.

Forstall, Christopher W. und Walter J. Scheirer. 2019. *Quantitative Intertextuality. Analyzing the Markers of Information Reuse*. Cham: Springer.

Genette, Gérard. 1993. *Palimpseste. Die Literatur auf zweiter Stufe*, übers. von Wolfram Bayer und Dieter Hornig. Frankfurt a. M.: Suhrkamp.

Greene, Ellen, Hg. 1996. *Re-Reading Sappho. Reception and Transmission*. Berkeley: University of California Press.

Heinsch, Cornelia. 2020. *„sappho gibt es nicht“. Die Rezeption Sapphos in deutschsprachiger Lyrik des 20. und 21. Jahrhunderts*. Baden-Baden: Ergon.

Hohl-Trillini, Regula und Sixta Quassdorf. 2010. „A ‚key to all quotations‘? A corpus-based parameter model of intertextuality.“ In *Literary and Linguistic Computing* 25/3: 269–286.

Hohl-Trillini, Regula. 2020. „WordWeb/IDEM: Datenbasierte Erfassung von Intertextualität durch eine Graphdatenbank zum frühneuzeitlichen englischen Theater.“ In *DHd 2020: Spielräume. Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2020), Paderborn*, hg. von Christof Schöch, 67–68, https://doi.org/10.5281/zenodo.3666690 (zugegriffen: 28. Dezember 2023).

Horstmann, Jan, Christian Lück und Immanuel Normann. 2023. „Textliche Relationen maschinenlesbar formalisieren: Systeme der Intertextualität.“ In *DHd 2023: Open Humanities, Open Culture. 9. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2023), Trier, Luxemburg*, hg. von Peer Trilcke, Anna Busch und Patrick Helling, unter Mitarbeit von Alistair Plum, Vivien Wolter, Joëlle Weis und Hendrik Chudoba, 1–4, https://doi.org/10.5281/zenodo.7715368 (zugegriffen: 28. Dezember 2023).

Kraus, Karl. 2021. *Dritte Walpurgisnacht*, hg. von Bernhard Oberreither, https://kraus1933.ace.oeaw.ac.at/ (zugegriffen: 28. Dezember 2023).

Manjavacas Arévalo, Enrique. 2021. „Computational Approaches to Intertextuality.“ Doktorarbeit, Universität Antwerpen.

Nantke, Julia und Frederik Schlupkothen. 2018. „Zwischen Polysemie und Formalisierung: Mehrstufige Modellierung komplexer intertextueller Relationen als Annäherung an ein ‚literarisches‘ Semantic Web.“ In *DHd 2018: Kritik der digitalen Vernunft. 5. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2018), Köln*, hg. von Georg Vogeler, [1–4], https://doi.org/10.5281/zenodo.4622553 (zugegriffen: 28. Dezember 2023).

Nantke, Julia und Frederik Schlupkothen. 2019. FormIt: „Eine multimodale Arbeitsumgebung zur systematischen Erfassung literarischer Intertextualität.“ In *DHd 2019: Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd 2019), Frankfurt am Main, Mainz*, hg. von Patrick Sahle, 289–291, https://doi.org/10.5281/zenodo.2596095 (zugegriffen: 28. Dezember 2023).

Nantke, Julia und Ben Sulzbacher. 2020. „Mehrstufige Annotation literarischer Intertextualität jenseits der Textoberfläche.“ In *DHd 2020: Spielräume. Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2020), Paderborn*, hg. von Christof Schöch, 65–66, https://doi.org/10.5281/zenodo.3666690 (zugegriffen: 28. Dezember 2023).

Oberreither, Bernhard. 2018. „Zwei Überlegungen zur Konzeption einer Linked-Data-Ontologie für die Literaturwissenschaften.“ In *digital humanities austria 2018. empowering researchers*, hg. von Marlene Ernst, Peter Hinkelmanns, Lina Maria Zangerl und Katharina Zeppezauer-Wachauer, unter Mitarbeit von Verena Höller, 134–139, https://doi.org/10.1553/dha-proceedings2018s134 (zugegriffen: 28. Dezember 2023).

Oberreither, Bernhard. 2023a. „A Linked Data Vocabulary for Intertextuality in Literary Studies, with some Considerations Regarding Digital Editions.“ In *Digitale Edition in Österreich. Digital Scholarly Edition in Austria*, hg. v. Roman Bleier und Helmut W. Klug, 69–87. Norderstedt: BoD.

Oberreither, Bernhard. 2023b. *SemanticKraus – Connecting Kraus-Scholarship to the Semantic Web*, https://semantickraus.acdh.oeaw.ac.at/ (zugegriffen: 10. Januar 2024).

Prins, Yopie. 1999. *Victorian Sappho*. Princeton: Princeton University Press.

Reynolds, Margaret, Hg. 2001. *The Sappho Companion*. London: Vintage.

Robinson, David M. 1924. *Sappho and Her Influence*. Norwood: Plimpton.

Rüdiger, Horst. 1933. *Sappho. Ihr Ruf und Ruhm bei der Nachwelt*. Leipzig: Dieterich.

Rüdiger, Horst. 1967 [1934]. *Geschichte der deutschen Sappho-Übersetzungen*. Nachdr. Nendeln; Liechtenstein: Kraus.

Saake, Helmut. 1972. *Sapphostudien. Forschungsgeschichtliche, biographische und literarästhetische Untersuchungen*. München; Paderborn; Wien: Schöningh.

Tennis, Joseph F. 2004. „URIs and Intertextuality. Incumbent Philosophical Commitments in the Development of the Semantic Web.“ In *Knowledge Organization and the Global Information Society. Proceedings of the Eighth International ISKO Conference, London*, hg. v. Ia Cecilia McIlwaine, 103–108. Würzburg: Ergon.

Untner, Laura, Hg. 2023. *Sappho. Texte zur literarischen Rezeption im deutschsprachigen Raum*. Würzburg: Königshausen & Neumann.

Wagner, Benno, Alexander Mehler und Hanno Biber. 2016. „Transbiblionome Daten in der Literaturwissenschaft. Texttechnologische Erschließung und digitale Visualisierung intertextueller Beziehungen digitaler Korpora.“ In *DHd 2016: Modellierung –*

Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. 3. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2016), Leipzig, hg. v. Elisabeth Burr, https://www.dhd2016.de/abstracts/sektionen-005.html (zugegriffen: 10. Januar 2024).

Weitere Quellen

CIDOC CRM, http://www.cidoc-crm.org/cidoc-crm/ (zugegriffen: 28. Dezember 2023).

FRBRoo, http://iflastandards.info/ns/fr/frbr/frbroo/ (zugegriffen: 28. Dezember 2023).

INTRO, https://w3id.org/lso/intro/currentbeta# (zugegriffen: 28. Dezember 2023).

NEWW Women Writers, https://womenwriters.resources.huygens.knaw.nl/ (zugegriffen: 28. Dezember 2023).

The Women Writers Project, https://www.wwp.northeastern.edu (zugegriffen: 28. Dezember 2023).

Processing Qualitative Interview Data - Development of a Software Platform to Support Open Data in the Humanities

Mollenhauer, Sabina
sabina.mollenhauer@uni-vechta.de
Universität Vechta, Deutschland

Introduction

Humanities research in Germany and Europe aligns with overarching strategies by NFDI and EOSC with technologies that support aims towards open science with digitally available shared open data (RfII, 2023). Despite this, humanities in Germany are not well represented digitally with only a small portion of researchers and data participating (Wuttke, 2022). Needs assessments in the humanities and social sciences underscore that a substantial portion of research data comprises textual and survey data (Imeri and Danciu, 2017, 10; Schopfel and Prost 2016, 100; Strunk 2018, 22–23). This dissertation thus delves into the complexities of digital humanities research data processing, particularly within the German context, with a focus on qualitative interview transcripts. It examines the challenges posed by limited software support for digital open data sharing.

While researchers acknowledge the importance of digital tools and value the reusability of research data, especially if tied to publications (Imeri and Danciu 2017, 19–20), barriers such as GDPR-compliance and resource-intensive de-identification processes hinder digital collection and preservation efforts (ibid., 28).

In essence, the dissertation project explores the user and system requirements (Maiden, 2008) for humanities researchers willing to share qualitative interview data and evaluates the extent to which these requirements can be met by targeted software support with the aim of developing such software.

State of Research and Software

In current practice, qualitative interview data undergoes predominantly manual processing, often aided by personal consultation (Adena et al., 2020). This includes steps such as transcription, de-identification, and modeling before sharing or publication. The decision-making process for GDPR compliant processing can theoretically be aided by interactive virtual assistance BERD@NFDI (Herklotz et al., 2021). Tools exist for de-identification through Named Entity Recognition or privacy-preserving data publishing, as well as for interoperability processing through metadata enrichment and data modeling (Pilán et al., 2022; Kuchma, 2018; Strathern et al., 2020; Prasser et al., 2020; Templ, 2008). However, these tools have not significantly increased the sharing of qualitative interview data in humanities research.

NFDI consortia aim to provide support and create technical infrastructure, yet an unfulfilled usability gap remains. NFDI4Culture provides user stories addressing digital data handling in the digital humanities, but it lacks coverage on the digitization of qualitative interview data (NFDI4Culture, 2023). NFDI4Memory has identified problem stories, including one concerning the anonymization of qualitative empirical data. However, this story lacks an assigned task area (Paulmann et al., 2022). Further research is needed to discover requirements that bridge this usability gap.

Research Project and Questions

The research project's overarching question focuses on identifying barriers faced by humanities researchers open to sharing their data and developing methods to mitigate these barriers through a technically guided workflow. An exploration of available infrastructure and services raises questions about existing technological solutions within the digital data pipeline and their adaptation for qualitative interview data processing.

Answering these questions will guide tool development within one specific aspect of the digital data pipeline. For instance, anonymization and automated data modeling are

complex processes in and of themselves. At the same time, another central aspect of the digital data pipeline is the integration of each of these individual steps into a user-centrically designed user interface. Further research must therefore emphasize identifying the most feasible solution to bridge these gaps in the digital data pipeline of qualitative interview data generated in humanities research.

To address these overarching questions, the dissertation research adopts an iterative process. Initial exploratory steps involve literature reviews of existing needs assessments in the humanities, coupled with expert interviews. The methodology, inherently qualitative and based on grounded theory (Truschkat et al., 2011), ensures adaptability to the evolving landscape of digital data pipelines for open data.

The iterative process of software development begins with data collection through qualitative expert interviews (Helfferich, 2022) and contextual inquiry, utilizing qualitative methods from psychology, sociology, and anthropology (Raven and Flanders, 1996; DeBellis and Haapala, 1995). In conjunction with classic software requirement engineering (Bednar and Welch, 2009; ISO/IEC/IEEE 29148, 2011), this process is referred to as a context-sensitive software development process (Siadat and Song, 2012).

Initial evaluation of existing tools and interviews with technical experts is carried out as system requirement analysis. This analysis, focusing on the diverging architectural paradigms in open research data, aims to ensure that the existing open data pipeline is optimized. The goal is for the developed tool to be part of a structure that facilitates its persistence, remaining relevant to future developments in existing architectural paradigms (Diepenbroek et al., 2023; Machado, Costa, and Santos, 2022).

Bibliographie

Adena, Maja, Christian Aßmann, Doris Bambey, Katarina Blask, Andreas Blätte, Michael Bosnjak, Daniel Buck, et al. “Consortium for the Social, Behavioural, Educational, and Economic Sciences (KonsortSWD).” Publisher: Zenodo, July 31, 2020. visited on 04/17/2023. https://doi.org/10.5281/zenodo.3968372. https://zenodo.org/records/3968372.

Bednar, Peter M and Christine Welch. 2008. Contextual inquiry and requirements shaping. In *Information Systems Development: Challenges in Practice, Theory, and Education Volume 1* (pp. 225-236). Boston, MA: Springer US.

DeBellis, Michael, and Christine Haapala. "User-centric software engineering." *IEEE Expert* 10, no. 1 (1995): 34-41.

Diepenbroek, Michael, Ivaylo Kostadinov, Bernhard Seeger, Frank Oliver Glöckner, Marius Alfred Dieckmann, Alexander Goesmann, Barbara Ebert, Sonja Schimmler, and York Sure-Vetter. "Towards a Research Data Commons in the German National Research Data Infrastructure NFDI: Vision, Governance, Architecture." In *Proceedings of the Conference on Research Data Infrastructure*, vol. 1. 2023.

Helfferich, Cornelia. "Leitfaden-und experteninterviews." In *Handbuch Methoden der empirischen Sozialforschung*, pp. 875-892. Wiesbaden: Springer Fachmedien Wiesbaden, 2022.

Herklotz, Markus, Lars Oberländer, Irene Schumm, and Renat Shigapov. 2021. iVA: Ein interaktiver Virtueller Assistent von BERD@ BW zur Aufbereitung von Rechtsfragen im Bereich Open Science. *Tage 2021*

Imeri, Sabine and Ida Danciu. 2017. Open Data. Forschungsdatenmanagement in den ethnologischen Fächern. Auswertung einer Umfrage des Fachinformationsdienstes Sozial- und Kulturanthropologie an der Universitätsbibliothek der Humboldt-Universität zu Berlin 2016. Teil I: Statistiken. http://www.evifa.de/cms/fileadmin/uploads/Umfrage_Bericht_Statistiken_1.0_14-06-2017.pdf (visited on 07/13/2023).

ISO/IEC/IEEE 29148: 2011. "ISO/IEC/IEEE International Standard—Systems and software engineering—Life cycle processes—Requirements engineering."

Kuchma, Iryna. 2018. "OpenAIRE Services for Open Science." *ITlib: Informacne Technologie a Kniznice* 4.

Machado, Inês Araújo, Carlos Costa, and Maribel Yasmina Santos. "Data mesh: concepts and principles of a paradigm shift in data architectures." *Procedia Computer Science* 196 (2022): 263-271.

Maiden, Neil. "User requirements and system requirements." IEEE Software 25, no. 2 (2008): 90-91.

NFDI4Culture. 2023. User Stories - NFDI4Culture. url: https://nfdi4culture.de/de/ressourcen/user-stories.html (visited on 07/13/2023).

Paulmann, Johannes, John Wood, Klaus Ceynowa, Fabian Cremer, Silvia Daniel, Daniel Fähle, Barbara Fichtl, Peter Haslinger, Torsten Hiltmann, Rüdiger Hohls, Ursula Lehmkuhl, Marina Lemaire, Gerald Maier, Ole Meiners, Gisela Minn, Katrin Moeller, Andreas Neuburger, Claudia Prinz, Matthias Razum, Harald Sack, Johannes Sauter, Hildegard Schäffler, Eva Schlotheuber, Stefan Schmunk, Arnost Stanzel, Helmuth Trischler and Thorsten Wübbena. 2022. "NFDI4Memory. Consortium for the historically oriented humanities. Proposal for the National Research Data Infrastructure (NFDI)."

Prasser, Fabian, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. 2020. "Flexible data anonymization using ARX—Current status and challenges ahead." Software: Practice and Experience 50, no. 7: 1277-1304.

Raven, Mary Elizabeth, and Alicia Flanders. 1996. "Using contextual inquiry to learn about your audiences." *ACM SIGDOC Asterisk Journal of Computer Documentation* 20, no. 1: 1-13.

RfII. 2023. RfII-Bericht „Föderierte Dateninfrastrukturen für die wissenschaftliche Nutzung“ – März 2023. url: https://rfii.de/download/rfii-bericht-foederierte-dateninfrastrukturen-fuer-die-

wissenschaftliche-nutzung-maerz-2023/ (visited on 04/17/2023).

Siadat, Seyed Hossein, and Minseok Song. "Understanding requirement engineering for context-aware service-based applications." (2012).

Schopfel, Joachim, and Hélène Prost. 2016. "Research data management in social sciences and humanities: A survey at the University of Lille (France)."

Strathern, Wienke, Moritz Issig, Kati Mozygemba, and Jürgen Pfeffer. 2020. QualiAnon - The Qualiservice tool for anonymizing text data. en. Tech. rep. TUM-I2087.

Strunk, Sonja. "Evaluationsbericht 2018: Bedarfserhebung Fachinformationsdienst (FID) Soziologie." (2018): 41.

Truschkat, Inga, Manuela Kaiser-Belz, and Vera Volkmann. "Theoretisches Sampling in Qualifikationsarbeiten: Die Grounded-Theory-Methodologie zwischen Programmatik und Forschungspraxis." *Grounded theory reader* (2011): 353-379.

Wuttke, Ulrike. 2022. "Wege bereiten, vermitteln und Denkräume schaffen! Reflexionen zu institutionellen und infrastrukturellen Erfolgsfaktoren für Digital Humanities an deutschen Universitäten auf Grundlage von Expert*inneninterviews." *Zeitschrift für digitale Geisteswissenschaften* 2022, no. 7

Quantitative Ansätze zur Untersuchung der frühneuzeitlichen Dramengeschichte

Giovannini, Luca
giovannini@uni-potsdam.de
Universität Potsdam, Deutschland / Universität Padua, Italien
ORCID: 0000-0003-2444-7192

Forschungsfrage

In den letzten Jahren hat sich die quantitative Forschung zum Drama als ein wichtiger Teil der computergestützten Literaturwissenschaft etabliert. Zum frühneuzeitlichen Drama gibt es allerdings noch wenig umfassende Studien zu verzeichnen, die über die Grenzen der nationalen Philologien hinausgehen und quantitative Beiträge zur Komparatistik liefern. Davon ausgehend ist Ziel des Promotionsvorhabens, eine quantitative Geschichte des frühneuzeitlichen europäischen Theaters zu skizzieren, die die Evolution der verschiedenen Nationalliteraturen vergleichend rekonstruiert.

Als Ausgangspunkt der Dissertation dient die u. a. von Moretti (1994) verbreitete These, dass die Entwicklung des europäischen Theaters in der frühen Neuzeit als ein Prozess der biologischen Artbildung interpretiert werden kann. Im Laufe des 17. Jahrhunderts, so Moretti, wurde ein europaweites Modell der Tragödie, das aus der Antike und dem Mittelalter übernommen wurde, durch nationale Varianten wie das deutsche Trauerspiel oder die französische *tragédie classique* ersetzt. Diese Varianten sollen sich inhaltlich, stilistisch und formal voneinander unterscheiden.

Dank der steigenden Textverfügbarkeit und den Fortschritten in den computationellen Methoden lässt sich diese bisher unhinterfragte literaturgeschichtliche These nun empirisch überprüfen. Daher lauten die konkreten Fragestellungen, wie eine solche Entwicklung der dramatischen Formen mit quantitativen Methoden nachzuvollziehen ist und ob der von Moretti beschriebene „Verzweigungsprozess“ nicht nur für die Tragödie, sondern auch für die Komödie und andere Gattungen stattgefunden hat.

Korpus

Das Promotionsvorhaben erfolgt im Umfeld des *DraCor*-Projekts (Fischer et al. 2017) und profitiert von den dort entwickelten Pipelines und Tools für die quantitative Dramenanalyse. Das für die Dissertation erstellte *Early Modern Drama Corpus* (*EmDraCor*) orientiert sich an der Bauweise bestehender *DraCor*-Sammlungen und umfasst derzeit 150 Theaterstücke, die in fünf europäischen Sprachen geschrieben und zwischen 1561 und 1710 entstanden sind.

Die *EmDraCor*-Texte wurden größtenteils aus digitalen Sammlungen dramatischer Werke sowie aus wissenschaftlichen Open-Access-Editionen abgeleitet. Lücken in der Textverfügbarkeit wurden durch die Erstkodierung von gescannten Stücken geschlossen. Die noch nicht in *DraCor* vorhandenen Theaterstücke wurden gemäß der flexiblen Pipeline, die in Börner et al. (2023a) beschrieben wird, in *DraCor*-kompatible XML-TEI-Dateien umgewandelt. Die Texte wurden danach nicht direkt zu *DraCor* hinzugefügt, sondern auf einer lokalen Instanz der Plattform eingesetzt, die aus einem Docker-Image (Börner et al. 2023b) erstellt wurde. Damit war es möglich, auf alle Funktionen der API zuzugreifen, ohne den umfangreichen Onboarding-Prozess durchlaufen zu müssen; damit könnte auch das textuelle Markup auf die spezifischen Bedürfnisse der Studie angepasst werden.

Methoden und Ausblick

Methodologisch inspiriert sich das Promotionsprojekt an den Forschungsansätzen des quantitativen Formalismus (Allison et al. 2011): Im Fokus steht die Struktur dramatischer Texte, d.h. eine der Komponenten, anhand derer die Entwicklung der Gattung Drama gezeigt werden kann. Da sich diese Dimension auf nicht sprachbedingte Elemente

bezieht, etwa Figurenkonstellationen oder Redeverteilung, kann man sie produktiv für eine komparative Studie verschiedener Nationaltraditionen einsetzen.

Als zentrale analytische Praxis für die Untersuchung der Variation des europäischen Dramas wird dann die Vektorisierung von Theaterstücken nach ihren strukturellen Merkmalen eingesetzt. Ähnlich wie bei *word embeddings* im *Natural Language Processing* zielt diese Methode darauf ab, die Stücke durch eine Reihe von zahlenbasierten Metriken in einen Vektor zusammenzufassen und damit verschiedene Berechnungen durchzuführen. Die Auswahl der Metriken für die Vektoren hängt von der erarbeiteten Operationalisierung des Begriffes „Drama" ab; hier wird versucht, die von Kretz (2015) genannten „elementaren Bausteine" der Gattung (Dialog, Figur, Handlung) in Betracht zu ziehen, um möglichst viele Textaspekte in den Vektoren abzubilden. Neben den Metriken, die von der DraCor-API ausgegeben werden und die überwiegend netzwerkbasiert sind (siehe u. a. Trilcke et al. 2015), werden auch einige im Kontext anderer Studien entwickelte Messwerte einbezogen (Szemes und Vida 2022, Trilcke et al. 2017, Algee-Hewitt 2017).

Im Rahmen des Promotionsvorhabens sind zwei Anwendungsmöglichkeiten für die vektorisierten Stücke vorgesehen. Zum einen können mithilfe verschiedener Abstandsmessungen (z. B. euklidischer Abstand oder Kosinus-Ähnlichkeit) die Distanzen zwischen den Vektoren berechnet werden, wobei ein größerer Abstand auf eine größere strukturelle Unterschiedlichkeit hinweisen soll. Zum anderen ist es möglich, die Vektoren durch Techniken wie die Hauptkomponentenanalyse (PCA) auf einer niedrigdimensionalen Ebene zu visualisieren, um Cluster zu identifizieren.

Erste Ergebnisse zeigen, dass ein Narrativ von kontinuierlicher Verzweigung zwischen literarischen Traditionen nicht ohne Einschränkungen vertretbar ist. Obwohl eine Tendenz zur Diversifizierung bemerkbar ist, ist die Gattungsevolution scheinbar durch komplexe und mehrschichtige Dynamiken geprägt. Auch wenn die Arbeit mit einem kleinen, aber sorgfältig kuratierten Korpus wie *EmDraCor* lediglich als erster Schritt zu einer umfassenden Analyse des frühneuzeitlichen Dramas betrachtet werden muss, wird dennoch schlussfolgernd angenommen, dass der auf die Vektorisierung von Theaterstücken basierte Arbeitsablauf eine effektive Methode zur Untersuchung der strukturellen Entwicklung dramatischer Formen darstellen kann.

Bibliographie

Algee-Hewitt, Mark. 2017. „Distributed Character: Quantitative Models of the English Stage, 1550–1900". New Literary History 48 (4): 751–82. https://doi.org/10.1353/nlh.2017.0038.

Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti, und Michael Witmore. 2011. „Quantitative formalism: an experiment". LitLab Pamphlets #1. https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Börner, Ingo, Frank Fischer, Luca Giovannini, Carsten Milling, Christopher Lu, Henny Sluyter-Gäthje, Daniil Skorinkin, und Peer Trilcke. 2023a. „Onboard onto DraCor: Prototyping workflows to homogeneize drama corpora". In DHd2023 Book of Abstracts. Belval/Trier: Universität Trier. https://zenodo.org/record/7715333.

Börner, Ingo, Peer Trilcke, Carsten Milling, Frank Fischer, und Henny Sluyter-Gäthje. 2023b. „Dockerizing DraCor: A Container-based Approach to Reproducibility in Computational Literary Studies". In *DH2023 Book of Abstracts*. Graz: Universität Graz. https://zenodo.org/record/8107836.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, und Peer Trilcke. 2019. 'Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama'. In *DH2019 Book of Abstracts*. Utrecht: Utrecht University. https://doi.org/10.5281/ZENODO.4284002.

Kretz, Nicolette. 2012. „Grundelemente (1): Bausteine des Dramas (Figur, Handlung, Dialog)". In: *Handbuch Drama*, herausgegeben von Peter W. Marx, 105–21. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-00512-0_9

Szemes, Botond, und Bence Vida. 2022. „Tragic and Comical Networks: Clustering Dramatic Genres According to Structural Properties". *Workshop on Computational Drama Analysis: Achievements and Opportunities*. Köln. https://arxiv.org/ftp/arxiv/papers/2302/2302.08258.pdf.

Trilcke, Peer, Frank Fischer, und Dario Kampkaspar. 2015. 'Digital Network Analysis of Dramatic Texts'. In: *DH2015 Book of Abstracts*. Sydney: University of Western Sydney. https://doi.org/10.5281/ZENODO.3627710.

Trilcke, Peer, Frank Fischer, Mathias Göbel, Dario Kampkaspar, und Christopher Kittel. 2017. „Netzwerkdynamik, Plotanalyse – Zur Visualisierung und Berechnung der ›progressiven Strukturierung‹ literarischer Texte". In: *DHd2017 Book of Abstracts*. Bern: Universität Bern. https://doi.org/10.5281/ZENODO.4622799.

Vernetzte Finanzen – Historische Finanzdokumente und aktuelle Herausforderungen der computergestützten Erschließung

Mischka, Bernadette
bernadette.mischka@ur.de
Universität Regensburg, Deutschland

Die Cash Book Collection des Rothschild Archiv London

Die Cash Book Collection (1810 – 1950) des Rothschild Archiv London umfasst insgesamt 325 Bände an handschriftlichen Aufzeichnungen zu täglichen Transaktionen des Unternehmens N M Rothschild & Sons. Die Cash Books sind genormte und standardisierte Werkzeuge im sogenannten double-entry-bookkeeping und sind sowohl Teil der Buchführung innerhalb von Institutionen, aber auch in externen Prüfungsprozessen (Fieldhouse, 1922, 31 ff.). Neben der Referenz der betreffenden Accounts im eigenen Buchführungssystem werden in genormten Tabellen jeweils das Datum, der Name der ein- oder auszahlenden Entität sowie der Vermerk von Discounts und der Einzahlungsart vermerkt. Zwischen den Ausgaben des Unternehmens, Handelskrediten und Staatsanleihen von Nationalstaaten zur Finanzierung von Kriegen oder von Infrastruktur finden sich private Ausgaben der Familie Rothschild, wie zum Beispiel Friseurbesuche, philanthropischen Tätigkeiten oder dem Erwerb von exotischen Pflanzen. Es lassen sich somit zum einen einzelne Finanzgeschäfte nachvollziehen und Indikatoren für das zeitgenössische Alltagsleben ablesen, aber gleichzeitig auch politische und soziale Netzwerke von Agenten sowie zeitgenössische Finanzierungsgeschäfte und die damit einhergehende Professionalisierung der Buchführung der wachsenden Industriestaaten im 19. Jahrhundert abbilden (Liedtke, 2006, 15 ff. sowie Flandreau & Zumer, 2016). Bestände wie die Cash Book Collection sind vor allem aufgrund des großen Volumens an manuell auszuwertenden Daten und dem damit verbundenen Ressourcenaufwand kaum für die Forschung zugänglich.

Historische Finanzdokumente in den Digital Humanities

Standardisierte und tabellarische Daten in historischen Finanz- und Verwaltungsdokumenten eignen sich als Big Data Bestände der Geschichte auf den ersten Blick besonders gut für computergestützte Auswertungsverfahren. Trotzdem waren die Quellengattungen bisher nur vereinzelt Bestandteil von automatisierten Erschließungsverfahren. Bisher waren vor allem digitale Editionen mittelalterlicher und frühneuzeitlicher Quellenbestände, wie zum Beispiel die Augsburger Baumeisterbücher[1] oder die Jahrrechnungen der Stadt Basel[2], Teil wissenschaftlicher Überlegungen und Projekte (Sarnowky, 2016, 7 ff.). Vor allem Herausforderungen in der computergestützten Erkennung komplexer Layouts sowie der unzureichenden Erkennungsrate von Handschriften stellen häufig ein Hindernis für eine automatisierte Auswertung von Wirtschaftsdokumenten im Sinne eines „Distant-Counting-Ansatzes“[3] dar. Anders als bei der Arbeit mit literarischen Texten sind Verfahren zur Korrektur von Zahlen im Post-Processing nur bedingt möglich und verhindert somit das automatisierte Auslesen von Zahlenwerten. Handschriftliche tabellarische Aufzeichnungen sind in den letzten Jahren zwar vermehrt in den Fokus von Erschließungsprojekten gerückt, jedoch durch ihre hohe Heterogenität und individuellen Eigenschaften meist mit explorativen Ansätzen und einem hohen Forschungsaufwand verbunden (Lehenmeier et al., 2020 sowie Constum et al. 2022). Es stellt sich somit die Frage, wie bestandserhaltende Einrichtungen intern mit handschriftlichen tabellarischen Finanzdokumenten arbeiten können oder ob diese weiterhin nur manuell analysierbar und somit durch den hohen Ressourcenaufwand größtenteils für die Forschung versperrt bleiben.

Citizen Science zur Erstellung von Ground Truth Daten

Das Dissertationsvorhaben möchte neben der Kontextualisierung des historischen Bestandes und seiner Möglichkeiten als Datengrundlage für die Geschichtswissenschaft, eine Erschließungsstrategie der handschriftlichen Tabellen mit aktuell publizierten Ansätzen und Tools erforschen. Insgesamt wurden für das Projekt 30 Jahre zwischen 1810 und 1915 in 89 Bänden des Bestandes digitalisiert. Eine erste große Herausforderung für das Projekt war die Generierung von Ground-Truth-Daten zum Training eines korpuseigenen Handschriftenmodells. Obwohl sich in den letzten Jahren immer mehr Finanzinstitutionen mit ihren historischen Beständen beschäftigt haben, waren keine publizierten Leitfäden oder Richtlinien zur Transkription von domänenspezifischen Sonderzeichen und Abkürzungen für den Dokumenttyp im 19. Jahrhundert vorhanden. Zur Generation von Ground-Truth-Daten zum Training eines Handschriftenmodells wurde in einem Citizen-Sci-

ence-Projekt[4] insgesamt 460 Seiten des Bestandes transkribiert und in einem Double-Keying-Verfahren zur Modellierung vorbereitet. Der hohe Organisationsaufwand der Betreuung sowie der Qualitätssicherung wurde hierbei ebenso deutlich wie das große Potenzial des Austausches mit den Teilnehmenden über die Dokumente, ihre Struktur und auftretende Sonderfälle in der Transkription für den Aufbau eines erweiterbaren und nachhaltigen Datensatzes.

Aktueller Stand und Ausblick

Aktuell befindet sich das Projekt in der Erschließungsphase des Bestandes. Es müssen sowohl die gewählten Ansätze der Handschriftenerkennung als auch der Tabellenerschließung evaluiert und anschließend eine sinnvolle Datenstruktur für die Dokumente im Archiv gefunden werden. Die Daten lassen sich sowohl mit Beständen innerhalb der unternehmensinternen Verwaltung, als auch mit historischen Ereignissen, Personen und Institutionen außerhalb des Unternehmens vernetzen. Für das Projekt stellt sich somit zudem die Frage einer gewinnbringenden Datenvisualisierung, die in den kommenden Jahren mit Entwicklungen und Neuerungen der Digital Humanities mitwachsen kann. Anhand des aktuellen Projektstandes lassen sich zum einen Erfahrungsberichte und -werte in der Generierung von Ground-Truth-Daten in einem Citizen-Science-Projekt erfassen, jedoch auch die Herausforderungen von Digital Humanities Projekten in Privatarchiven hinsichtlich Urheberrechten und möglicher Datenstrukturen abbilden. Durch die domänenspezifischen Eigenschaften der Dokumente kann das Projekt Hinweise und Leitfäden für weitere Forschungsvorhaben liefern, die jedoch mit einer möglichst großen Forschungscommunity diskutiert und evaluiert werden müssen.

Fußnoten

1. https://augsburger-baumeisterbuecher.de/
2. https://gams.uni-graz.at/context:srbas
3. Idee der automatisierten Auswertung der Projektdaten nach Vorbild des Distant-Reading-Ansatzes nach Franco Moretti.
4. https://cashbooks.app.uni-regensburg.de/

Bibliographie

Constum, Thomas, Nicholas, Kempf, Thierry, Paquet, Pierrick, Tranouez, Clément, Chatelain, Sandra, Brée, Francois, Merveille. 2022. “Recognition and Information Extraction in Historical Handwritten Tables: Toward Understanding Early 20 th Century Paris Census”. *Document Analysis Systems 2022 proceedings.*143-157. Cham: Springer.

Fieldhouse, Arthur. 1922. „The Student’s Elementary Commercial Book-Keeping. Accounting and Banking“. London: Simpkin, Marshall & Co.

Flandreau, Marc, Frédérik, Zumer. 2016. „Media Manipulation in Interwar France: Evidence from the Archive of Banque de Paris et des Pays-Bas, 1914–1937“. *Contemporary European History 25,1*, 11-36. New York: Cambrisge University Press.

Lehenmeier, Constantin, Manuel, Burghardt, Bernadette, Mischka. 2020. „Layout Detection and Table Recognition – Recent Challenges in Digitizing Historical Documents and Handwritten Tabular Data”. *Digital Libraries for Open Knowledge. TPDL 2020. Lecture Notes in Computer Science, Vol 12246*, 229-242. Cham: Springer.

Liedtke, Rainer. 2006. “N M Rothschild & Sons. Kommunikationswege im europäischen Bankenwesen im 19. Jahrhundert“. Köln: Böhlau Verlag.

Sarnowsky, Jürgen. 2016. “Einführung”. *Konzeptionelle Überlegungen zur Edition von Rechnungen und Amtsbüchern des späten Mittelalters*. 7-12. Göttingen: V&R unipress.

Zur Perspektive in Erzähltexten. Ein Ansatz der Computational Literary Studies.

Sluyter-Gäthje, Henny
sluytergaeth@uni-potsdam.de
Universität Potsdam, Deutschland
ORCID: 0000-0003-2969-3237

1. Projektübersicht

Das ästhetische Überformen einer Geschichte zu einer Erzählung setzt die Gestaltung der Perspektive voraus, aus der die Geschichte erzählt wird. Die Perspektive beeinflusst, wie mittelbar eine Geschichte erzählt wird (z.B. Lubbock, 2006 [1921]; Friedman, 1955; Stanzel, 1985) und ist, je nach narratologischer Modellierung, dadurch bestimmt, wer die wahrnehmende und wertende Instanz ist (z.B. Uspenskij, 1975; Rimmon-Kenan, 1983; Schmid, 2014), wer für die Vermittlung von Wissen zuständig ist (z.B. Friedman, 1955), bzw. in welchem Verhältnis das Wissen der Figuren und der Erzählinstanz steht (Genette, 2010), ob Raum und Zeit an Figuren der erzählten Welt geknüpft sind (z.B. Uspenskij, 1975; Rimmon-Kenan, 1983; Schmid, 2014) und inwiefern die Sprache durch die Figuren oder die Erzählinstanz geprägt ist (z.B. Uspenskij, 1975; Schmid, 2014). Die Ausgestaltung der Perspektive beeinflusst dementsprechend maßgeblich die Bewertung von Geschehnis-

sen in der erzählten Welt und prägt somit auch die literaturwissenschaftliche Interpretation, etwa weil sie auf Ebene der *histoire* auf den Wirklichkeitsstatus von Dingen und Ereignissen in der erzählten Welt Einfluss nimmt oder die Zuverlässigkeit des Erzählens bedingen kann.

Die automatisierte Identifikation von Perspektive in Erzähltexten, zu der es zum jetzigen Zeitpunkt sowohl aus den *Computational Literary Studies* (CLS) als auch aus dem *Natural Language Processing* (NLP) nur reduzierte Ansätze gibt, die z.B. in der Unterscheidung einer Ich- oder Er-Erzählung liegen (z.B. Eisenberg und Finlayson, 2016; Chen und Bunescu, 2022), würde es erlauben, eine Vielzahl an Texten in kurzer Zeit untersuchbar und so Muster, z.B. im Hinblick auf Epochen, erkennbar zu machen. Die Entwicklung eines solchen Systems ist das Ziel dieses Promotionsprojekts.

Daraus ergeben sich zwei Fokusse:

1. Die Grundlage für die Implementierung eines automatischen Systems ist ein mit Perspektiven-Informationen angereichertes Korpus, das für regelbasierte oder *Machine Learning*- Systeme als Entwicklungs- und Evaluationskorpus genutzt werden kann. Um die Annotation eines Korpus zu ermöglichen, muss die theoretische Modellierung von Perspektive operationalisiert werden, d.h. abstrakte Begriffe wie ‘Wahrnehmen’ oder ‘Wissen’ werden in einem Prozess der Formalisierung konkretisiert und konzentriert, sodass Indikatoren abgeleitet werden können, die an der Textoberfläche untersuchbar sind. Durch den Prozess werden einerseits existierende Perspektiventheorien auf den Prüfstand gestellt, andererseits erhoffe ich mir, sowohl durch den Annotationsprozess als auch durch die statistische Auswertung der Annotationen neue Erkenntnisse zum Phänomen ‘Perspektive’ gewinnen zu können. Darin begründet sich der methodologisch-theoretische Ansatz der Arbeit.
2. In der Implementierung eines Systems zur automatischen Identifizierung von Perspektive besteht der zweite, computerlinguistische Teil der Arbeit. Dieser umfasst einerseits die Frage, in welche Unteraufgaben sich die Perspektivenerkennung gliedern lässt und wie sich diese gestalten, andererseits den Vergleich von verschiedenen Ansätzen.

2. Vorgehen

Das Vorhaben ist in drei Arbeitspakete unterteilt: Die Erarbeitung der narratologischen Modellierung, die Operationalisierung und Annotation sowie die Implementierung eines Systems zur automatischen Identifizierung.

Am Anfang steht die Erarbeitung narratologischer Modellierungen von Perspektive unter dem Gesichtspunkt der Operationalisierbarkeit. Dabei werden etablierte Definitionen auf den Grad der Abstraktheit bzw. in Bezug auf ihre Textnähe geprüft. In dem Projekt wird mit der Modellierung von Schmid (2014) gearbeitet, der fünf Parameter (räumlich, zeitlich, sprachlich, perzeptiv, ideologisch), die die Perspektive bedingen, unterscheidet. Der sprachliche und ideologische Parameter sind weiter unterteilt, für den zeitlichen und räumlichen Parameter werden textuelle Indikatoren zur Erkennung genannt. Aufgrund dieser Ausdifferenziertheit und der Textnähe bildet die Modellierung nach Schmid eine gute Grundlage für die Operationalisierung. Gius (2015) erarbeitete darauf aufbauend bereits ein erstes, rudimentäres Tagset zur Bestimmung von Perspektive.

Ausgehend von Schmids Modellierung werden – orientiert an Reiter (2021) – Annotationsrichtlinien formuliert, die auf ein Korpus von 6 deutschsprachigen Prosatexten angewendet werden. Hierfür wird dasselbe Korpus verwendet wie für die Untersuchung von Ereignissen (Vauth et al., 2023)[1], da es über eine größere Zeitspanne reicht (1797-1915) und sowohl Autoren als auch Autorinnen beinhaltet. Darüber hinaus können am Projektende die bestehenden Annotationen korreliert und so eine weitere Bandbreite an Forschungsfragen untersucht werden. Die Annotation wird von zwei Annotatorinnen in CATMA (Gius et al., 2023) durchgeführt, die sich an dem von Gius und Jacke (2017) entwickelten Zyklus orientieren. Zum jetzigen Zeitpunkt wurde bereits ein Annotationsdurchgang und eine Überarbeitung der Richtlinien durchgeführt.

Schließlich folgt der computerlinguistische Teil der Arbeit, in dem ich zuerst nach Fragestellungen in der CL, den CLS und dem NLP suche, die derjenigen der Identifizierung von Perspektive bzw. ihren Unteraufgaben ähneln (z.B. Sentiment Analysis). Auch suche ich nach Systemen, die ggf. für die Perspektivenerkennung nutzbar gemacht werden können. Es folgt die Implementierung bzw. Erweiterung von Systemen zur Perspektivenerkennung. Zu diesem Zeitpunkt ist eine Vielzahl an Ansätzen denkbar, angefangen mit regelbasierten Systemen bis hin zu *Deep Learning* -Systemen, die mit *Large Language Models* arbeiten.

Fußnoten

1. Der EvENT-Datensatz ist verfügbar unter: https://github.com/forTEXT/EvENT_Dataset

Bibliographie

Chen, Mike und Razvan Bunescu. 2021. “Changing the narrative perspective: From deictic to anaphoric point of view.” *Information Processing & Management 58.4: 102559*. DOI: https://doi.org/10.1016/j.ipm.2021.102559 .

Eisenberg, Joshua und Mark Finlayson. 2016. “Automatic Identification of Narrative Diegesis and Point of View.” In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)* , 36-46. DOI: 10.18653/v1/W16-5705 .

Friedman, Norman. 1955. "Point of View in Fiction: The Development of a Critical Concept." *PMlA* 70.5: 1160-1184. URL: https://www.jstor.org/stable/459894 .

Genette, Gérard. 2010 [frz. 1972, 1983]. *Die Erzählung* . 3., durchgesehene und korrigierte Auflage, München: Fink.

Gius, Evelyn. 2015. *Erzählen über Konflikte. Ein Beitrag zur digitalen Narratologie.* Berlin / Boston: De Gruyter.

Gius, Evelyn und Janina Jacke. 2017. "The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis." *International Journal of Humanities and Arts Computing* 11.2: 233-254. DOI: 10.3366/ijhac.2017.0194.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Mareike Schumacher, Dominik Gerstorfer. 2023. CATMA 7 (Version 7.0). Zenodo. DOI: 10.5281/zenodo.1470118 .

Lubbock, Percy. 2006 [1921]. *The Craft of Fiction.* Project Gutenberg. https://www.gutenberg.org/ebooks/18961 (letzter Zugriff am 18.07.2023).

Reiter, Nils. 2020. "Anleitung zur Erstellung von Annotationsrichtlinien." In *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt,* hg. von Nils Reiter, Axel Pichler und Jonas Kuhn, 193-203. Berlin / Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110693973 .

Rimmon-Kenan, Shlomith. 1983. *Narrative Fiction. Contemporary Poetics.* London / New York: Methuen.

Schmid, Wolf. 2014. *Elemente der Narratologie.* 3., erweiterte und überarbeitete Auflage. Berlin / Boston: De Gruyter.

Stanzel, Franz K. 1985. *Theorie des Erzählens.* 3. durchgesehene Auflage. Göttingen / Zürich: Vandenhoeck und Ruprecht.

Uspenskij, Boris A. 1975. *Poetik der Komposition: Struktur des künstlerischen Textes und Typologie der Kompositionsform.* Frankfurt am Main: Suhrkamp.

Vauth, Michael, Evelyn Gius und Hans Ole Hatzel. 2023. forTEXT/EvENT_Dataset: v.1.2 (v1.2). Zenodo. https://doi.org/10.5281/zenodo.8063719 .

Posterpräsentationen

A Low-Cost Reflectance Transformation Imaging Dome

Winslow, Sean
sean.winslow@uni-graz.at
Karl-Franzens-Universität Graz, Österreich
ORCID: 0000-0002-5114-0594

Krottmaier, Sina
sina.krottmaier@edu.uni-graz.at
Karl-Franzens-Universität Graz, Österreich
ORCID: 0000-0002-7966-7996

Tosques, Fabio
fabio.tosques@uni-graz.at
Karl-Franzens-Universität Graz, Österreich

Zuanni, Chiara
chiara.zuanni@uni-graz.at
Karl-Franzens-Universität Graz, Österreich
ORCID: 0000-0003-4027-0278

Reflectance Transformation Imaging (RTI) has many uses in imaging heritage items such as coins, seals, and small archaeological artifacts. It allows for imaging in a way that preserves and highlights small surface details, which are often of high significance for the study of such objects. While RTI can be done by hand, this process is time-consuming and best reserved for larger objects; smaller objects are generally imaged with a dome which provides a set of lights in enough directions that all the directions of light can be reconstructed programmatically. These domes are not readily and inexpensively available; commercial examples can cost in the neighborhood of €10.000–30.000. This means that many scholars and institutions with limited need for RTI have no access to them either for production of images or for students to learn and experiment. A number of projects for self-made RTI domes are documented (primarily online), but most of these domes require specialized electronic skills, soldering, and are camera-specific.

The SimpleRTIDome project is putting together a dome which will require no soldering, combining 3D printed parts with ready-made parts from the microcontroller and single-board computer (SBC) space, designed to be assembled by individuals with no particular “Maker” skills. All of the parts for the prototype dome came to around €200, and the released version will target a total cost in parts and printing of under €300. The use of a single-board computer to drive the project, as opposed to specialized electronics or hacky workarounds, means that the dome has access to the full range of open-source software, which is used to implement a workflow that sees a python script control the workflow of the system, controlling all lights and managing the file structure. The use of gphoto2 means that a broad range of DSLRs and even consumer-grade digital cameras are already supported without any additional work by the end user, and the prototype camera is also being designed to work with the sub-€50 Raspberry Pi Camera Module 3, so that a camera can be permanently housed within the device, allowing easy use and experimentation (on small items).

Working off of the template developed for the small teaching dome, the same controller and setup can be adapted for a larger, higher-quality dome suitable for research-quality imaging of objects larger than the coins which are the target of the teaching dome. This dome, despite being higher quality and larger will still be much less expensive than commercial alternatives (target: €500) and able to be assembled with basic craft skills, as it will similarly be solderless.

Our poster will show the steps involved in the design and construction of the dome, as well as the necessary coding snippets to successfully run the application. Our aim is to promote our cost-effective workflow, which can be easily replicated by anyone interested, making RTI easier to integrate into teaching and heritage research. We hope, by presenting this project, the prototypes, and the GitHub documenting the project and providing the relevant instructions for procuring parts and assembly, we can make this interesting and innovative heritage imaging technology available to a wider community of scholars.

Bibliographie

Allen, Richard Benjamin. 2023. *RTIPy*. Accessed July 19, 2023, https://github.com/BeebBenjamin/RTIPy .

Corregidor, et al. (2020) “Arduino -controlled Reflectance Transformation Imaging to the study of cultural heritage objects”. SN Applied Sciences 2:1586 | https://doi.org/10.1007/s42452-020-03343-4

Cultural Heritage Imaging. 2002-2023. „Reflectance Transformation Imaging (RTI)“. *Cultural Heritage Imaging* (blog). Accessed July 19, 2023, https://culturalheritageimaging.org/Technologies/RTI/ .

Kinsman, Ted. 2016. „An Easy To Build Reflectance Transformation Imaging (RTI) System“. *JBC* 40, No. 1: 10–14. Accessed July 17, 2023, https://s3.cad.rit.edu/cadgallery_production/storage/media/uploads/projects/1533/documents/275/kinsman-rti-6625-47322-1-pb.pdf .

Österreichisches Archäologisches Institut. 2021. *RTI-Dome - Beleuchtungskuppel & Software*. YouTube. Accessed July 19, 2023, https://www.youtube.com/watch?v=lpPBOkwjqkI.

Pawlowicz, Leszek. 2017. „Affordable Reflectance Transformation Imaging Dome“. Affordable Reflectance

Transformation Imaging Dome. Accessed July 19, 2023, https://hackaday.io/project/11951/instructions .

Winslow, Sean. *SimpleRTIDome*, 2023. Accessed July 19, 2023, https://github.com/larkvi/SimpleRTIDome .

Zaman, Tim. 2015. „Reflectance Transformation Imaging (RTI) Dome“. TimZaman.com. Accessed July 19, 2023, http://www.timzaman.nl/rti-dome .

Annotieren, Visualisieren, Explorieren – ein integrativer Ansatz zur Erschließung von Lyrik in Text und Rezitation

Ketschik, Nora
nora.ketschik@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland

Schauffler, Nadja
nadja.schauffler@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland

Blessing, André
andre.blessing@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland
ORCID: 0000-0001-7573-578X

Gärtner, Markus
markus.gaertner@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland
ORCID: 0000-0002-9548-8461

Rheinwald, Florin
rheinwfn@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland

Bernhart, Toni
toni.bernhart@ilw.uni-stuttgart.de
Uni Stuttgart, Deutschland

Kinder, Anna
anna.kinder@dla-marbach.de
Deutsches Literatur Archiv Marbach, Deutschland

Koch, Julia
julia.koch@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland

Richter, Sandra
sandra.richter@dla-marbach.de
Uni Stuttgart, Deutschland; Deutsches Literatur Archiv Marbach, Deutschland

Sturm, Rebecca
rebecca.sturm@dla-marbach.de
Deutsches Literatur Archiv Marbach, Deutschland

Viehhauser, Gabriel
gabriel.viehhauser-mery@ilw.uni-stuttgart.de
Uni Stuttgart, Deutschland

Vu, Thang
thang.vu@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland

Kuhn, Jonas
jonas.kuhn@ims.uni-stuttgart.de
Uni Stuttgart, Deutschland

Einleitung

Die literaturwissenschaftliche Forschung hat in den letzten Jahren verstärktes Interesse an akustischen Äußerungen von Literatur gezeigt (Binczek und Wirth, 2020; Lehnert et al., 2022; Richter et al., 2023). Dennoch befindet sich die synergetische Betrachtung von Text und Ton in den Digitalen Geisteswissenschaften noch in den Anfängen. Bisher fand Literatur vor allem in ihrer schriftlichen Form Beachtung, während die auditive Realisierung vernachlässigt wurde. Das Projekt »textklang« der Universität Stuttgart und des Deutschen Literaturarchivs (DLA) Marbach widmet sich diesem Forschungsdesiderat, indem es systematisch die Beziehung zwischen literarischen Texten, insbesondere Lyrik, und ihrer lautsprachlichen Realisierung untersucht. Dabei werden digitale phonetische und sprachtechnologische Verfahren, computerlinguistische Ansätze sowie Explorations- und Visualisierungstechniken mit hermeneutischen Methoden der literaturwissenschaftlichen Textanalyse kombiniert. Die Zusammenführung von Text- und Klangebene verlangt nach speziellen Workflows und Tools. In diesem Beitrag stellen wir zwei Anwendungen zur

Datenexploration vor, die es erlauben, das multimodale Lyrikkorpus auf verschiedenen Ebenen zu erforschen.

Der Daten-Workflow in »textklang« (Abb. 1) reicht vom Sammeln der heterogenen Daten über deren Annotation auf Text- und Audioebene zur Exploration und Nachnutzung. Der Workflow zeigt, dass die Explorationstools auf unterschiedlichen Datenquellen basieren: Das MD-Dashboard visualisiert die Metadaten, der Textklang-Explorer ergänzt diese um Text- und Audio-Daten, die zuvor umfänglich annotiert wurden. Die Anwendungen stehen an der Schnittstelle zur (Nach)Nutzung, indem sie die Entwicklung und Überprüfung konkreter Forschungsfragen befördern.

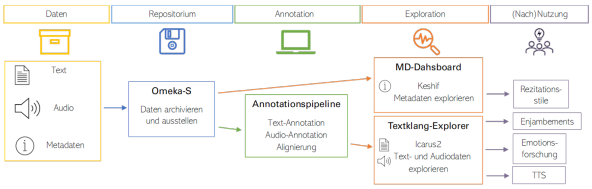


Abb. 1: Daten-Workflow im Projekt »textklang«.

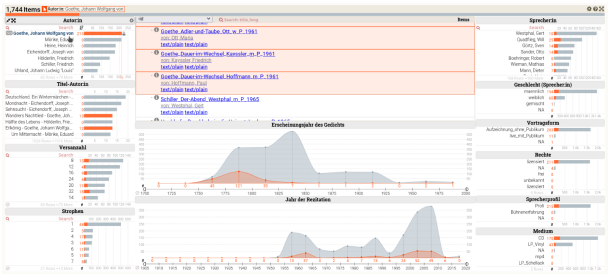


Abb. 2: MD-Dashboard1 mit Markierung aller Goethe-Gedichte.

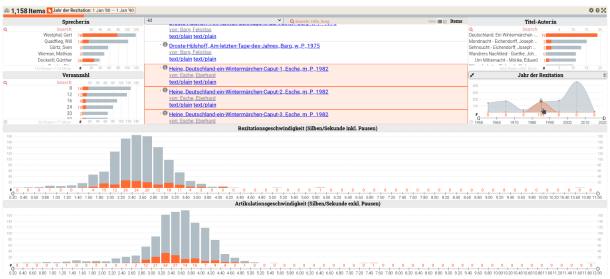


Abb.3: MD-Dashboard2 zur Sprechgeschwindigkeitsanalyse.

Korpus

Das Audio-Korpus speist sich aus den umfangreichen Beständen des DLA und umfasst derzeit 1744 Rezitationen von gut 1000 Gedichten, hauptsächlich aus der Zeit der Romantik. Metadaten, wie Verfasser:in und Entstehungsjahr der Gedichte, Sprecher:in und Aufnahmejahr der Rezitation, werden zusammen mit einer Transkription und der Audioaufnahme im Online-Repositorium Omeka-S (Omeka, 2022) abgelegt. Omeka-S ist eine open-source Plattform für die Erstellung, Nutzung und Ausstellung multimodaler digitaler Sammlungen. Die Text- und Audiodateien werden anschließend exportiert und durchlaufen umfangreiche automatische Annotationen (basierend auf der Pipeline des GRAIN-Korpus, Schweitzer et al., 2018), wodurch morphosyntaktische Informationen wie Wortarten, Lemmata und Syntax, sowie phonetische Informationen wie Intonationsereignisse und Silbendauern erfasst werden (vgl. Schauffler et al., 2022a). Zudem werden formale Aspekte wie Anfang und Ende von Versen und Strophen annotiert.

Exploration

MD-Dashboard. Wir verwenden das webbasierte Visualisierungstool Keshif (Yalçın et al., 2016) für die Exploration der Metadaten. Diese können je nach Forschungsinteresse interaktiv ausgewählt und kombiniert werden (Abb. 2 und Abb. 3). Die aus Omeka-S importierten Metadaten ergänzen wir um berechnete Text- und Audioattribute wie Versanzahl und Rezitationsgeschwindigkeit.

Durch die Kombination verschiedener Metadatenattribute können Zusammenhänge und Verteilungen im Korpus sichtbar gemacht werden. So können wir zum Beispiel herausfinden, welche Autoren zu welchen Zeiten und/oder von welchen Sprecher:innen rezitiert wurden. Auch diachrone Zusammenhänge, zum Beispiel zwischen Sprechgeschwindigkeit und Aufnahmejahr, lassen sich erfassen. Das MD-Dashboard ermöglicht den direkten Zugriff auf einzelne Einträge in Omeka-S und das Abspielen von (nicht-lizensierten) Aufnahmen.

Textklang-Explorer. Der Textklang-Explorer ermöglicht die Abfrage verschiedener Annotationen auf Text- und Audioebene in Verbindung mit den Metadaten. Es handelt sich um eine benutzerfreundliche Webanwendung, die auf ICARUS2 (Gärtner, 2018, 2020) als Anfrageschnittstelle basiert. Dadurch können komplexe Suchanfragen an das Korpus gestellt werden, beispielsweise zur Morphosyntax (POS-Tags, Lemmata), zur formalen Struktur der Gedichte (stehen bestimmte Phänomene am Anfang oder Ende von Versen/Strophen) oder zur prosodischen Realisierung bestimmter Abschnitte. Eine Kombination der Suchanfragen ermöglicht es, die verschiedenen Annotationsebenen gezielt zusammenzuführen (Abb. 4).

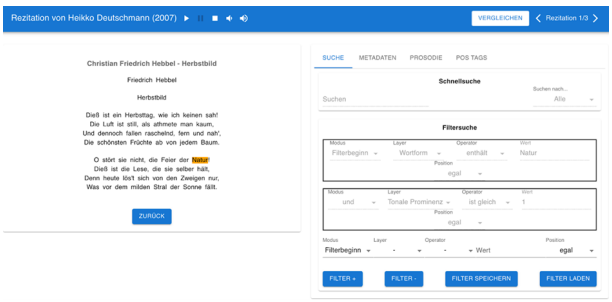


Abb. 4: Suchanfrage im Textklang-Explorer3, bei der alle tonal prominenten Versionen des Wortes “Natur” gesucht werden (rechts), und eines der Sucher-gebnisse mit markiertem Token (links).

Darüber hinaus besteht die Möglichkeit, die Primärdaten – d.h. die Transkripte der Gedichte sowie die Rezitationen – einzusehen und (sogar wortweise) abzuspielen, sodass makroanalytische Queries mit mikroanalytischen Untersuchungen kombiniert werden können. Auch der Vergleich verschiedener Rezitationen von Gedichten wird unterstützt, zum Beispiel durch die Visualisierung prosodischer Unterschiede. Zudem wird eine Schnittstelle für den Download der Daten integriert, die es ermöglicht, die Audiodateien etwa im Kontext sprachtechnologischer Ansätze (TTS) weiterzunutzen (Koch et al., 2022).

Die vorgestellten Ressourcen und Tools begegnen der Herausforderung, multimodale Daten zu erforschen, und eröffnen durch die Kombination der verschiedenen Informationsebenen neue Forschungsfragen. Ein Beispiel ist die Erforschung des Zusammenspiels zwischen Aspekten der äußeren Gedichtform und der sprachlichen Realisierung in einer Studie zur Stilfigur des Enjambements (Schauffler et al., 2022b; Schauffler et al., 2023). Auch Fragen zu Merkmalen von Rezitationsstilen und zur Identifikation von Sprechweisen (z.B. metrisch, prosaisch) können adressiert werden. In unserem Poster stellen wir die genannten Tools vor und diskutieren anhand von Anwendungsfällen die Herausforderungen und Chancen der integrativen Analyse von Text und Ton.

Fußnoten

1. https://clarin03.ims.uni-stuttgart.de/md_dashboard/.
2. https://clarin03.ims.uni-stuttgart.de/md_dashboard/speech.html .
3. https://96e90d20-0328-4a38-87e5-93bc6af-a5877.ma.bw-cloud-instance.org/textklang.

Bibliographie

Binczek , Natalie und Uwe Wirth (Hrsg.). 2020. Handbuch Literatur & Audiokultur, Berlin, Boston: de Gruyter.

Gärtner , Markus. 2018. “ICARUS2: 2nd generation of the Interactive platform for Corpus Analysis and Research tools.” University of Stuttgart. Distributed via: http://hdl.handle.net/11022/1007-0000- 0007-C635-E.

Markus Gärtner. 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.

Koch, Julia, Florian Lux, Nadja Schauffler, Toni Bernhart, Felix Dieterle, Jonas Kuhn, Sandra Richter, Gabriel Viehhauser , and Ngoc Thang Vu. 2022. “PoeticTTS – Controllable Poetry Reading for Literary Studies.” In Proceedings of Interspeech 2022. DOI: 10.48550/arXiv.2207.05549.

Lehnert, Nils, Ina Schenker und Andreas Wicke (Hrsg.). 2022. Gehörte Geschichten. Phänomene des Auditiven, de Gruyter https://doi.org/10.1515/9783110741773

Omeka. 2022. “Omeka S.” Omeka. Distributed via GitHub: https://github.com/omeka/omeka-s, 3.2.

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, Andreas Kozlik. 2023. “Der Klang der Lyrik. Zur Konzeptualisierung von Sprecher und Stimme, auch für die computationelle Analyse.” In Poema. Jahrbuch für Lyrikforschung / Annual for the Study of Lyric Poetry / La recherche annuelle en poésie lyrique 1 (2023), 39–51. https://doi.org/10.38072/2751-9821/p4

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. 2022a. “»textklang« - Towards a Multi-Modal Exploration Platform for German Poetry.” In Proceedings of the 13th edition of the Language Resources and Evaluation Conference (LREC 22), 5345–5355.

Schauffler, Nadja, Fabian Schubö , Toni Bernhart, Gunilla Eschenbach, Julia Koch, Sandra Richter, Gabriel Viehhauser , Thang Vu, Lorenz Wesemann und Jonas Kuhn. 2022b. “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, Julia Koch, Nora Ketschik , Toni Bernhart , Felix Dieterle , Gunilla Eschenbach, Anna Kinder, Sandra Richter, Gabriel Viehhauser , Thang Vu, and Jonas Kuhn. 2023. “Final lengthening in line end perception of re-synthesized recitations of German poems.” In Proceedings of the 20th International Congress of Phonetic Sciences, [in Vorbereitung].

Schweitzer, Katrin, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska , Arndt Riester, Ina Rösiger , Antje Schweitzer, Sabrina Stehwien , and Jonas Kuhn. 2018. “German radio interviews: The GRAIN release of the SFB732 silver standard collection.” In Proceedings of the

Eleventh International Conference on Language Resources and Evaluation (LREC 18), 2887–2895.

Yalçın , Mehmet Adil, Niklas Elmqvist , and Benjamin B. Bederson . 2016. “Keshif: Out-of-the-box visual and interactive data exploration environment.” In Proceedings of IEEE VIS 2016 Workshop on Visualization in Practice: Open Source Visualization and Visual Analytics Software.

Ansätze und Tools für Historische Text Reuse Detection Fragmentierter Text Reuse am Beispiel ripuarischer Inkunabeln des 15. Jahrhunderts

Ostrowski, Alina
alina.ostrowski@uni-passau.de
Universität Passau, Deutschland; Universität zu Köln, Deutschland
ORCID: 0000-0002-0910-8566

Text Reuse Detection (TRD) ist ein Teilgebiet des Natural Language Processing (NLP) mit dem Ziel, wiederverwendete Passagen (Text Reuse, TR) in distinkten Texten zu identifizieren. Seit etwa zehn Jahren wird TRD vermehrt für historische Forschung erprobt und eingesetzt, wobei typische Charakteristika historischer Sprachdaten (z.B. geringe sprachliche Standardisierung, fehlende digitale Ressourcen) die Durchführung erschweren. Der vorliegende Beitrag gibt einen Überblick über Ansätze und Tools zur Historischen TRD (HTRD). Anhand eines konkreten Forschungsvorhabens wird die Anwendbarkeit dreier HTRD-Tools (*Passim*, *BLAST* und *TextPAIR*) auf kurzen TR in mittelalterlichen, volkssprachigen Texten diskutiert.

Oft zitiert ist *Tracer* (Büchler, 2013; vgl. Hiltmann et al., 2021), eine Software-Suite, die für einzelne TRD-Schritte unterschiedliche Algorithmen zur Verfügung stellt und mit Fingerprinting arbeitet.[1] *Passim* (Smith et al., 2015) und *TextPAIR* (Gladstone, 2018) finden Dokumente mit TR auf Basis der Überlappung ihrer Zeichen- bzw. Wort-n-Gramme. Die Dokumentenpaare werden anschließend zwecks Eingrenzung der korrespondierenden Passagen aligniert. Auch das R-Paket *textreuse* (Mullen, 2020) folgt diesem Schema, jedoch werden zur Berechnung der Dokument-Ähnlichkeit Min-Hashes und der Jaccard-Koeffizient zweier Texte benutzt. Einen sprachagnostischen Ansatz wählten Vesanto et al. (2017) und „codierten“ die natürlichsprachlichen Texte als Aminosäuren, die sie dann mit dem aus der Bioinformatik stammenden *BLAST* (Basic Local Alignment Search Tool) auf ähnliche Sequenzen hin untersuchten. *Tesserae* wurde speziell für die Erforschung von Allusionen in lateinischer Lyrik entwickelt (Coffee et al., 2013) und verwendet besonders weiche Ähnlichkeitskriterien sowie eine Scoring-Funktion zur nachträglichen Ergebnisfilterung. HTRD mit stochastischen Sprachmodellen wurde bisher seltener versucht (vgl. Liebl/Burghardt, 2022).

Die meisten HTRD-Methoden wurden für große Textmengen in klassischen und neuzeitlichen Sprachen mit moderater orthographischer Varianz entwickelt (z.B. Smith et al., 2015; Gladstone/Cooney, 2020), in denen oft längerer TR (>1 Satz) vorlag. Für viele historische Anwendungsfälle sind diese Prämissen jedoch nicht gegeben. Die Eignung der vorhandenen Ansätze für abweichende Szenarien soll darum am Beispiel zweier ripuarischer Inkunabeln (#500.000 Tokens) untersucht werden. Es handelt sich um Werke aus dem Kontext der Kölner Stadtgeschichtsschreibung, die nachweislich Parallelstellen enthalten (Anonym, 1490[2]; Anonym, 1499[3]; vgl. Meier, 1998, S. 78f.). Auf die Texte wurde mit *Transkribus* Handwritten Text Recognition (HTR) angewandt, die stellenweise eine erhöhte Fehlerrate aufweist. Zudem liegt ein hoher Grad an intra- und intertextueller orthographischer Varianz vor. Die bereits bekannten Fälle von TR im Textkorpus legen nahe, dass dieser eher fragmentiert, d.h. kurz, nicht-wörtlich und teils in Syntax, Reihenfolge oder verwendeten Lexemen abgeändert ist, was die Anwendung von TRD allgemein erschwert (vgl. Moritz et al., 2016). Hinzukommt, dass für die ripuarische Sprache keine NLP-Ressourcen, wie trainierte Lemmatisierer oder annotierte Korpora[4] vorliegen.

Erste Tests zeigten, dass insbesondere die sprachliche Varianz der Texte sowie die Kürze des TR Probleme für die HTRD darstellen: *TextPAIR* und *BLAST* fanden mit Standardeinstellungen zwar zahlreiche, aber überwiegend triviale, *Passim* gar keine textuellen Ähnlichkeiten. Im Folgenden werden in stark verkürzter Form die Durchführung und die Ergebnisse eines systematischen Vergleichs der drei Programme vorgestellt.

Zunächst wurden für ein Evaluationsset[5] 19 wörtliche bis schwach-wörtliche TR-Fälle in einem Auszug aus dem Untersuchungskorpus manuell annotiert. Außerdem wurden zwei naive Maßnahmen zur sprachlichen Vereinheitlichung umgesetzt: Erstens, regelbasierte Orthographie-Normalisierung und zweitens, „Pseudo-Lemmatisierung” durch das Clustern von Wörtern mit einem hohen Alignment-Wert bei Verwendung des Needleman-Wunsch-Algorithmus. Mit jeder Kombination dieser Präprozessierungsarten wurde ein Evaluationsset aus den annotierten Textteilen erstellt, die wiederum die Eingabedaten für Testdurchläufe jedes HTRD-Programms mit diversen Kombinationen aus Parameterwerten bildeten. Für diese wurde anschließend der F1-Score anhand eines Abgleichs zwischen den Ergebnistreffern und den 19 TR-Fällen berechnet.

Tab. 1: Beste Testdurchläufe (nach F1, Evaluationswerte gerundet)

Tool	*F1*	*Rec.*	*Prec.*	*Treffer*	*davon Goldtreffer*	*Input*	*Normal.*	*Lemmat.*	*Länge der n-Gramme*
BLAST	0,31	0,26	0,38	16	5	Volltexte	Nein	Nein	7 Zeichen
Passim	0,42	0,26	1,00	5	5	Volltexte	Nein	Ja	9 Zeichen
TextPAIR	0,26	0,16	0,75	4	3	Volltexte	Nein	Ja	3 Tokens

Tab. 2: Beste Testdurchläufe (absoluter Recall, Evaluationswerte gerundet)

Tool	*F1*	*Rec.*	*Prec.*	*Treffer*	*davon Goldtreffer*	*Input*	*Normal.*	*Lemmat.*	*Länge der n-Gramme*
BLAST	0,14	0,47	0,08	123	9	Volltexte	Nein	Ja	5 Zeichen
Passim	0,15	0,79	0,08	119	15	Sätze	Nein	Ja	5 Zeichen
TextPAIR	0,22	0,32	0,17	41	6	Volltexte	Nein	Ja	3 Tokens

Passim erreichte die besten Evaluationswerte, gefolgt vom ebenfalls mit Zeichen-n-Grammen arbeitenden *BLAST* (Tab. 1)[6]. Für *TextPAIR* und *Passim* stellte sich die Anwendung der Pseudo-Lemmatisierung als erfolgreich heraus, wohingegen das sprachagnostische *BLAST* mit nicht-vorverarbeiteten Texten erfolgreicher war. Insgesamt ist der F1-Score aller Programme trotz optimierter Parameterwerte niedrig (<0,5) und die Precision bei erhöhtem Recall sehr gering (Tab. 2). In Gesamtdurchläufen mit diesen optimierten Parameterwerten wurden im Korpus zwar zuvor unbekannte Parallelstellen gefunden, doch wegen der Kürze der n-Gramme waren über 70% aller Ergebnisse triviale Ähnlichkeiten (Namen, Mehrwort-Ausdrücke, Phrasen).

Es lässt sich festhalten, dass die getesteten, auf n-Gramm-Vergleichen basierenden HTRD-Ansätze zwar in der Lage sind, Parallelstellen in ripuarischen Texten zu erkennen, und somit einen Mehrwert für die Textanalyse bieten, doch dass für die zuverlässige Erkennung von komplexem TR in vormodernen, deutschsprachigen Texten weitere Forschung nötig ist oder gänzlich andere Ansätze verwendet werden müssen.

Fußnoten

1. *Tracer* war trotz angemessener Versuche nicht (mehr) zugänglich.
2. Veröffentlichung in Vorbereitung.
3. Volltext bereitgestellt vom Projekt „Koelhoffsche Chronik 1499 digital“ (https://www.uni-muenster.de/Geschichte/histsem/LG-G/Forschen/koelhoffschechronik.html; Bruch, 2023).
4. Sprachlich entfernt verwandt aber mit geringer Type-Überlappung: ReN-Team, 2021.
5. Veröffentlichung in Vorbereitung.
6. Die n-Gramm-Länge steht stellvertretend für zahlreiche untersuchte Programm-Parameter. 30% aller *TextPAIR*-Durchläufe konnte wegen eines internen Programmfehlers nicht berücksichtigt werden.

Bibliographie

Anonymus. 1490. *Der Doernenkrantz van Collen* . Köln: Johann Koelhoff d. J. Digitalisat: https://tudigit.ulb.tu-darmstadt.de/show/inc-ii-674 (zugegriffen: 5.12.2023).

Anonymus. 1499. *Die Cronica van der hilliger Stat van Coellen*. Köln: Johann Koelhoff d. J. Digitalisat: https://sammlungen.ulb.uni-muenster.de/hd/content/titleinfo/7159780 (zugegriffen: 5.12.2023).

[*BLAST*:] **Vesanto, Aleksi et al.** *Text Reuse Detection with BLAST*. Ohne Version (Stand 2019). URL: https://github.com/avjves/textreuse-blast.

Bruch, Julia. 2023. „Mit Studierenden edieren: Digitale Editionen als Chance für die Lehre.“ *DigiTRiP*. https://digitrip.hypotheses.org/1278 (zugegriffen: 19.07.2023).

Büchler, Marco. 2013. *Informationstechnische Aspekte des Historical Text Re-use*. PhD diss., Universität Leipzig. urn:nbn:de:bsz:15-qucosa-108515.

Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde und Sarah L. Jacobson. 2013. „The Tesserae Project: intertextual analysis of Latin poetry.“ *Literary and Linguistic Computing* 28 (2): 221–28. 10.1093/llc/fqs033.

Franzini, Greta, Marco Passarotti, Maria Moritz und Marco Büchler. 2018. „Using and Evaluating TRACER for an Index Fontium computatus of the Summa Contra Gentiles of Thomas Aquinas.“ In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-It 2018*, hrsg. von E. Cabrio, A. Mazzei und F. Tamburini, 1–11. https://books.openedition.org/aaccademia/3369 (zugegriffen: 19.07.2023).

Gladstone, Clovis. 2018. „TextPAIR: a new high-performance sequence aligner.“ *ARTFL Project Research Blog*. https://artfl.blogspot.com/2018/12/textpair-new-high-performance-sequence.html (zugegriffen: 19.07.2023).

Gladstone, Clovis und Charles Cooney. 2020. „Opening new paths for scholarship: Algorithms to track text reuse in Eighteenth Century Collections Online.“ In *Digitizing Enlightenment: Digital Humanities and the Transformation of Eighteenth-Century Studies*, hrsg. von S. Burrows und G. Roe, S. 353–374. Oxford University Studies in the Enlightenment 2020:07. Liverpool.

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* 47 (1): 122–56. 10.13109/gege.2021.47.1.122.

Liebl, Bernhard und Manuel Burghardt. 2022. „The Vectorian API: A Research Framework for Semantic Textual Similarity (STS) Searches.“ In *Digital Humanities 2022: Conference Abstracts*, hrsg. von Yifan Wang, 654–56. Tokio.

Meier, Robert. 1998. *Heinrich van Beeck und seine „Agrippina“: Ein Beitrag zur Kölner Chronistik des*

15. Jahrhunderts. Mit einer Textdokumentation. Kölner historische Abhandlungen 41. Köln.

Moritz, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni und Marco Büchler. 2016. „Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse.“ In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1849–59. 10.18653/v1/D16-1190.

Mullen, Lincoln. 2020. *textreuse: Detect Text Reuse and Document Similarity*. https://docs.ropensci.org/textreuse, https://github.com/ropensci/textreuse.

[*Passim*:] **Smith, David A. et al.** *passim*. Version 2.0.0 alpha 2 (Stand 2022). URL: https://github.com/dasmiq/passim/releases/tag/v2.0.0-alpha.2.

ReN-Team. 2021. „Reference Corpus Middle Low German/Low Rhenish (1200–1650); Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)“ (Version 1.1) [Data set]. 10.25592/uhhfdm.9195.

Smith, David A., Ryan Cordell und Abby Mullen. 2015. „Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers.“ *American Literary History* 27 (3): E1-E15.

[*TextPAIR*:] **ARTFL-Project.** *TextPAIR (Pairwise Alignment for Intertextual Relations)*. Version 2.1 (Stand Aug. 2023). URL: https://github.com/ARTFL-Project/text-pair/releases/tag/v2.1.0.1.

[*Transkribus*:] **Kahle, Philip, Sebastian Colutto, Günter Hackl und Günter Mühlberger.** 2017. „Transkribus - a Service Platform for Transcription, Recognition and Retrieval of Historical Documents.“ In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 19–24. 10.1109/ICDAR.2017.307.

Vesanto, Aleksi, Filip Ginter, Hannu Salmi, Heli Rantala, Asko Nivala und Tapio Salakoski. 2017. „A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora.“ In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 330–33. https://aclanthology.org/W17-0249 (zugegriffen: 19.07.2023).

#arthistoCast – der Podcast zur Digitalen Kunstgeschichte Wissenschaft auf die Ohren

Klusik-Eckert, Jacqueline
jklusik@gmail.com
HHU Düsseldorf, Deutschland
ORCID: 0000-0002-0969-2520

Im Wissenschaftspodcast #arthistoCast dreht sich alles um die Digitale Kunstgeschichte. Dabei geht es um den Einsatz digitaler Methoden in der kunsthistorischen Forschung, also um die Frage, wie technische Entwicklungen für das Fach genutzt werden können und wie sich die Forschung im Zuge der Digitalisierung verändert hat.

Wissenschaft auf die Ohren. Wo sind die Geisteswissenschaften?

Podcasts liegen im Trend, auch wenn sie an sich nichts neues sind. Das allgemeine Interesse an Podcasts als Informationsquelle für wissenschaftliche Themen in der Bevölkerung ist in den letzten Jahren deutlich gestiegen (2018: 6 %, 2021: 16 %) („WiD -Wissenschaftsbarometer 2021“, 2021). Als Form des Audiobloggings sind Wissenschaftsformate aus den Podcasts-Charts nicht mehr wegzudenken. Sie sind als Medienformate für die Wissenschaftskommunikation etabliert, gleichwohl es keine genaue Definition dieser Kategorie gibt oder eine einschlägige Erforschung des Impacts dieser Formate für die Wissenschaftskommunikation erforscht ist (Leander, 2020, 1–2). Doch wie sieht es mit geisteswissenschaftlichen Themen aus? Während man bei naturwissenschaftlichen Themen aus einer Fülle an Sendungen wählen kann, muss man für geisteswissenschaftliche Formate auf die Suche gehen.[1] Auch wenn in den populären Podcasts der Rundfunkanstalten, die sich allgemein mit wissenschaftlichen Themen auseinandersetzten, geisteswissenschaftliche Forschungsfragen angesprochen werden, dominieren die MINT-Fächer das Feld der Wissenschaftspodcast (Moltmann, 2020, 3; Dernbach, 2022).[2]

Wer sind die Akteure?

Die Kategorie „Wissenschaft“ im Podcastuniversum wird maßgeblich von den Wissenschaftsredaktionen der öffentlichen Rundfunks- und Fernsehanstalten dominiert. Wissenschaftspodcasts, die als Medium der Wissenschaftskommunikation direkt von Forscher*innen produziert und moderiert werden, sind nicht nur rar, sie verfügen auch nicht über die gleiche Sichtbarkeit wie professionell produzierte Formate.[3] In seiner Studie über das Podcastverhalten direkt aus der heraus Wissenschaft stellt Lewis MacKenzie fest, dass dieses Engagement „as an extra commitment beyond their regular duties as a scientific researcher, science educator or science communicator“ ist (2019, 15). Diese Beobachtung stützt auch der Wert einer Umfrage 2021. Hier haben lediglich fünf Prozent oder weniger der befragten Wissenschaftler*innen angegeben, einen Podcast für ihre Inhalte einzusetzen (Ziegler et al., 2021).

Warum noch einen Podcast?

Die Gründe für den Einsatz von Podcasts als Medium der Wissenschaftskommunikation sind vielseitig und eindrücklich (Dernbach, 2022, 308–9). Gleichwohl das Format sicherlich nicht zu den innovativsten zählt.

Neben der Popularisierung des Podcasthörens an sich ist da zunächst die Zugänglichkeit zu nennen. Veröffentlich über ein RSS-Format können die Sendungen in die unterschiedlichen Distributoren (Spotify, Apple Podcast, Overcast, etc.) bereitgestellt werden. Die Hörer*innen können gemäß eigener Hörgewohnheiten die Plattform der Wahl verwenden. Auch der Zeitpunkt oder der Ort ist deutlich flexibler als Video- oder Textformate.

Darüber hinaus lässt sich in dem Format das Spagat zwischen wissenschaftlich-anspruchsvollen auch teils komplexen Inhalten und unterhaltsamer Aufbereitung eben dieser gut umsetzten (Lee, McLoughlin, und Chan, 2008). Daher hat sich in einige Wissenschaftsformaten das sokratische Gespräch mit eindeutiger Rollenverteilung meist etabliert, wie man es auch an den Digital Humanities Podcast RaDiHum20 und Coding Codices feststellen kann.[4]

Motivation für einen Wissenschaftspodcast zur Digitalen Kunstgeschichte

Auf informativer und unterhaltsamer Weise soll im #arthistoCast – der Podcast zur Digitalen Kunstgeschichte die Hemmschwelle für Kunsthistoriker*innen abgebaut werden, die sich bislang kaum mit digitalen Methoden oder Verfahren auseinandergesetzt haben. Darüber hinaus wird auch für Digital Literacy sensibilisiert. Neben den gängigen Forschungsthemen gibt es ausreichend Raum für Theorie, Methoden und die praktische Umsetzung dieser geben. Neben Diskussionen zu aktuellen Debatten der Digitalen Kunstgeschichte werden Institutionen vorgestellt werden, die durch ihre Aktivität einen Mehrwehrt für die Digitale Kunstgeschichte leisten. Dadurch soll der Brückenschlag zwischen der Hochschule und den ebenfalls forschenden GLAM-Institutionen geleistet werden.[5]

Im Zentrum stehen dabei Diskussionen und damit die Vorstellung unterschiedlicher Perspektiven innerhalb einzelner Themenbereiche, die meist noch in Aushandlungsphasen sind. Dadurch kann ein Transfer von Ergebnissen aus einem Spezialist*innenkreis für die breite kunsthistorische Forschung gelingen und ermächtig die Hörer*innen zur Teilhabe an den aktuellen Debatten.

Dies gelingt, in dem in jeder Folge Expert*innen aus unterschiedlichen Fach- und Themenbereichen zu Gast sind, um über ihre Arbeit und ihre Erfahrungen mit digitalen Methoden und Technologien zu sprechen. Dabei geht es nicht nur um gute Lösungsansätze und etablierte Systeme, sondern auch um aktuelle Herausforderungen und Möglichkeiten, die mit der Anwendung digitaler Methoden in der Kunstgeschichte einhergehen.

Jacqueline Klusik-Eckert führt die Zuhörer*innen dabei durch den Begriffsdschungel der Technikwelt und hilft, jeder und jedem einen Einstieg in die Themenfelder zu finden.

Wer ist verantwortlich?

Der Podcast wird von Jacqueline Klusik-Eckert im Auftrag des Arbeitskreises Digitale Kunstgeschichte produziert. Unterstützt wird sie dabei von der Redaktion bestehend aus Mitgliedern des Arbeitskreis Digitale Kunstgeschichte: Peter Bell, Lisa Dieckmann, Peggy Große, Waltraud von Pippich und Holger Simon. Finanziert wird #arthistoCast – der Podcast zur Digitalen Kunstgeschichte von NFDI4Culture, also dem Konsortium in der Nationalen Forschungsdateninfrastruktur (NFDI). Weiterhin wird der Podcast vom Deutschen Verband für Kunstgeschichte unterstützt.

Wie nachhaltig kann ein Podcast sein?

Gemäß einer guten wissenschaftlichen Praxis wurde bei #arthistoCast darauf geachtet, dass die Folgen nicht nur über die gängigen Podcast-Distributoren verfügbar sind. Zu einer nachhaltigen Verwendung der Inhalte und einer Zitierfähigkeit trägt ein eigens erdachtet Hosting-Konzept bei. Die Folgen werden im Repositorium heidICON mit Metadaten langzeitarchiviert und DOI referenziert. Weiterhin werden die Hörer*innen im begleitenden Block des Fachinformationsdienst für Kunstgeschichte arthistoricum.net zu Diskussionen und Beiträgen aufgefordert.

Links

Homepage des Podcasts: https://www.arthistoricum.net/themen/podcasts/arthistocast
RRS Feed: https://feeds.zencastr.com/f/G1QaI9Vc.rss
HeidICON: https://doi.org/10.11588/heidicon/1738702

Fußnoten

1. Die Recherche über die gängigen Chartportale und Metasammlungen gestaltet sich als schwierig, da es selten ein Label „Geisteswissenschaften“ gibt. In den Charts der Kategorie „Wissenschaft“ findet sich kaum ein dezidiert geisteswissenschaftler Podcasts, vgl. https://podwatch.io/charts/wissenschafts-podcasts/ (Stand 21.11.2023, Veränderungen sind tagesaktuell). Die Plattform www.podcasts.de listet unter der Kategorie „Wissenschaft“ über 3000 Formate im deutschsprachigen Raum auf, wobei auch hier die Natur- und Humanwissenschaf-

ten dominieren, vgl. https://www.podcast.de/beste-podcasts/wissenschaft-16 (Stand 21.11.2023).
2. Vgl. https://wissenschaftspodcasts.de/podcasts/?cat=geisteswissenschaft (Stand 19.07.2023)
3. Die Plattform wissenschaftspodcasts.de möchte dem Abhilfe verschaffen. Mit einem ehrenamtlich engagierten Team wird hier versucht die Bandbreite der Podcasts für den deutschen Sprachraum abzudecken, vgl. https://wissenschaftspodcasts.de/ (Stand 19.07.2023)
4. Vgl. https://radihum20.de/
5. Das fachliche Tun steht hierbei über dem Versuch einer Definition der Digitalen Kunstgeschichte (Kaufmann, Satilmis, und Mieg, 2019, 4)

Bibliographie

Dernbach, Beatrice. 2022. „Hineinhören in die wunderbare Welt der Wissenschaft. Podcasts als Medium der Wissenschaftskommunikation“. In *Podcasts* , herausgegeben von Vera Katzenberger, Jana Keil, und Michael Wild, 307–32. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-38712-9_12.

Kaufmann, Margrit E., Ayla Satilmis, und Harald A. Mieg. 2019. „Einleitung: Forschendes Lernen in den Geisteswissenschaften: Ansätze, Impulse und Herausforderungen“. In *Forschendes Lernen in den Geisteswissenschaften* , herausgegeben von Margrit E. Kaufmann, Ayla Satilmis, und Harald A. Mieg, 1–18. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21738-9_1.

Leander, Lisa. 2020. „Wissenschaft im Gespräch: Wissensvermittlung und -aushandlung in Podcasts“. *kommunikation@gesellschaft* 21 (2): 1–24. https://doi.org/10.15460/kommges.2020.21.2.621.

Lee, Mark J. W., Catherine McLoughlin, und Anthony Chan. 2008. „Talk the Talk: Learner-generated Podcasts as Catalysts for Knowledge Creation“. *British Journal of Educational Technology* 39 (3): 501–21. https://doi.org/10.1111/j.1467-8535.2007.00746.x.

MacKenzie, Lewis E. 2019. „Science Podcasts: Analysis of Global Production and Output from 2004 to 2018“. *Royal Society Open Science* 6 (1): 180932. https://doi.org/10.1098/rsos.180932.

Moltmann, Rebecca. 2020. „Vom ‚Verfertigen der Gedanken‘: Zum Potential von Podcasts für die geisteswissenschaftliche Wissenschaftskommunikation“. *kommunikation@gesellschaft* 21 (2). https://doi.org/10.15460/kommges.2020.21.2.624.

„WiD -Wissenschaftsbarometer 2021“. 2021. Berlin: Wissenschaft im Dialog.

Ziegler, Richarda, Liliann Firscher, Jens Ambrasat, Gregor Fabian, Philipp Niemann, und Cecilia Buz. 2021. „Wissenschaftskommunikation in Deutschland. Ergebnisse einer Befragung unter Wissenschaftler:innen“. Wissenschaft im Dialog, Deutsches Zentrum für Hochschul- und Wissenschaftsforschung, Nationales Institut für Wissenschaftskommunikation. https://www.wb.dzhw.eu/pdf/2021_WisskommBefragung_Ergebnisbroschuere_WiD_DZHW_NaWik.pdf.

Auffinden und Analysieren komplexer Textvarianten in Hannah Arendts Denk- und Schreibwerkstatt mit LERA

Etling, Fabian
fabian.etling@fu-berlin.de
Freie Universität Berlin, Deutschland

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0000-0001-8718-1198

Grote, Brigitte
brigitte.grote@fu-berlin.de
Freie Universität Berlin, Deutschland

Das Editionsvorhaben *Hannah Arendt. Kritische Gesamtausgabe*[1] macht sämtliche Schriften der Autorin mit Ausnahme der Briefe zugänglich. Dafür werden von einem internationalen Herausgeber:innenteam ca. 21.000 Seiten der von Hannah Arendt zu Lebzeiten veröffentlichten Werke (Monographien, Essaysammlungen, Artikel, Interviews etc.) sowie unveröffentlichter Dokumente aus ihrem Nachlass (Typoskripte, Notizen etc.) bearbeitet und mit einem detaillierten Stellenkommentar versehen. Die Website[2] der Edition komplementiert die in den Druckbänden veröffentlichten konstituierten Texte[3], indem sie die zahlreichen Vorfassungen und Umarbeitungen der Autorin integriert. Die Website versteht sich nicht nur als Publikationsort der digitalen Edition, sondern darüber hinaus als erweiterte Forschungsplattform[4]: Die Entstehungsstadien ihrer Schriften sollen, soweit Zeugnisse vorliegen, nachvollzogen werden können. Der Editionsvorgang soll dabei bis hin zu einzelnen Entzifferungs- und Deutungsentscheidungen nachprüfbar bleiben (Etling und Kieslich, 2024). Eine Herausforderung stellen dementsprechend die Kollationierung und textkritische Auseinandersetzung mit den großen Textmengen dar – eine editorische Arbeit, die manuell nur mit erheblichem Aufwand leistbar ist. Den Leser:innen der Edition sollen durch die Wiedergabe eines

vollständigen Textvergleichs die textkritischen Entscheidungen transparent gemacht sowie die Identifikation und Beurteilung varianter Stellen sowohl auf einer Makroebene (Textsegmente) als auch auf einer Mikroebene (Zeichenmengen innerhalb der Segmente) ermöglicht werden.[5]

Mit LERA[6] existiert ein frei verfügbares digitales Kollationierungswerkzeug, welches die Berechnung und Präsentation von Textvarianten zwischen Textfassungen sowie deren Analyse mittels interaktiver Visualisierungen ermöglicht (Pöckelmann et al., 2023). Der Textvergleich erfolgt in zwei Stufen: Auf eine automatische Alignierung von größeren Textsegmenten wie Absätzen folgt ein Detailvergleich auf Tokenebene, sodass auch große Textmengen von der Software verarbeitet und dargestellt werden können. LERA ist für den Vergleich mehrerer Textfassungen ausgelegt, wobei im Gegensatz zum Leitquellenprinzip ein gleichrangiger Vergleich verfolgt wird.[7] Spezifische Textvarianten können durch das integrierte Filtersystem dynamisch vom Textvergleich ausgenommen werden. Neben der Volltextpräsentation via synoptischer Gegenüberstellung und Variantenapparat zur Analyse von Einzelstellen (Close Reading) bietet LERA auch verschiedene, kombinierte Distant-Reading-Visualisierungen an, welche die Textvarianten aggregieren und so eine explorative Analyse ermöglichen (Schütz und Pöckelmann, 2016).

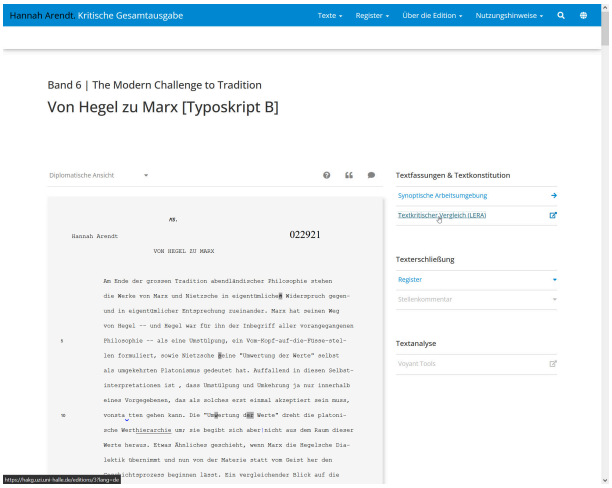


Abb.1: Die Einzeltextansicht der Arendt-Edition verweist für einen textkritischen Fassungsvergleich auf LERA

Die Bedarfe der Arendt-Edition legen den Einsatz von LERA als Forschungswerkzeug und Visualisierungsinstrument nahe. Zunächst wurde exemplarisch die Anbindung für drei Gruppen von Textfassungen umgesetzt, um projektspezifische Anforderungen an LERA zu bestimmen. In einem ersten Schritt wird LERA durch die Editor:innen bei der Erforschung der Unterschiede und Gemeinsamkeiten zwischen Textfassungen genutzt und kann so z. B. Erkenntnisse zur Textgenese liefern und die Kollationierungsarbeit unterstützen. Als ‚semi-automatisches' Werkzeug liefert LERA Vorschläge für Textalignierung und variante Segmente, die in der Bedienoberfläche weiterbearbeitet und so in einem iterativen Prozess vervollständigt und präzisiert werden können. Vor allem bei großen Textmengen lassen sich Abweichungen zwischen Einzeltexten gegenüber einem manuellen Abgleich mit wesentlich geringerem Aufwand identifizieren. Die synoptische Vergleichsansicht ermöglicht hierbei durch die farbcodierte Hervorhebung der varianten Textstellen zudem das Aufdecken von Transkriptionsfehlern. Die semi-automatisch erzeugten Analysedaten werden in einem zweiten Schritt genutzt, um im Webportal die Varianten für die forschenden Leser:innen der Edition zu visualisieren. Ausgehend von der Einzelansicht eines Textes kann die Vergleichsansicht von LERA für abweichende Textfassungen aufgerufen werden (vgl. Abb. 1). Die Vergleichsansicht bietet durch die Alignierung der Segmente einen Überblick über die in Arendts Fall oft großen Textmengen sowie über theoretisch beliebig viele Fassungen und erlaubt damit, Abweichungen auf einer Makro- und Mikroebene schnell zu identifizieren und zu untersuchen (vgl. Abb. 2). Durch die Kooperation mit LERA gelingt der Edition ein erster Schritt hin zu einer virtuellen Forschungsumgebung, die vertiefte Einblicke in Arendts Denk- und Schreibprozesse erlaubt.

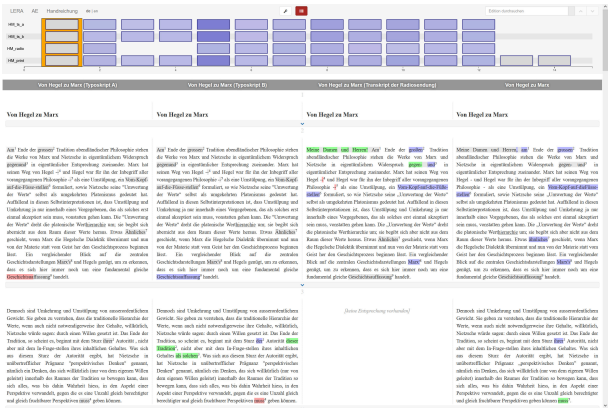


Abb. 2: Synoptische Gegenüberstellung von Textfassungen unter Hervorhebung der Textvarianten mit LERA

Der Einsatz im Kontext der Arendt-Edition zeigt zugleich auch neue Bedarfe und Verbesserungspotenziale für LERA auf. So wurden die Mechanismen für die Verarbeitung von XML-Strukturen verfeinert, z. B. als Ausgangspunkt für eine Segmentierung beim Datenimport oder als Kriterien für eine jeweils spezifische Behandlung verschiedener XML-Markupkonstrukte im Textvergleich. Damit einhergehend wurden die Konfigurationsmöglichkeiten von LERA erweitert, insbesondere die Steuerung der Tokenisierung (z. B. bei Worttrennungen oder Inline-Markup) sowie die Filtereinstellungen für Textvergleich und -wiedergabe, und das User Interface hinsichtlich der Zugänglichkeit und Bedienbarkeit optimiert.

Die Anforderung der Arendt-Edition, den XML-Datenexport aus LERA in den Datenbestand der Edition zurückspielen und z. B. für die Anzeige von textkritischen Apparaten einsetzen zu können, motiviert die derzeitige Weiterentwicklung des LERA-Exportformats gemäß der

TEI-Richtlinien mit dem Ziel, auch für Dritte eine vielfältige Nachnutzung durch einen interoperablen Datenstandard zu gewährleisten. Dieses und die Erweiterung von LERA für den strukturierten Vergleich mehrsprachiger Textfassungen, die nicht strikt parallele Texte sind, sind Gegenstand eines aktuellen DFG-Projekts[10]. Die Synergien, die sich aus der Kooperation herausgebildet haben, bieten über die beteiligten Projekte hinaus einen Mehrwert für die Forschungslandschaft im Bereich der digitalen Editionen.

Fußnoten

1. Siehe https://www.arendteditionprojekt.de. Seit 2020 wird das Editionsprojekt als Langfristvorhaben durch die DFG gefördert: https://gepris.dfg.de/gepris/projekt/421689821.
2. Siehe https://hannah-arendt-edition.net .
3. Siehe https://www.wallstein-verlag.de/reihen/hannah-arendt-kritische-gesamtausgabe.html.
4. Vgl. auch aktuelle Editionsvorhaben wie *Arthur Schnitzler digital* (https://www.schnitzler-edition.net), die *Historisch-kritische Edition von Goethes Faust* (https://faustedition.net) oder auch die *Uwe Johnson Werkausgabe* (http://www.uwe-johnson-werkausgabe.de), die allesamt Werkzeuge zur Analyse und Exploration der publizierten Texte anbieten.
5. Ein Textvergleich auf der Makroebene soll den Text in seiner Gesamtheit betrachten und sich auf anhand bestimmter Vorgaben ermittelte Segmente beziehen, z. B. Absatzstrukturen. Ein Vergleich auf der Mikroebene soll innerhalb der Segmente, die sich gemäß Makrovergleich entsprechen, bestimmte Token betrachten, z. B. Wörter, Zahlen und Satzzeichen.
6. LERA (Locate, Explore, Retrace and Apprehend complex text variants) wird seit 2013 im Rahmen interdisziplinärer Forschungsprojekte an der Martin-Luther-Universität Halle-Wittenberg entwickelt. Ein Demonstrator der Software mit verschiedenen Textbeispielen ist unter https://lera.uzi.uni-halle.de verfügbar.
7. Im Gegensatz zu LERA oder auch CollateX (Dekker und Middell, 2011) arbeiten viele Kollationierungswerkzeuge, welche für mehr als zwei zu vergleichende Textfassungen gleichzeitig ausgelegt sind, nach dem klassischen Leitquellenprinzip, bei dem eine Textfassung als Basistext für den Vergleich ausgewählt werden muss. So zum Beispiel eComparatio (Bräckel et al., 2019), Juxta (Wheeles und Jensen, 2013) oder TUSTEP (Ott, 2000).
8. Siehe https://hannah-arendt-edition.net/v1/ae1254820.
9. Siehe https://hakg.uzi.uni-halle.de/editions/3.
10. Siehe https://www.cedis.fu-berlin.de/services/projektentwicklung/aktuell/semi-automatische-kollationierung/index.html.

Bibliographie

Bräckel, O., H. Kahl, F. Meins und C. Schubert. 2019. “eComparatio - A Software Tool for Automatic Text Comparison.” In Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution. De Gruyter Saur, 221–238. https://doi.org/10.1515/9783110599572-013 .

Dekker, R. H. und G. Middell. 2011. “Computer-supported collation with CollateX: managing textual variance in an environment with varying requirements. ” Supporting Digital Humanities 2011, Kopenhagen, 17.–18.11.2011.

Etling, F. und I. Kieslich. 2024. “Transformation als inhärentes Textphänomen . ” Angenommen als Beitrag zur 20. internationalen Tagung der Arbeitsgemeinschaft für germanistische Edition zum Thema Edition als Transformation. Bedingungen, Formen, Interessen und Ziele editorischer Präsentationen, Wuppertal, 21.–24. Februar 2024.

Ott, W. 2000. “Strategies and tools for textual scholarship: the Tübingen system of text processing programs (TUSTEP).” In *Literary and Linguistic Computing*, Volume 15, Issue 1, 93–108. https://doi.org/10.1093/llc/15.1.93.

Pöckelmann, M., A. Medek, J. Ritter und P. Molitor. 2023. “LERA - An interactive platform for synoptical representations of multiple text witnesses.” In *Digital Scholarship in the Humanities*, Volume 38, Issue 1, 330–346. https://doi.org/10.1093/llc/fqac021.

Schütz, S. und M. Pöckelmann. 2016. “LERA - Explorative Analyse komplexer Textvarianten in Editionsphilologie und Diskursanalyse.” DHd2016, 3. Jahrestagung der Digital Humanities im deutschsprachigen Raum, Leipzig, 07.–12.03.2016. https://doi.org/10.5281/zenodo.4645364 .

Wheeles, D. und K. Jensen. 2013. “Juxta Commons.” DH2013, Proceedings of the 24. international annual conference of Digital Humanities, Lincoln, 16. – 19.07.2013. 545–546.

Über den Text hinaus: Die Edition eines Historiogramms

Cugliana, Elisa
elisa.cugliana@uni-koeln.de
Bergische Universität Wuppertal, Deutschland
ORCID: 0000-0002-6460-2954

Krottmaier, Sina
sina.krottmaier@uni-graz.at
Universität Graz, Österreich

ORCID: 0000-0002-7966-7996

Rouxel, Lennart

lennart.rouxel@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland

Peter von Poitiers, französischer Gelehrter an der Pariser Universität, schuf im ausgehenden 12. Jahrhundert mit dem *Compendium historiae in genealogia Christi* ein theologisches Lehrwerk, welches die Bibelgeschichte und die in ihr vorkommenden zentralen Personen in ihrer Abstammung und Chronologie als Historiogramm (Worm, 2021) präsentiert. Im Zentrum dieses Diagramms liegt die genealogische Abstammungslinie Christi, basierend auf ihren Beschreibungen im Alten Testament, die sich von Adam und Eva über die sieben großen Patriarchen zu Jesus erstreckt. Die Namen der Personen sind in Medaillons dargestellt, deren Verbindungen neben ihrer genealogischen Bedeutung auch als geistige Verbindung oder Tradition verstanden werden müssen. Neben der Abstammungslinie Christi sind weitere Nebenlinien dargestellt; so findet man beispielsweise die Abfolge der jüdischen Hohepriester oder der assyrischen Könige. Die systematisierende Natur des Werkes wird um Diagramme, die sich exegetisch mit bedeutsamen Bibelstellen befassen (etwa der innere Aufbau der Arche), sowie durch Illustrationen, die biblische Konzepte vermitteln sollen, ergänzt. Hinzu kommen zahlreiche textuelle Erläuterungen zu bestimmten Personen, bedeutenden Ereignissen und zur Bibelgeschichte. Eine von Peter von Poitiers selbst verfasste Handschrift ist nicht überliefert, jedoch bezeugen zahlreiche frühdatierte Kopien eine rasche Ausbreitung seines Werkes in ganz Europa.

Im Rahmen unseres Projekts „History as a Visual Concept: Peter of Poitiers' Compendium historiae“ (Worm et al., 2023) mit einer Laufzeit von 3 Jahren (2023-2025) soll dieses Werk digital erschlossen werden. Für das Edieren von textuellen Werken gibt es neben den Verfahren der klassischen Textkritik inzwischen etablierte digitale Methoden (z.B. TEI für Text als Sprache und Werkstruktur oder RDF für die semantische Dimension) und auch Text-Bild Beziehungen sind in der digitalen Editorik keine Neuheit mehr (s. Foys et al., 1996-2003; Stronks et al., 2003-2006; Schmidt und Šimek, 2011-2015). Meist wird jedoch entweder der Text oder das Bild als wichtiger erachtet.

Für ein Werk, welches vor allem aus bildhaften Narrativen unterschiedlicher Natur, essenziell einem Graphen, aber auch aus diversen Textabschnitten besteht, bedarf es jedoch sowohl für die Erschließung des Werkes selbst als auch für die Präsentation der Edition anderer Zugänge. Um den vielfältigen Ebenen des *Compendiums* gerecht zu werden, streben wir den Ansatz einer mehrdimensionalen Edition an, welche die graphische und bildhafte Natur des *Compendiums* berücksichtigt.

Geplant ist, dass eine abstrahierte Visualisierung des Werkes - der sog. „Navigations-Graph“, (s. Abb. 1) - als Einstieg in und Navigation durch die Edition dient. Dieser „NavGraph“ repräsentiert eine rekonstruierte, kanonische Fassung des *Compendiums* und enthält den Stammbaum von Adam und Eva bis Christus (dz. 448 Knoten), die exegetisch bedeutsamen Diagramme (5) sowie die deskriptiven Texte (dz. 166) und erklärende Beischriften zu Gruppen und Ereignissen (dz. 83), die Graphen und Diagramme begleiten.

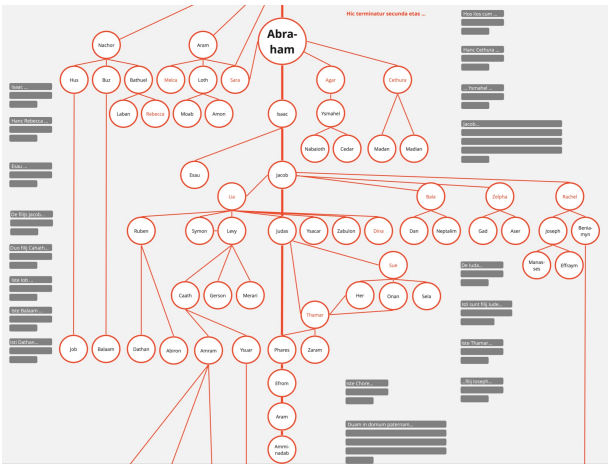


Abb. 1 Ausschnitt aus dem NavGraph

Durch einen Klick auf die einzelnen Objekte sollen User:innen in die nächsttiefere Ebene der Edition eintauchen, die sowohl die eigentliche Edition bietet als auch den Varianzen der Überlieferung gewidmet ist. Hier sollen zusammengehörende Knoten des Stammbaums in ihrem Zusammenhang gezeigt (s. Abb. 2), ein kritischer Text geboten und die textuelle sowie visuelle Varianz repräsentativer, vorab definierter Überlieferungsgruppen des *Compendiums* gezeigt werden.

Abb. 2 Beispieldarstellung der strukturellen Varianten von den „Kindern von Adam und Eva“

Das Grundgerüst für dieses Vorhaben stellt ein Netzwerk aus IDs dar, die für sämtliche Inhalte (Knoten und Knotengruppen, Diagramme, Textblöcke, Illustrationen, sowie deren Varianten) während der TEI-Kodierung des *Compendiums* vergeben werden. So lassen sich die jeweiligen Elemente in den einzelnen Manuskripten adressieren und 'auf einer Ebene' gegenüberstellen.

Essenziell für diese Tiefenerschließung ist die vorangegangene Breitenerschließung der Überlieferung. Die dabei entstehende Datenbank der Handschriftenbeschreibungen aller bisher bekannten Textzeugen gibt einerseits einen Überblick über die Verteilung und Attribute bekannter Handschriften, andererseits werden dadurch die editorischen Selektionsprozesse transparent gemacht. Die Daten-

bank erweitert die von Piggin (2019) erstellte statische Liste der Handschriften um weitere Kategorien (z.B. zu den visuellen Komponenten) und ergänzt die Sammlung um bis dato noch nicht erfasste *Compendia*. So konnte Piggins Liste von 220 Handschriften bereits jetzt auf rund 300 erweitert werden - Tendenz steigend. Im Zusammenspiel mit der Tiefenerschließung entsteht somit nicht nur eine offen zugängliche, kritische Ausgabe, charakterisiert durch einen visuellen Ansatz zur graphischen Metastruktur des Werkes, sondern auch ein Überblick über die Überlieferungsgeschichte, den sozio-kulturellen Kontext und die Rezeption des *Compendiums*.

Das eingereichte Poster wird den Stand des Projekts nach einem Jahr Laufzeit vorstellen und dabei auf den intendierten Workflow von Tiefen- sowie Breitenerschließung und die damit verbundenen technischen Lösungsansätze (Stichwort: NavGraph) eingehen.

Das Poster soll besonders Kolleg:innen ansprechen, die sich mit ähnlich komplexen multimodalen Quellen beschäftigen und bei der Modellierung nebst Rückgriff auf Standards wie XML/TEI, SVG und RDF auch eine quellennahe visuelle Umsetzung anstreben.

Bibliographie

Foys, Martin K., James F. Caccamo, Jody Evenson, und Erica L. Pittman. 1996-2003. „The Bayeux Tapestry Digital Edition“. Digitale Edition. The Bayeux Tapestry Digital Edition. https://web.archive.org/web/20230130090658/http://www.sd-editions.com/bayeux/online/index.html#/facsimile%3DBayeux%26panel%3D1 (zugegriffen: 04. Dezember 2023).

Piggin, Jean-Baptiste. 2019. „Peter’s Stemma.“ In *Piggin.net* (blog). https://web.archive.org/web/20231114105046/https://www.piggin.net/stemmahist/petercatalog.htm (zugegriffen: 04. Dezember 2023).

Schmidt, Peter, und Jakub Šimek. 2011-2015. „Welscher Gast digital“. Digitale Edition. Welscher Gast digital. https://web.archive.org/web/20231028193254/https://digi.ub.uni-heidelberg.de/wgd/ (zugegriffen: 04. Dezember.2023).

Stronks, E., P. Boot, J. Tilstra, H. Brandhorst, M. de Gruijter, D. Stiebral, H. van Baren, G. Huijing, und J.A. Blans. 2003-2006. „Emblem Project Utrecht“. Digitale Edition. Emblem Project Utrecht - Dutch Love Emblems of the Seventeeth Century. https://web.archive.org/web/20230329003122/https://emblems.hum.uu.nl/ (zugegriffen: 04. Dezember 2023).

Worm, Andrea. 2021. *Geschichte und Weltordnung: graphische Modelle von Zeit und Raum in Universalchroniken vor 1500*. Berlin: Deutscher Verlag für Kunstwissenschaft.

Worm, Andrea, Patrick Sahle und Roman Bleier. 2023. „Geschichte als visuelles Konzept: Peter von Poitiers' Compendium historiae. https://web.archive.org/web/20231204065548/https://compendium-historiae.uni-graz.at/de/ (zugegriffen: 04. Dezember 2023).

CANSpiN: Zur computergestützten Analyse narrativen Raums im Roman des 19. und 20. Jahrhunderts

Lemke, Marc
marc.lemke@uni-rostock.de
Universität Rostock, Deutschland
ORCID: 0009-0004-8065-8191

Henny-Krahmer, Ulrike
ulrike.henny-krahmer@uni-rostock.de
Universität Rostock, Deutschland
ORCID: 0000-0003-2852-065X

Kellner, Nils
nils.kellner@uni-rostock.de
Universität Rostock, Deutschland

Mit unserer Einreichung stellen wir den aktuellen Arbeitsstand des Projekts „Computational Approaches to Narrative Space in 19th and 20th Century Novels“ (CANSpiN) vor, das im Rahmen des DFG-Schwerpunktprogramms „Computational Literary Studies“ (SPP 2207) von April 2023 bis März 2026 gefördert wird. Ziel des Vorhabens ist es, computergestützte Methoden zur Erkennung und Analyse narrativen Raums in literarischen Texten zu entwickeln und diese Methoden für die Untersuchung literaturhistorischer Fragen zum Verhältnis von Raum und nationaler Identität in deutsch- und spanischsprachigen Romanen des 19. und 20. Jahrhunderts zur Anwendung zu bringen. Dieser Zielstellung entspricht die Zusammensetzung der Projektgruppe, die aus Wissenschaftler:innen der Germanistik, Romanistik, den Digital Humanities und der Mathematik besteht.

Ausgangspunkt der Überlegungen ist die Definition des narrativen Raums: Darunter verstehen wir im weiteren Sinne die Räumlichkeit eines narrativen Textes nach Schumacher (2022b), die durch raumreferentielle sprachliche Ausdrücke bestimmt ist. In einem engeren narratologischen Sinne begreifen wir narrativen Raum als den Raum der erzählten Welt, wie er durch die Erzählung konstituiert ist (Genette, 2010). Wie dieser Raum strukturell und funktional zu beschreiben ist, dazu existieren bereits zahlreiche

literaturwissenschaftliche Vorschläge (Dennerlein, 2009; Piatti, 2008; Ryan, 2014; Lotman, 1977; Renner, 2004). Wie raumanalytische Zugänge formalisiert und auf konkrete Textmerkmale abgebildet werden können, dazu bieten Arbeiten der Computational Literary Studies schon erste Antworten (Viehhauser und Barth, 2017; Barth, 2021, 2022; Viehhauser, 2020; Schumacher, 2022a, 2022b).

Auf diesem Forschungsstand setzen wir auf: In einem offenen explorativen Verfahren erproben wir verschiedene raumanalytische Kategorien-Sets hinsichtlich ihrer literaturwissenschaftlichen Aussagekraft und ihrer computergestützten Operationalisierbarkeit (Arbeitspaket 1). Für jedes dieser Sets entwickeln wir Annotationsrichtlinien, um die verwendeten Korpora computergestützt zu annotieren und mit quantitativen Methoden vergleichend zu untersuchen. So haben wir ein erstes Kategoriensystem CANSpiN.CS1 für die Annotation von Raumreferenzen in Erzähltexten definiert und mit Hilfe des Tools CATMA (Gius et al., 2023) erprobt. Es ist geplant, auf diesem Wege eine Ground Truth aufzubauen, mit der im Tool NTEE (Lemke et al., 2023) Modelle für die Erkennung von Raumentitäten trainiert werden können, welche die semi-automatisierte, vollständige Annotation der Textkorpora im Projekt erlauben.

Das Interesse des Projekts richtet sich dabei insbesondere auch auf die Möglichkeiten und Grenzen des Deep Learnings mit Sprachmodellen der BERT-Architektur (Devlin et al., 2019) für die computergestützte Annotation literaturwissenschaftlich definierter Textphänomene: Um diese auszuloten, werden die im Arbeitspaket 1 erzeugten Modelle mit Methoden eines Explainable AI-Ansatzes (XAI) untersucht (Arbeitspaket 2), was konkrete Maßnahmen und Vorschläge zur Optimierung der Annotationsprozesse erwarten lässt, aber auch Erkenntnisse über die technischen Grenzen dieses Vorgehens (Rogers et al., 2021).

Die Korpora werden derzeit ausgehend von verschiedenen Vorarbeiten zusammengestellt, aus deutschsprachigen Romanen des 19. Jahrhunderts (Zeit: 1790-1910, Ziel: etwa 200 Romane, Quellen: ELTeC-deu (Konle et al., 2021), DTA (Berlin-Brandenburgischen Akademie der Wissenschaften, 2023), TextGrid Repository (TextGrid Konsortium, 2014)), spanischsprachigen Romanen des 19. Jahrhunderts (Zeit: 1790-1870, Ziel: etwa 100 Romane, Quellen: ELTeC-spa (Navarro-Colorado et al., 2021), Biblioteca Virtual Miguel de Cervantes (Centro de Humanidades Digitales en la Universidad de Alicante, 2021)), spanischsprachigen Romanen aus Lateinamerika des 19. Jahrhunderts (Zeit: 1830-1910, Ziel: etwa 200 Romane, Quellen: Corpus de novelas hispanoamericanas del siglo XIX (Conha19) (Henny-Krahmer, 2021)) und deutschsprachigen Romanen des 20. Jahrhunderts (Zeit: 1950-2000, Ziel: etwa 100 Romane, Quellen: TEI-Dateien der Uwe Johnson-Werkausgabe, E-Books (u.a. zu den Autor:innen Heinrich Böll, Christine Brückner, Uwe Johnson, Walter Kempowski, Christa Wolf)). Für ein einheitliches Format der Texte haben wir uns für das Schema ELTeC Level 0 (Burnard, 2012) entschieden, das wir für unsere Zwecke im CANSpiN-Projekt anpassen.

Die Auswahl der Korpora gründet auf unserer Arbeitshypothese, dass narrativer Raum in Romanen sprach- und zeitübergreifend ein für die Analyse zugängliches Phänomen ist, das unterschiedlich ausgeformt, also dargestellt und funktionalisiert sein kann. Durch die quantitativen Analysen aus Arbeitspaket 1 erwarten wir in dieser Hinsicht neue Erkenntnisse zu literaturhistorischen Fragestellungen (Arbeitspaket 3), etwa zur Darstellung und Semantik von Raum in Romanen im Kontext der Nationenbildung im 19. Jahrhundert in Deutschland und Lateinamerika (Sommer, 1993; Peñaranda Medina, 1994; Hanway, 2003; Viel, 2009; Ferrer, 2018) und der Herausbildung zweier deutscher Identitäten in der Nachkriegsliteratur des 20. Jahrhunderts (Bond, 1996; Erll et al., 2003; Helbig et al., 2007; Westphal, 2007; Nies, 2018).

Bibliographie

Barth, Florian. 2021. “Konzept und Klassifikation literarischer Raumentitäten.” In *INFORMATIK 2020*, hg. von Ralf Heinrich Reussner, Anne Koziolek und Robert Heinrich, 1281–1293. Bonn: Gesellschaft für Informatik. https://dx.doi.org/10.18420/inf2020_120.

Barth, Florian. 2022. “Von der literaturwissenschaftlichen Theorie zu maschinellen Erkennung: Operationalisierung von Raumentitäten und Settings.” In *Digitale Verfahren in der Literaturwissenschaft*, hg. von Jan Horstmann und Frank Fischer. Sonderausgabe #6 von *Textpraxis. Digitales Journal für Philologie*. https://doi.org/10.17879/64059429732.

Berlin-Brandenburgischen Akademie der Wissenschaften (Hg.). 2023. *Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*. Berlin. https://www.deutschestextarchiv.de.

Bond, Greg. 1996. “'weil es ein Haus ist, das fährt.' Rauminszenierungen in Uwe Johnsons Werk.” In *Johnson-Jahrbuch* 3, 72–96. Göttingen: Vandenhoeck & Ruprecht.

Burnard, Lou. 2012. *COST-ELTeC/Schemas*. Version 0.8.1. https://doi.org/10.5281/zenodo.4679363.

Centro de Humanidades Digitales en la Universidad de Alicante (Hg.). 2021. *Biblioteca Virtual Miguel de Cervantes*. Alicante. https://www.cervantesvirtual.com.

Dennerlein, Katrin. 2009. *Narratologie des Raumes*. Berlin und New York: De Gruyter.

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), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1423.

Erll, Astrid, Marion Gymnich und Ansgar Nünning (Hg.). 2003. *Literatur – Erinnerung – Identität: Theoriekonzeptionen und Fallstudien*. Trier: VWT.

Ferrer, José Luis. 2018. *La invención de Cuba: Novela y nación (1837-1846)*. Madrid: Editorial Verbum.

Genette, Gérard. 2010. *Die Erzählung*. 3., durchges. und korrigierte Aufl. Paderborn: Fink.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Mareike Schumacher und Dominik Gerstorfer. 2023. *CATMA*. Version 7.0.0. https://doi.org/10.5281/zenodo.1470118.

Hanway, Nancy. 2003. *Embodying Argentina: Body, Space and Nation in 19th Century Narrative*. Jefferson, NC: McFarland.

Helbig, Holger, Kristin Felsner, Sebastian Horn und Therese Manz (Hg.). 2007. *Weiterschreiben: zur DDR-Literatur nach dem Ende der DDR*. Berlin: Akademie Verlag.

Henny-Krahmer, Ulrike (Hg.). 2021. *Corpus de novelas hispanoamericanas del siglo XIX (conha19)*. Version 1.0.1. https://doi.org/10.5281/zenodo.4766987.

Konle, Leonard, Fotis Jannidis, Carolin Odebrecht und Lou Burnard. 2021. *German Novel Corpus (ELTeC-deu)*. Version 1.0.0. https://doi.org/10.5281/zenodo.4662482.

Lemke, Marc, Konrad Sperfeld und Jochen Zöllner. 2023. *NEISS TEI Entity Enricher*. Version 1.1.1. https://github.com/NEISSproject/tei_entity_enricher.

Lotman, Jurij Michailowitsch. 1977. *Die Struktur literarischer Texte*. München: Fink.

Navarro-Colorado, Borja, Lou Burnard und Carolin Odebrecht. 2021. *Spanish Novel Corpus (ELTeC-spa)*. Version 0.9.1. https://doi.org/10.5281/zenodo.4662603.

Nies, Martin (Hg.). 2018. *Raumsemiotik. Räume – Grenzen – Identitäten.* https://web.archive.org/web/20220124091647/https://www.kultursemiotik.com/wp-content/uploads/2019/10/Raumsemiotik-Martin-Nies-Hg._SMKS-Online-No.4-2018-red.pdf.

Peñaranda Medina, Rosario. 1994. *La novela modernista hispanoamericana: estrategias narrativas*. Valencia: Universitat de Valencia.

Piatti, Barbara. 2008. *Die Geographie d er Literatur. Schauplätze, Handlungsräume, Raumphantasien*. Göttingen: Wallstein.

Renner, Karl Nikolaus. 2004. "Grenze und Ereignis. Weiterführende Überlegungen zum Ereigniskonzept von Jurij M. Lotman." In *Norm – Grenze – Abweichung: Kultursemiotische Studien zu Literatur, Medien und Wirtschaft*, 357–81. Passau: Verlag Karl Stutz. https://web.archive.org/web/20230717115613/https://www.kultursemiotik.com/wp-content/uploads/2015/01/Renner_Grenze-und-Ereignis.pdf.

Rogers, Anna, Olga Kovaleva und Anna Rumshisky. 2021. "A Primer in BERTology: What We Know About How BERT Works." In *Transactions of the Association for Computational Linguistics* 8, 842–866. https://doi.org/10.1162/tacl_a_00349.

Ryan, Marie-Laure. 2014. "Space." In *The Living Handbook of Narratology*, hg. von Peter Hühn, Jan Christoph Meister, John Pier und Wolf Schmid. Hamburg: Hamburg University. https://web.archive.org/web/20221210041153/https://www-archiv.fdm.uni-hamburg.de/lhn/node/55.html.

Schumacher, Mareike. 2022a. "'Wie Wölkchen Im Morgenlicht' – Zur automatisierten Metaphern-Erkennung und der Datenbank literarischer Raummetaphern LaRa." In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*, 232–236. https://doi.org/10.5281/zenodo.6304590.

Schumacher, Mareike. 2022b. *Orte und Räume im Roman. Ein Beitrag zur digitalen Literaturwissenschaft*. Berlin und Heidelberg: Metzler. https://doi.org/10.1007/978-3-662-66035-5.

Sommer, Doris. 1993. *Foundational Fictions. The National Romances of Latin America*. Berkeley, LA: University of California Press.

TextGrid Konsortium. 2014. *TextGrid: Virtuelle Forschungsumgebung für die Geisteswissenschaften*. Göttingen: TextGrid Konsortium. https://www.textgrid.de.

Viehhauser, Gabriel. 2020. "Zur Erkennung von Raum in narrativen Texten. Spatial Frames und Raumsemantik als Modelle für eine digitale Narratologie des Raums." In *Reflektierte Algorithmische Textanalyse*, 373–388. Berlin und Boston: De Gruyter. https://doi.org/10.1515/9783110693973-015.

Viehhauser, Gabriel und Florian Barth. 2017. "Towards a Digital Narratology of Space." In *Digital Humanities 2017. Conference Abstracts*, 643–646. Montréal. https://web.archive.org/web/20230615201230/https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf.

Viel, Bernhard. 2009. *Utopie der Nation: Ursprünge des Nationalismus im Roman der Gründerzeit*. Berlin: Matthes & Seitz.

Westphal, Nicola. 2007. *Literarische Kartografie. Erzählter Raum in den Romanen Uwe Johnsons*. Göttingen: Vandenhoeck & Ruprecht.

Community Building auf Citizen Science-Projektplattformen: Ja, nein, vielleicht?

Heinisch, Barbara

barbara.heinisch@univie.ac.at
Universität Wien, Österreich
ORCID: 0000-0002-1362-4088

Die Zusammenarbeit mit Bürger:innen in den Geisteswissenschaften blickt auf eine lange Tradition zurück. Auch heutzutage erfreut sich die Zusammenarbeit zwischen professionellen Forschenden und Bürger:innen in vielen Disziplinen der (digitalen) Geisteswissenschaften, wie Lexikografie, Geschichte oder Kunst zunehmender Beliebtheit.

Die Beiträge der Bürger:innen in Form von Citizen Science (ECSA 2020) umfassen in den digitalen Geisteswissenschaften beispielsweise Datensammlung, Transkription oder Annotation. Das Ausmaß der Einbindung von Angehörigen der Öffentlichkeit in Citizen Science-Projekten in den digitalen Geisteswissenschaften rangiert dabei zwischen der Erledigung von Mikroaufgaben, wie der Transkription eines Satzes bis hin zur gemeinsamen Gestaltung von ganzen Forschungsprojekten.

Um Teilnehmende für geisteswissenschaftliche Forschungsprojekte zu gewinnen, setzen Forschende teilweise auf Citizen Science-Projektplattformen, wie *Bürger schaffen Wissen*, *Österreich forscht* oder *Zooniverse*. Auch spezialisierte Plattformen, die nur ein ganz bestimmtes Thema (z.B. Linguistik oder Kunstgeschichte) oder eine Methode (z.B. Transkription oder Übersetzung) abdecken, sind in der geisteswissenschaftlichen Citizen Science-Landschaft zu finden.

Das Potenzial von Citizen Science liegt in der Öffnung der Wissenschaft, der Demokratisierung der Wissenschaft (Irwin 1995; Scanlon / Herodotou 2022), der Berücksichtigung einer Vielfalt von Epistemologien in der Forschung (Jaeger et al. 2022), der Erhöhung der wissenschaftlichen Bildung und des Wissens über Wissenschaft selbst (Ceccaroni et al. 2017; Queiruga-Dios et al. 2020), der Stärkung des Vertrauens in die Wissenschaft und dem transformativen Potenzial zur Erreichung der Ziele für nachhaltige Entwicklung (Fritz et al. 2019).

Dieses transformative Potenzial lässt sich auf den genannten Citizen-Science-Plattformen jedoch nur in begrenztem Umfang realisieren. Dies liegt an der Natur der Plattformen, die die Art der Zusammenarbeit mit den Teilnehmenden bereits prädeterminiert. Diese Plattformen eignen sich besonders gut für das Crowdsourcing kleinerer Aufgaben, die von einer großen Anzahl von Personen erledigt werden können. Diese Plattformen bieten normalerweise verschiedene Citizen Science-Projekte an, zu denen Freiwillige beitragen können.

Diese Plattformen, die entweder Projekte aus einer Vielzahl von wissenschaftlichen Disziplinen oder ein bestimmtes Themengebiet abdecken, konzentrieren sich oft auf Aktivitäten in Form von Kleinstaufgaben, die leicht online allein von zu Hause aus erledigt werden können und in der Regel kaum Interaktion zwischen den Wissenschafter*innen und Teilnehmenden erfordern. Allerdings wird gerade Community Building als ein wesentlicher Faktor für den Erfolg eines Citizen Science-Projekts verstanden (Golumbic et al. 2020).

Daher wurden drei allgemeine Citizen Science-Plattformen (*European Citizen Science, Zooniverse und Bürger schaffen Wissen*), drei thematisch spezialisierte Citizen-Science-Plattformen (*LanguageARC* (Fiumara et al. 2020) im Bereich Linguistik, *ARTigo* (Bry / Schefels 2016) im Bereich Kunstgeschichte, sowie *MicroPasts* (Bonacchi et al. 2019) im Bereich kulturelles Erbe und Archäologie) und methodisch spezialisierte Plattformen (*FromThePage* für Transkription, *SPOTTERON* für Georeferenzierung) im Hinblick auf Community Building-Aspekte gemäß Golumbic et al. (2020) analysiert. Gegenstand der Analyse sind somit Foren, Chats, diverse (eingebettete) soziale Medien, sowie reguläre Austauschformate (wie beispielsweise Feedbackrunden) über diese Plattformen .

Die untersuchten Plattformen wurden anhand ihrer Größe und der Dauer ihres Bestehens ausgewählt, da davon auszugehen ist, dass größere und ältere Plattformen eher auf die Bedürfnisse der Benutzer:innen eingehen. Außerdem wird ein Vergleich einer globalen, einer europaweiten und einer nationalen Plattform angestellt. Die themenspezifischen Plattformen repräsentieren die Fachgebiete, die in den *Fields of Science and Technology* angeführt sind. Von den vier darin aufgelisteten Geisteswissenschaften (a) Geschichte, Archäologie, b) Sprach- und Literaturwissenschaften, c) Philosophie, Ethik, Religion und d) Kunstwissenschaften) wurde je eine themenspezifische Plattform ausgewählt, mit Ausnahme von Philosophie, Ethik oder Religion, da zum Zeitpunkt der Erhebung keine spezialisierte Plattform zu finden war.

Da auf diesen Plattformen Crowdsourcing eine zentrale Rolle zukommt, stellt sich die Frage, inwieweit diese Plattformen die Teilnehmenden (sowie die Forschenden) dabei unterstützen, eine Gemeinschaft zu bilden und inwieweit diese Aufgabe den Projekten selbst überlassen wird. Die übergeordnete Forschungsfrage somit ist: Wie wird Community Building auf den untersuchten Plattformen umgesetzt?

Erste Zwischenergebnisse verdeutlichen, dass allgemeine Citizen Science-Projektplattformen kein übergeordnetes Community Building, wie beispielsweise Austauschmöglichkeiten über Projektgrenzen hinweg, aufgrund ihres Aufbaus bezwecken. Community Building findet in diesem Fall primär auf Projektebene statt, entweder mittels projekteigener Kommunikations- und Informationstechnologie oder, wie im Falle von *Zooniverse*, auf der Zooniverse-Projektseite selbst. Thematisch spezialisierte Citizen Science-Plattformen hingegen verfügen teils über Community-Features und die Möglichkeit, auf die Bedürfnisse der wissenschaftlichen Projekte einzugehen. Methodisch spezialisierte Citizen Science-Projektplattformen fungieren hierbei wiederum verstärkt auf Projektebene, in dem Community-Funktionalitäten für jedes Projekt angeboten werden, um den Austausch und die Interaktion zwischen den Benutzer:innen der jeweiligen App im Citizen Science-Projekt zu stärken.

Dadurch wird deutlich, dass Community Building von Teilnehmenden in Projekten, die auf Citizen Science-Projektplattformen geführt werden, in erster Linie auf Projektebene selbst (und nur teilweise über diese Plattformen) stattfindet. Daher bleibt offen, ob ein Community Building von Teilnehmenden in Citizen Science-Projekten auch über Projektgrenzen hinaus einen Mehrwert bieten könnte.

Bibliographie

Bonacchi, Chiara/Bevan, Andrew/Keinan-Schoonbaert, Adi/Pett, Daniel/Wexler, Jennifer (2019):

“Participation in heritage crowdsourcing”, in: Museum Management and Curatorship 34, 2: 166–182.

Bonney, Rick/Ballard, Heidi/Jordan, Rebecca/ McCallie, Ellen/Phillips, Tina/Shirk, Jennifer/ Wilderman, Candie C. (2009): *Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report*. Center for Advancement of Informal Science Education. http://files.eric.ed.gov/fulltext/ED519688.pdf.

Bry, Francois / Schefels, Clemens (2016): *An Analysis of the ARTigo Gaming Ecosystem With a Purpose*, 1–10. https://www.en.pms.ifi.lmu.de/publications/PMS-FB/PMS-FB-2016-6/PMS-FB-2016-6-paper.pdf.

Ceccaroni, Luigi/Bowser, Anne/Brenton, Peter (2017): “Civic Education and Citizen Science”, in: Ceccaroni, Luigi / Piera, Jaume (eds.): *Analyzing the role of citizen science in modern research*. Hershey PA, Information Science Reference, 1–23.

ECSA (2020): *ECSA’s characteristics of citizen science*. https://ecsa.citizen-science.net/sites/default/files/ecsa_characteristics_of_citizen_science_-_v1_final.pdf.

Fiumara, James/Cieri, Christopher/Wright, Jonathan/Liberman, Mark (2020): “LanguageARC: Developing Language Resources Through Citizen Linguistics”, in: *Proceedings of the LREC 2020 Workshop “Citizen Linguistics in Language Resource Development”*. Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020, 1–6.

Fritz, Steffen/See, Linda/Carlson, Tyler/Haklay, Mordechai/Oliver, Jessie L./Fraisl, Dilek/Mondardini, Rosy/Brocklehurst, Martin/Shanley, Lea A./Schade, Sven/Wehn, Uta/Abrate, Tommaso/Anstee, Janet/ Arnold, Stephan/Billot, Matthew/Campbell, Jillian/ Espey, Jessica/Gold, Margaret/Hager, Gerid/He, Shan/ Hepburn, Libby/Hsu, Angel/Long, Deborah/Masó, Joan/McCallum, Ian/Muniafu, Maina/Moorthy, Inian/ Obersteiner, Michael/Parker, Alison J./Weisspflug, Maike/West, Sarah (2019): “Citizen science and the United Nations Sustainable Development Goals”, in: Nature Sustainability 2, 10: 922–930.

Golumbic, Yaela N. / Baram-Tsabari, Ayelet / Koichu, Boris (2020): Engagement and communication features of scientifically successful citizen science projects. In: Environmental Communication 14 (4), 465–480.

Irwin, Alan (1995): *Citizen science. A study of people, expertise and sustainable development*. London [u.a.], Routledge.

Jaeger, Johannes/Masselot, Camille/Greshake Tzovaras, Bastian/Senabre Hidalgo, Enric/Haklay, Muki/Santolini, Marc (2022): *An Epistemology for Democratic Citizen Science*. https://doi.org/10.31219/osf.io/j62sb.

Queiruga-Dios, Miguel Ángel/López-Iñesta, Emilia/ Diez-Ojeda, María/Sáiz-Manzanares, María Consuelo/ Vázquez Dorrío, José Benito (2020): “Citizen Science for Scientific Literacy and the Attainment of Sustainable Development Goals in Formal Education”, in: Sustainability 12, 10: 1–18.

Scanlon, Eileen/Herodotou, Christothea (2022): “Advancing the Democratization of Research: Citizen Science”, in: Kizilcec, René F. (ed.): *Proceedings of the Ninth ACM Conference on Learning Scale*, New York City, Association for Computing Machinery, 280–283.

Compiling Controlled Vocabularies of Contributor and User Roles for a Platform of Open Educational Resources

Steiner, Petra
petra.steiner@tu-darmstadt.de
TU Darmstadt, Deutschland
ORCID: 0000-0001-8997-2620

Hastik, Canan
canan.hastik@fst.tu-darmstadt.de
TU Darmstadt, Deutschland
ORCID: 0000-0003-1729-4642

Fuhrmans, Marc
marc.fuhrmans@tu-darmstadt.de
TU Darmstadt, Deutschland
ORCID: 0000-0002-9826-018X

Introduction and Objectives

The project DALIA - Knowledge Base for "FAIR data usage and supply” is funded by the German Federal Ministry of Education and Research (BMBF) and by the EU's Reconstruction and Resilience Facility. It aims to provide an infrastructure for educational resources, especially those which are being produced in the NFDI consortia (Sure-Vetter et al., 2021), for relevant RDM networks, and in the long-term for an international, interdisciplinary community. Its main field of interest lies in related FAIR (Wilkinson et al., 2016) data science and research data management.

As a technical link, the DALIA platform for teaching and training materials is being developed. It is based on the DALIA Knowledge Graph which serves as an interlink of the heterogeneous and subject-specific teaching and

learning materials, and assists in making them visible, findable and accessible for users from a wide range of disciplines, career and competence levels. It will implement the knowledge as an ontology in an RDF triplestore which will be reusable and interlinked with other Linked Open Data (LOD) projects.

To this aim, the development of controlled vocabularies which are interdisciplinary usable is an essential foundation. The question of how this can be realized will be answered in feedback loops between developers, contributors, and finally frontend users: DALIA addresses content providers and curators for educational purposes with learning content repositories or future data competence centers using their own community-specific taxonomies, e.g. TaDiRAH (Borek et al., 2021). The sources and resources are heterogenous, and their providers and users have established different metadata standards and formats. Therefore, our metadata must provide solutions for many different applications and be open and extendible.

Structure of the Metadata Categories

Our aim is to create a hierarchy of simple metadata categories with mappings to other metadata schemas and a small (closed) core set for content curation, searching and harvesters. This then can serve as best practice and quality assessment for internal and external project partners and providers.

The metadata categories consist of descriptive metadata, administrative metadata, structural metadata, legal metadata, technical metadata, and usage, quality, and statistical metadata. Here, we introduce the controlled vocabularies concerning the human entities: contributors and their roles in accordance with the user roles of the platform.

The format style is similar to DCTAP (*Dublin Core Tabular Application Profile*) (Coyle et al., 2023) which is a set of elements and definitions for setting up metadata schemas for applications and their validations. It is a table format which is exportable and convertible to RDF structures, providing among others the terms, definitions, cardinalities, mappings to other metadata standards, term types, data types, domains and ranges.

The categories are influenced by Hoebelheinrich et al. (2022), the IEEE Standard for Learning Object Metadata (IEEE Computer Society 2020), the DataCite Metadata Working Group (2021), the DDI Alliance Controlled Vocabulary for Contributor Role (2019), Re3data.org (Strecker et al., 2021), Allgemeines Metadatenprofile for Bildungsressourcen (Pohl et al., 2023), and other standards. Emphasis has been placed on compatibility to the DataCite contributor types, which form the basis for modeling of human contributions for many terminologies, e.g. the Metadata4Ing ontology (Arndt et al., 2022) developed within NFDI4Ing.

Metadata for Human Entities in DALIA

Metadata for human entities mostly apply to administrative and descriptive metadata with their entities of different contributor and learner types. Based on the abovementioned resources and the requirements, we compiled a hierarchical synopsis for contributors including contributor roles. We added or substituted missing or paraphrase definitions, and decided on the term types, to make the classification and ontology entities consistent. Typical contributor types for administrative metadata are author, editor, data manager, data curator but also sponsor and work package leader. The categories of the user types, such as teachers, students, or researchers, are mostly based on the metadata profile by Pohl et al. (2023) which refers to the educational audience roles of the LRMI Concept Schemes (Barker and Sutton, 2017), and the RDA Minimal Application Profiles from Biernacka and Hoebelheinrich (2023). These synopses will be included into the ontology work. It is part of a general compilation of six metadata classes comprising about 30 entries for the minimal set and approximately 400 entries for the extended version. This extended version will be used as an inventory for harmonization.

Acknowledgements

This project with the federal label 16DWWQP07A is funded by the German Federal Ministry of Education and Research (BMBF) and by the EU's Reconstruction and Resilience Facility.

Bibliographie

Arndt, Susanne, Benjamin Farnbacher, Marc Fuhrmans, Stephan Hachinger, Johanna Hickmann, Nils Hoppe, et al. 2022. *Metadata4Ing: An ontology for describing the generation of research data within a scientific activity*: Zenodo. https://zenodo.org/record/7706017. DOI: 10.5281/zenodo.7706017.

Barker, Phil. and Sutton, Stuart, eds. 2017. *LRMI Concept Scheme. LRMI Educational Audience Role Vocabulary*. https://www.dublincore.org/specifications/lrmi/concept_schemes/educationalAudienceRole/.

Biernacka, Katarzyna, Nancy J., and Hoebelheinrich, Nancy J. 2023. *ETHRD-IG Virtual Session: Learning Resources Minimal Metadata Application Profile*. https://docs.google.com/presentation/d/12CCLNMCC9mGjKID4vxOGUJIhoX4XoLTipJsSLLC6mZ8.

Borek, Luise, Canan Hastik, Vera Khramova, Klaus Illmayer, and Jonathan D. Geiger. 2021. “Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR.” In: *Information between data and knowledge. Information Science and its*

Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th-10th March 2021, ed. Thomas Schmidt und Christian Wolff, 321–332. Glückstadt: Verlag Werner Hülsbusch, Fachverlag für Medientechnik und -wirtschaft. (Schriften zur Informationswissenschaft, 74). DOI: 10.5283/EPUB.44951.

Coyle, Karen, Tom Baker, Phil Barker, John Huck, Ben Riesenberg, and Nishad Thalhath. 2023. *DC Tabular Application Profiles (DC TAP) - Primer*. https://github.com/dcmi/dctap/blob/main/TAPprimer.md.

DataCite Metadata Working Group. 2021. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4*. Members of the Metadata Working Group: Madeleine de Smaele, Isabel Bernal Martínez, Robin Dasler, Jan Ashton, Sophie Roy, Martin Fenner et al. DOI: 10.14454/3w3z-sa82.

DDI Alliance. 2019. *DDI Alliance Controlled Vocabulary for Contributor Role* (Version 1.0.2) [Controlled vocabulary]. urn:ddi:int.ddi.cv:ContributorRole:1.0.2. https://vocabularies.cessda.eu/vocabulary/ContributorRole?lang=en.

Hoebelheinrich, Nancy J., Katarzyna Biernacka, Michelle Brazas, Leyla Jael Castro, Nicola Fiore, Margareta Hellström, et al. 2022. *Recommendations for a minimal metadata set to aid harmonised discovery of learning resources*. DOI: 10.15497/RDA00073.

IEEE Computer Society. 2020. *IEEE Standard for Learning Object Metadata*. IEEE Std 1484.12.1™-2020. https://ieeexplore.ieee.org/servlet/opac?punumber=9262116. DOI: 10.1109/IEEESTD.2020.9262118.

Pohl, Adrian, Axel Klinger, Boris Hartmann, Carl Schuurbiers, Fabian Steeg, Manuel Oellers, et al. 2023. *Allgemeines Metadatenprofil für Bildungsressourcen (AMB). Entwurf vom 06. November 2023*. https://dini-ag-kim.github.io/amb/draft/.

Strecker, Dorothea, Roland Bertelmann, Helena Cousijn, Kirsten Elger, Lea Maria Ferguson, David Fichtmueller, et al. 2021. *Metadata Schema for the Description of Research Data Repositories: version 3.1*. Unter Mitarbeit von Florian Fritze, Claudio Fuchs, Agnes Kirchhof, Jens Klump, Claudia Kramer, Jessica Rücknagel et al. DOI: 10.48440/RE3.010.

Sure-Vetter, York, Eva Lübke, Sophie Kraft, and Hendrik Seitz-Moskaliuk. 2021. "Nationale Forschungsdateninfrastruktur (NFDI) e. V. - Satzungsvorstellung". https://doi.org/10.5281/ZENODO.5735196.

Wilkinson, Mark D., Michel Dumontier, Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, 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.

Computerspiele als Darstellungsmedium für immaterielles Kulturerbe: Theoretische Überlegungen

Piontkowitz, Vera
vera.piontkowitz@uni-leipzig.de
Universität Leipzig, Deutschland
ORCID: 0000-0003-3605-3609

Einleitung

Methoden der Digitalisierung haben den Schutz des materiellen Kulturerbes zweifellos vorangebracht. Zunehmend werden auch Lösungen für die Langzeitarchivierung sogenannter *born digitals* gefordert: So formulierte etwa die Gesellschaft für Informatik den „Erhalt des digitalen Kulturerbes" als eine von fünf Grand Challenges der Informatik (Gesellschaft für Informatik e.V., 2014). Dazu gehören auch Computerspiele, deren Archivierung in Fachkreisen zunehmend in den Fokus rückt (Nylund et al., 2021). Ist es aber umgekehrt möglich, dass Computerspiele selbst als Archive dienen? Das Poster untersucht die grundsätzliche Eignung von Computerspielen als Präsentationsmedium für immaterielles Kulturerbe (IKE) – also jenen Ausdruck von Kultur, der gelebt wird und nicht unmittelbar in materieller Form dargestellt werden kann.[1] Zu diesem Zweck erfolgt ein *Mapping* der Bedarfe für den Schutz von IKE auf die Eigenschaften und Vorteile von digitalen Spielen. Anhand von Beispielen aus dem Spiel *Assassin's Creed: Viking Age Discovery Tour* (Ubisoft Montreal, 2021), das viele IKE-Artefakte enthält, wird die Argumentation unterstützt.

Immaterielles Kulturerbe und Computerspiele

Spätestens seit der Verabschiedung der Convention for the Safeguarding of the Intangible Cultural Heritage durch die UNESCO, bemühen sich Kulturinstitutionen, -träger:innen und weitere Akteur:innen um den Schutz von IKE. Bestehende Lösungsansätze sind so vielfältig wie IKE selbst (Hou et al., 2022). Der Ansatz, Computerspiele für diese Aufgabe zu nutzen, ist nicht gänzlich neu: Laiti et al. (2021) untersuchten etwa am Beispiel der Sami-Community, wie das Kulturerbe einer Gemeinschaft im Rahmen von Game Jams in Computerspiele übersetzt werden kann. LaPensée (2021) reflektiert die Entwick-

lung des Spiels *When Rivers Were Trails* in Kollaboration mit nordamerikanischen Indigenen. Im Rahmen des Forschungsprojekts *i-Treasures* entwickelten Forscher:innen ein Framework zur Implementierung von Computerspielen mit Motion-Capture-Daten (Dimitropoulos et al., 2018). Einer Analyse der Repräsentation von Kulturerbe in Computerspielen widmen sich Balela und Mundy (2015) und führen ein geeignetes Analyseframework ein. Die Projekte zeigen, dass Computerspiele ein großes Potenzial für die Darstellung von IKE haben. Doch worauf gründet sich dieses Potenzial? Dazu werden im Folgenden drei Kernargumente dargelegt, die basierend auf umfangreicher Literaturrecherche zu IKE, Computerspielen und deren Schnittstelle, sowie einer Spielanalyse auf Basis des oben genannten Frameworks von Balela und Mundy (2015) ermittelt wurden.

Multimodalität und Prozeduralität

In Computerspielen können verschiedene semiotische Modi verwendet werden, um Informationen zu vermitteln. Dazu gehören Text, Bild, Ton, haptisches Feedback, und Prozeduralität, die es Computerspielen erlaubt, komplexe Konzepte und Prozesse darzustellen, indem Regeln und Aktionen im Spiel manifestiert werden (Hawreliak, 2019). Die Verwendung von Prozeduralität trägt dazu bei, IKE darzustellen und zu vermitteln, indem sie den Spieler:innen die Auswirkungen ihrer Handlungen in bestimmten sozialen und kulturellen Kontexten verdeutlicht (Majewski, 2019).

Kontext und Materialität

Materielles und immaterielles Kulturerbe sind untrennbar miteinander verbunden, und die Erhaltung eines Objekts ohne die Erhaltung des damit verbundenen IKE lässt es ohne Kontext für Interpretationen (Bonn et al., 2016). Der Kontext von IKE kann verschiedene Formen annehmen, beispielsweise bestimmte Räumlichkeiten, Landschaften, Zeiten, Kleidung, oder Werkzeuge. Computerspiele bieten die Möglichkeit, IKE in einem simulierten physischen Kontext zusammen mit dem dazugehörigen materiellen Erbe darzustellen.

Körperlichkeit und Körperbewegung

Es liegt in der Natur der Sache, dass IKE niemals statisch ist - es wird immer auf eine bestimmte Art und Weise *gemacht*; dementsprechend ist der menschliche Körper immer an seiner Herstellung beteiligt und ist für das Verständnis und die Darstellung von IKE unabdingbar (Wulf, 2015). Computerspiele können die Rolle des menschlichen Körpers reflektieren und darstellen: durch die Spieler:innen selbst und die Übersetzung von Körperbewegungen über das Eingabemedium, über den Player Character, und schließlich durch Non-Playable Characters, also nicht spielbare Charaktere, die sich bewegen und mit denen Spieler:innen teilweise auch interagieren können.

Fazit und Ausblick

Diese Überlegungen zeigen das Potenzial von Computerspielen für den Schutz von IKE. Darauf aufbauend gilt es nun herauszufinden, wie dies adäquat und in Übereinstimmung mit Kulturträger:innen umgesetzt werden kann oder bereits umgesetzt wird. Durch das Poster erhoffe ich mir weiterführende Diskussionen über die Präsentation von IKE in Spielen, über die kolonialistische Konnotation solcher „Archive“, sowie über alternative, dekoloniale Modelle der intraludischen Archivierung.

Das Poster eröffnet neue Perspektiven an der Schnittstelle von Cultural Heritage Studies, Game Studies und Digital Humanities. Innerhalb der Digital Humanities leistet das Poster einen Beitrag zu den Teilbereichen der „Digitized Humanities“ sowie der „Humanities of the Digital“ (Roth, 2019), da Computerspiele sowohl den Untersuchungsgegenstand als auch das Medium der Digitalisierung für IKE darstellen.

Fußnoten

1. Dazu zählen etwa mündlich überlieferte Traditionen, darstellende Künste oder traditionelles Handwerk (UNESCO 2003).

Bibliographie

Balela, Majed S., und Darren Mundy. 2015. „Analysing Cultural Heritage and its Representation in Video Games“. In *Proceedings of the 2015 DiGRA International Conference*. Lüneburg, Germany.

Bonn, Maria, Lori Kendall, und Jerome McDonough. 2016. „Preserving Intangible Heritage: Defining a Research Agenda“. *Proceedings of the Association for Information Science and Technology* 53 (1): 1–5. https://doi.org/10.1002/pra2.2016.14505301009.

Dimitropoulos, Kosmas, Filareti Tsalakanidou, Spiros Nikolopoulos, Ioannis Kompatsiaris, Nikos Grammalidis, Sotiris Manitsaris, Bruce Denby, u. a. 2018. „A Multimodal Approach for the Safeguarding and Transmission of Intangible Cultural Heritage: The Case of i-Treasures“. *IEEE Intelligent Systems* 33 (6): 3–16. https://doi.org/10.1109/MIS.2018.111144858.

Gesellschaft für Informatik e.V. 2014. „Die Grand Challenges der Informatik“.

Hawreliak, Jason. 2019. *Multimodal Semiotics and Rhetoric in Videogames*. Routledge Studies in Multimodality 8. S.l.: Routledge.

Hou, Yumeng, Sarah Kenderdine, Davide Picca, Mattia Egloff, und Alessandro Adamou. 2022. „Digitizing Intangible Cultural Heritage Embodied: State

of the Art“. *Journal on Computing and Cultural Heritage* 15 (3): 1–20. https://doi.org/10.1145/3494837.

Laiti, Outi, Sabine Harrer, Satu Uusiautti, und Annakaisa Kultima. 2021. „Sustaining Intangible Heritage through Video Game Storytelling: The Case of the Sami Game Jam“. *International Journal of Heritage Studies* 27 (3): 296–311. https://doi.org/10.1080/13527258.2020.1747103.

LaPensée, Elizabeth. 2021. „When Rivers Were Trails: Cultural Expression in an Indigenous Video Game“. *International Journal of Heritage Studies* 27 (3): 281–95. https://doi.org/10.1080/13527258.2020.1746919.

Majewski, Jakub. 2019. „Playing with intangible heritage: video game technology and procedural re-enactments“. In *Safeguarding Intangible Heritage: Practices and Politics*, herausgegeben von Natsuko Akagawa und Laurajane Smith. London; New York, NY: Routledge/Taylor and Francis Group.

Nylund, Niklas, Patrick Prax, und Olli Sotamaa. 2021. „Rethinking Game Heritage – towards Reflexivity in Game Preservation“. *International Journal of Heritage Studies* 27 (3): 268–80. https://doi.org/10.1080/13527258.2020.1752772.

Roth, Camille. 2019. „Digital, Digitized, and Numerical Humanities“. *Digital Scholarship in the Humanities* 34 (3): 616–32. https://doi.org/10.1093/llc/fqy057.

Ubisoft Montreal. 2021. „Assassin’s Creed Discovery Tour: Viking Age“. Montreal.

UNESCO. 2003. „Text of the Convention for the Safeguarding of the Intangible Cultural Heritage“. https://ich.unesco.org/en/convention.

Wulf, Christoph. 2015. „Performativity and Dynamics of Intangible Cultural Heritage“. In *Globalization, Culture, and Development*, herausgegeben von Christiaan De Beukelaer, Miikka Pyykkönen, und J. P. Singh, 132–46. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137397638_10.

Curation and Analysis of 'XVIIIe siècle: Bibliographie'

Schöch, Christof
schoech@uni-trier.de
Trier Center for Digital Humanities, Universität Trier, Deutschland
ORCID: 0000-0002-4557-2753

Introduction

Subject bibliographies support the creation of quantitative overviews (e.g. Luhmann and Burghardt, 2021) and qualitative surveys (e.g. Schöch et al., 2023) of a given field. In turn, bibliometrics, i.e. the statistical analysis of bibliographic datasets, provides insights into the history and characteristics of publication activities, but also into the methods and pitfalls of such analyses and the risks of relying on bibliometric indicators for performance assessment (e.g. van Raan, 2019). The value of open bibliographic data is also articulated clearly and programmatically (e.g. Peroni et al., 2015).

Motivated by this background, the present contribution describes one particular bibliography, the *XVIIIe siècle: Bibliographie*, its transformation into a structured dataset, and some results from an analysis of this dataset. The project website provides more information about the project and all resources mentioned below: https://christofs.github.io/BIB18/.

The poster aims to showcase the unique resource represented by the *XVIIIe siècle: Bibliographie*, to provide insight into strategies for the curation of bibliographic data (as well as some of the challenges involved in the process), and to highlight key findings about the publication habits of the research community whose work is documented by the bibliography.

The *XVIIIe siècle: Bibliographie*

The *XVIIIe siècle: Bibliographie* has been published, since 1992, and in over 550 instalments, by Canadian literary scholar Benoît Melançon. The bibliography is focused on scholarly publications about the Eighteenth century, both in French an in other languages. Members of the community can submit publications using an online form, making this a community-driven resource. This process, however, means there is no attempt at complete or systematic coverage and may also be a source of biases. In early 2023, Benoît Melançon published a complete set of all references for the years 1992-2022, containing about 64.400 entries, as a CSV file, along with a statement on his motivations for "liberating" this data (Melançon, 2023).

Transformation to BiBTeX for Zotero

The tabular format provided by Melançon was first transformed into BibTeX using a custom Python script. This BibTeX file was imported into Zotero, where the bibliography can now be consulted. Challenges in the transformation notably include the correct splitting of largely unstructured lists of author and editor names. As a consequence, manual improvement of the data is ongoing, but now considerably facilitated by Zotero’s interface.

Analysis of the Data

The bibliographic data in BiBTeX format was also exported, from Zotero, to several other formats, among them Zotero RDF. Using this XML-based format, several analyses have already been performed concerning e.g. the authors, publication types, languages, and patterns of collaboration that can be observed in the bibliography. These and more analyses are provided on a dedicated website, and only a summary of results is presented here.

For example, the four scholars most frequently mentioned as authors or editors of publications are Michel Porret, Jacques Berchtold, Michel Delon and Catriona Seth, all with over 300 publications over a period of 31 years. Also, the dataset is clearly dominated by French-language publications, which make up about 74% of the entries (Figure 1).

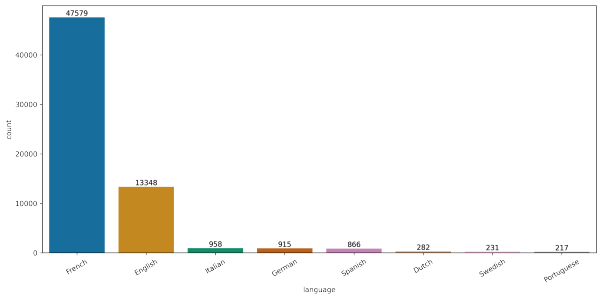


Figure 1: Distribution of languages in the *XVIIIe: Bibliographie* dataset.

With respect to collaboration patterns, it is notable that co-authorship is clearly an exception (only 6% of publications have more than one author), but that co-editorship is very widespread (60% of publications have two or three editors, while 32% have only one editor). Editors frequently publishing together have been identified and co-editing clusters determined using community detection in Gephi (Figure 2).

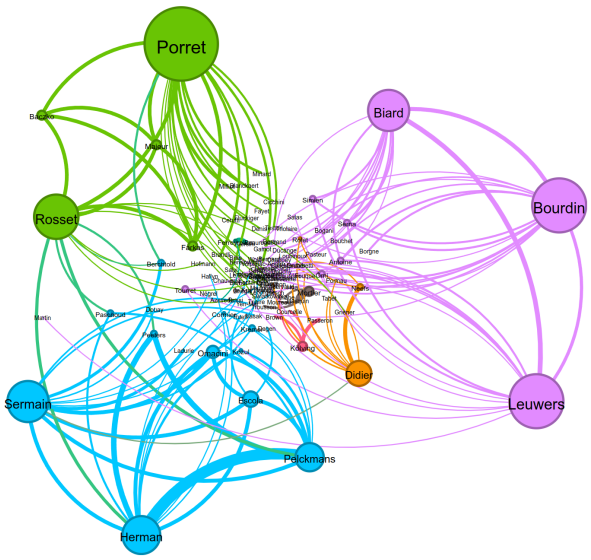


Figure 2: Co-editing network for the *XVIIIe siècle: Bibliographie* dataset, with communities shown by color.

Publication of the data and results

All data analyses have been performed using Python in Jupyter Notebooks. The scripts, results and explanatory prose have been published directly from the notebook files as a website using Quarto for rendering and Github pages as the publication platform.

The following tools were used in this project: Zotero 6.3, https://www.zotero.org/, Python 3.10, https://www.python.org/, Gephi 0.10.1, https://gephi.org/, Quarto 1.3, https://quarto.org/ and Github pages, https://pages.github.com/.

Future Work

Further corrections and refinements will be applied to the dataset on Zotero. When preparing data for analysis, the missing possibility to systematically and explicitly model part-whole relationships between contributions to edited volumes and the edited volumes themselves in Zotero requires attention, as it is a potential source of inflated numbers for editorship in the data. As for analyses, one area of future work concerns thematic trands and patterns, something that requires using external resources such as catalogs to scrape abstracts and/or keywords in order to enable analyses based on more than just the titles. Also, it would be of interest to identify potential biases in the dataset, for example with respect to author gender. Based on author gender identification (which, however, has its own problems; see e.g. Karimi et al. 2016), the distribution of author genders found in the Bibliographie could be compared to that in larger resources like the *Bibliography of the Modern Lan-*

guage Association, when filtering the latter for research on the Eighteenth Century.

Note on language

Because of the international audience of the *Bibliographie*, the website and this abstract have been written in English. The poster would, however, be produced in German for DHd2024.

Bibliographie

Luhmann, Jan, and Manuel Burghardt. 2021. "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 (2): 148–71. https://doi.org/10.1002/asi.24533.

Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. "Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods." In Proceedings of the 25th International Conference Companion on World Wide Web, 53–54. WWW'16. https://doi.org/10.1145/2872518.2889385.

Melançon, Benoît. 2023. "Libération des données, 21 février 2023." Site de Benoît Melançon / XVIIIe siècle: bibliographie (blog). 2023. http://mapageweb.umontreal.ca/melancon/donnees_biblios_1_550.html.

Peroni, Silvio, Alexander Dutton, Tanya Gray, and David Shotton. 2015. "Setting Our Bibliographic References Free: Towards Open Citation Data." Journal of Documentation 71 (2): 253–77. https://doi.org/10.1108/JD-12-2013-0166.

Raan, Anthony van. 2019. "Measuring Science: Basic Principles and Application of Advanced Bibliometrics." In Springer Handbook of Science and Technology Indicators, edited by Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, 237–80. Springer Handbooks. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_10.

Schöch, Christof, Julia Dudar, and Evgeniia Fileva, eds. 2023. Survey of Methods in Computational Literary Studies. CLS INFRA. https://doi.org/10.5281/zenodo.7892112.

Daidalos: Wie viel Methodenkompetenz braucht ein User?

Beyer, Andrea
beyeranz@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID: 0000-0002-8468-6411

Schulz, Konstantin
schulzkx@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID: 0000-0002-3261-9735

DH-Methodenkompetenzen in der Klassischen Philologie

In der deutschsprachigen Klassischen Philologie gibt es eine Kompetenzlücke zwischen analogen und digitalen Methoden der Forschung. Das mit den Methoden der Digital Humanities (DH) verknüpfte Verständnis von Text als Datensatz, d.h. als digitales Forschungsobjekt, ist zwar international durchaus etabliert, wenn u. a. literaturwissenschaftliche (z. B. zu Autorschaft Ochab & Essler, 2019) oder philologische Fragen (z. B. zu Textrekonstruktionen Assael et al., 2022) mit DH-Methoden untersucht werden. Doch spielen vor allem Methoden des Natural Language Processing (NLP) in der deutschen klassisch-philologischen Forschungscommunity nahezu keine Rolle. Dies stellt ein Infrastrukturprojekt wie Daidalos vor ein zentrales Problem: Wie erreichen und binden wir dauerhaft User, die trotz ihres Interesses an NLP-lastigen Forschungsfragen mangels eigener Kompetenz keine entsprechenden Forschungsfragen entwickeln können?

Das interdisziplinäre Daidalos-Projekt

Im DFG-Projekt Daidalos entwickeln wir eine Forschungsinfrastruktur, die vorhandene Tools, z. B. LatinCy (Burns, 2023), nachnutzt. Zugleich bietet sie dem Ansatz des Forschenden Lernens (Huber & Reinmann, 2019) folgend Lernmaterialien, wie User selbst Forschungsfragen entwickeln und die DH-Forschungsmethoden sinnvoll nutzen können. Sie soll Forschende der Klassischen Philologie sowie angrenzender Disziplinen befähigen,

1. ein spezifisches Textkorpus zusammenzustellen, zu analysieren und zu adaptieren,

2. die Analyseergebnisse zu visualisieren, zu speichern, zu teilen und zu exportieren,
3. über Ergebnisse und Methodik zu reflektieren, Forschungsfragen zu verfeinern oder aus neuen Perspektiven zu betrachten sowie
4. die eigene Digital Literacy auszubauen und über den NLP-Einsatz zu reflektieren.

Neben der Möglichkeit zur Nutzung eigener, bereits digitalisierter Texte sollen die User u.a. dazu befähigt werden, traditionelle literaturwissenschaftliche Fragestellungen mit Hilfe von maschineller Sprachverarbeitung zu verfolgen. Dazu kann beispielsweise die Erstellung von Goldstandardannotationen (Uzuner et al., 2010) gehören, die Nachverfolgung von literarischen Motiven über Text- und Gattungsgrenzen hinweg (Cordes, 2020) oder auch die Nutzung der Textsammlungen für lexikographische Zwecke (Heid et al., 2008). Data Sheets sollen dabei über relevante Metadaten und statistische Kennzahlen der kuratierten Textkorpora Auskunft geben. Die Zuordnung von Forschungsfragen zu passenden Methoden wird als Taxonomie modelliert und in einer Datenbank gespeichert. Sie ergibt sich aus den einzelnen Use Cases und bietet einen Fundus für methodische Empfehlungen bei neuen Forschungsfragen.

Um unterschiedliche Voraussetzungen und Lernausgangslagen der User berücksichtigen zu können, bietet die Software drei verschiedene Niveaustufen und Lernangebote, die von einer starken Lenkung durch Demo-Workflows über den detailliert erläuterten Einsatz von Jupyter Notebooks bis zu einer freien Konfiguration der Einstellungen sowie der Möglichkeit einer Methodentriangulation reichen werden.

Community of Practice und Forschungstandems

In Vorbereitung des DFG-Projektes haben wir bereits einen überregionalen Arbeitskreis Digital Classics aufgebaut, den wir zu einer Community of Practice (CoP) im Sinne des situierten Lernens (Lave & Wenger, 1991) erweitern und mit DH-Workshops unterstützen. Die Workshops werden durch Befragungen und Auswertung der Arbeitsergebnisse evaluiert. Auch sonst berücksichtigen wir bei der Gestaltung unserer CoP empirisch belegte Erfolgskriterien: „leadership roles, personalized learning, guiding principles, organizational support, social learning and purpose“ (Trust & Horrocks, 2019). Dazu haben wir Forschungstandems gebildet, wie sie u. a. im Design-Based Research Ansatz (Bakker, 2018) üblich und auch in unserer Fachcommunity (Freund & Janssen, 2020) bekannt sind. Diese Tandems bestehen aus erfahrenen und eher unerfahrenen Usern von DH-Methoden, die gemeinsam kleine Forschungsfragen entwickeln, mithilfe „digitalgestützter“ Methoden bearbeiten und eine abschließende Open-Access-Publikation in einer Zeitschrift mit Peer Review veröffentlichen (vgl. Abbildung 1). Zugleich dienen diese Arbeiten als authentische user stories (Wautelet et al., 2014), die für die Entwicklung der Infrastruktur (Bedienbarkeit und Funktionalität) genutzt und als kuratierte Lernmaterialien inkl. Nutzungsstatistiken aufbereitet werden. Außerdem wollen wir auf der Basis der Arbeit mit den Forschungstandems eine fachspezifische Digital Literacy modellieren und diese als Teil zukunftsfähiger Forschungskompetenz mithilfe der CoP in der deutschsprachigen Fachcommunity nicht nur in die Forschung, sondern auch in die Lehre einführen. Somit bilden die Forschungstandems den Kern unseres Daidalos-Projektes, da die Forschenden ihrerseits als Multiplikatoren tätig werden sollen und wir dadurch auf eine Akzeptanz unserer Forschungsinfrastruktur in der Breite hinwirken wollen.

Diese Akzeptanz ist umso wichtiger, als die verschiedenen Blickwinkel auf antike Texte in der Klassischen Philologie, Linguistik und Informatik zu konzeptuellen Herausforderungen führen. Dies betrifft etwa die Frage nach Distant und Close Reading, nach Open und Closed Access für Publikationen sowie nach eher individueller oder eher kollaborativer Arbeitskultur. Um die erfolgreiche Überwindung dieser Hürden zu dokumentieren, soll die Infrastruktur durch qualitative Nutzungsstudien evaluiert werden.

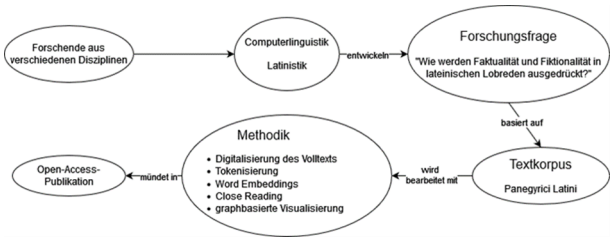


Abbildung 1: Exemplarisches Forschungstandem, dessen Arbeit mit der Daidalos-Infrastruktur ermöglicht und unterstützt werden soll.

Bibliographie

Assael, Yannis, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag und Nando de Freitas. 2022. “Restoring and attributing ancient texts using deep neural networks.” *Nature* 603: 280–283.

Burns, Patrick J. 2023. “ *LatinCy: Synthetic Trained Pipelines for Latin NLP*.” https://arxiv.org/abs/2305.04365v1, 2023.

Bakker, Arthur. 2019. “ *Design Research in Education: A Practical Guide for Early Career Researchers.*” New York: Routledge.

Cordes, Lisa. 2020. “Wenn Fiktionen Fakten schaffen. Faktuales und fiktionales Erzählen in den spätantiken Panegyrici Latini.” In D. Breitenwischer, H.-M. Häger, & J. Menninger (Hgg.). *Faktuales und fiktionales Erzählen II. Geschichte – Medien – Praktiken*: 31–56. Würzburg: Ergon.

Eshet-Alkalai, Yoram. 2004. “Digital Literacy: A Conceptual Framework for Survival Skills in the Digital era.” *Journal of Educational Multimedia and Hypermedia* 13(1): 93-106.

Freund, Stefan und Leoni Janssen. 2020. „Forschendes Lernen im Praxissemester unter den Bedingungen kleiner Fächer: Ein Praxiskonzept für die Begleitung von Studienprojekten im Praxissemester am Beispiel des Faches Latein.” *Die Materialwerkstatt. Zeitschrift für Konzepte und Arbeitsmaterialien für Lehrer*innenbildung und Unterricht* 2(2): 66–74.

Heid, Ulrich, Fabienne Fritzinger, Susanne Hauptmann, Julia Weidenkaff, und Marion Weller. 2008. “Providing corpus data for a dictionary for German juridical phraseology.” *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing, KONVENS*: 131–144.

Huber, Ludwig und Gabi Reinmann. 2019. „ *Vom forschungsnahen zum forschenden Lernen an Hochschulen: Wege der Bildung durch Wissenschaft*.“ Wiesbaden: Springer.

Lave, Jean und Etienne Wenger. 1991. “ *Situated Learning: Legitimate Peripheral Participation*.” Cambridge: Cambridge University Press.

Ochab, Jeremi K. und Holger Essler. 2019. „Stylometry of literary papyri.“ *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*: 139–142.

Trust, Torrey und Brian Horrocks. 2019. “Six Key Elements Identified in an Active and Thriving Blended Community of Practice.” *TechTrends* 63: 108–115.

Uzuner, Ozlem, Imre Solti, Fei Xia und Eithon Cadag. 2010. “Community annotation experiment for ground truth generation for the i2b2 medication challenge.” *Journal of the American Medical Informatics Association* 17(5): 519–523.

Wautelet, Yves, Samedi Heng, Manuel Kolp und Isabelle Mirbel. 2014. “Unifying and extending user story models.” *International Conference on Advanced Information Systems Engineering*: 211–225.

Das kleine Wörterbuch der Redeeinleiter

Brunner, Annelen

brunner@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Tu, Ngoc Duyen Tanja

tu@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Weimer, Lukas

weimer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0001-6919-3646

Das Poster[1] präsentiert eine Online-Ressource, die es erlaubt, 523 Redeeinleiter dynamisch zu durchsuchen. Sie bietet damit niedrigschwelligen Zugang zu empirischen Daten aus einem umfangreich manuell annotierten Korpus von fiktionalen und nicht-fiktionalen Texten, die zwischen 1850 und 1919 erschienen sind. Durch die Einbindung in eine große Forschungsinfrastruktur wird Zugänglichkeit garantiert und Nachnutzung vereinfacht.

Gegenstand

Redeeinleiter sind sprachliche Ausdrücke, die relativ zu einer direkten oder indirekten Rede- oder Gedankenwiedergabe in Voran-, Mittel- oder Nachstellung stehen und diese einleiten (Breslauer, 1996; Michel, 1966; Jäger, 1968). Auch wenn Verben wie *sagen* und *denken* sehr häufig sind, können Redeeinleiter auch kreative Ausdrücke sein und unterschiedliche Wortarten umfassen, z.B.

(1) *Der Untersuchungsrichter* ***legte*** *ihm* ***ans Herz*** *, daß, wenn er nicht angeben wolle, wo er den Einbruch verübt habe, sein Kopf sich schon als abgetan betrachten könne.* [aus: Hermann Sudermann: Miks Bumbullis (1917)]

(2) *»Kratzt der Alte einmal wieder«,* ***brummte*** *er, »und stört die ordentlichen Leute in ihrer Nachtruhe.«* [aus: Franz Grillparzer: Der arme Spielmann (1840)]

(3) […] *daß ihnen aber nicht einmal immer das geliefert wird, was sie zu fordern ein Recht haben, bezeugen häufige* ***Klagen*** *von Auswanderern* […] . [aus: Deutsche Auswanderer-Zeitung, Unbekannte:r Autor:in (1852)]

Durch ihre Vielfältigkeit sind Redeeinleiter ein interessanter Untersuchungsgegenstand, sowohl aus linguistischer als auch literaturwissenschaftlicher Perspektive. Sie eignen sich, um dynamische Prozesse im lexikalischen Inventar der Sprache zu untersuchen (Tu, erscheint 2024) und spielen eine wichtige Rolle bei der Einbettung von Figurenrede in den narrativen Kontext (McHale 2014, Abschnitt 2).

Datengrundlage und Funktionalitäten

Grundlage für die Ressource ist das Kernkorpus „Redewiedergabe“ (RW-Korpus; Brunner et al., 2020a), welches im Rahmen eines DFG-Projekts entstand. Das Korpus umfasst ca. 49.000 Tokens und enthält Ausschnitte aus Erzählungen sowie Zeitungs- und Zeitschriftenartikeln aus dem Zeitraum 1850-1919. Das Textmaterial ist balanciert nach Dekaden sowie dem Merkmal fiktional vs. nicht-fiktional

und wurde aufwendig manuell nach Formen von Rede-, Gedanken- und Schriftwiedergabe annotiert (Brunner et al., 2020b): Eine Konsensannotation wurde auf Basis zweier unabhängiger Annotationen erstellt. Zwar ist das RW-Korpus sowohl in TEI-konformem XML-Format als auch in einem spaltenbasierten Textformat vollständig frei verfügbar(https://zenodo.org/records/3739239), es bedarf jedoch technischer Kenntnisse, spezialisierte Informationen wie die über Redeeinleiter zu extrahieren. Die vorgestellte Ressource bietet hierfür einen bequemen und niedrigschwelligen Zugang.

Alle 3059 Einleiter-Vorkommen von direkter und indirekter Rede- oder Gedankenwiedergabe wurden mit ihren Attributen extrahiert und zu einer Häufigkeitsliste zusammengefasst. Für jeden der 523 Redeeinleiter-Typen bietet die Ressource einen Überblick über die Vorkommensverteilung nach den Dimensionen „Medium“ (Rede- oder Gedankenwiedergabe), „Wiedergabetyp“ (direkt oder indirekt), „Position“ (initial, medial oder final) und „Textsorte“ (fiktional oder nicht-fiktional). Abbildung 1 illustriert dies für den Einleiter *sagen*. Abbildung 2 gibt einen Überblick über die Attributverteilung des gesamten Inventars.

Abbildung 1: Vorkommensverteilung für den Redeeinleiter *sagen*.

Medium	Redewiedergabe:	535
	Gedankenwiedergabe:	1
Wiedergabetyp	direkt:	398
	indirekt:	138
Position	initial:	291
	medial:	123
	final:	122
Textsorte	fiktional:	435
	nicht-fiktional:	101

Für jede mögliche Attributkombination, mit der ein Redeeinleiter vorkommt, wurde ein zufällig gewählter Korpusbeleg extrahiert. Dieser besteht aus dem Satz, in dem der Redeeinleiter vorkommt, sowie dem vorangehenden und dem nachfolgenden Satz. Eine Verlinkung verknüpft ihn mit dem entsprechenden Dokument im RW-Korpus. Die Liste der Redeeinleiter kann zudem nach Attributwerten gefiltert werden und Redeeinleiter sowie die zugehörigen Belege können als Excel- oder CSV-Tabellen exportiert werden.

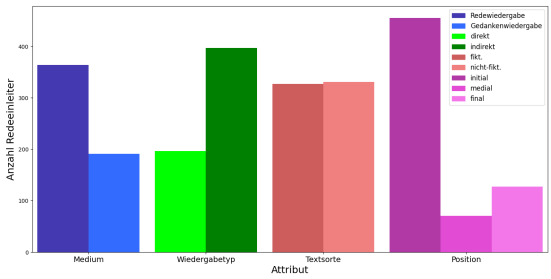


Abbildung 2: Verteilung des gesamten Einleiter-Inventars über die Attribute; gezählt wurden Redeeinleiter-Typen, wobei solche, für die mehrere Merkmale zutreffen, mehrfach gezählt wurden (z.B. bei Vorkommen in allen drei Positionen Zählung für jede der Positionen einmal).

Mit Hilfe der Ressource bekommt man nicht nur einen grundsätzlichen Überblick darüber, welche Einleiter in welchen quantitativen Verhältnissen verwendet werden, sondern sie erlaubt auch Kontrastierungen in unterschiedlichen Dimensionen. Die Filterfunktion ermöglicht zudem, gezielt solche Einleiter anzuzeigen, die in mehreren Kontexten vorkommen können (z.B. als Rede- und als Gedankeneinleitung). Neben den interessanten Beobachtungen, die die Ressource zu dem Datenbestand des RW-Korpus selbst erlaubt, bietet es sich an, diese mit Daten aus anderen Textgrundlagen zu vergleichen (z.B. moderne Literatur).

Zugänglichkeit und Nachhaltigkeit

Das „Kleine Wörterbuch der Redeeinleiter“ ist unter der Adresse https://www.owid.de/plus/redeeinleiter über OWIDplus abrufbar. OWIDplus ist eine Plattform für multilinguale lexikalisch-lexikografische Daten, für quantitative lexikalische Auswertungen und für interaktive lexikalische Anwendungen und zudem Teil des Portfolios, welches das Leibniz-Institut für Deutsche Sprache in das NFDI-Konsortium Text+ einbringt. Als Teil des breiten Tool-, Kompetenz- und Datenportfolios von Text+ erhöhen sich Auffindbarkeit und Verbreitung (z. B. durch die Text+ Registry, Genêt et al., 2023) der Ressource, wodurch es einer breiteren Wissenschaftscommunity entlang der FAIR-Kriterien (Wilkinson et al., 2016) zur Verwendung und Nachnutzung bereitgestellt werden kann. Gleichzeitig sind durch die institutionelle Anbindung an das IDS als zertifiziertes Datenzentrum nachhaltige Zugänglichkeit und Archivierung gesichert.

Fußnoten

1. Contributor Roles: Annelen Brunner (Data Curation, Writing – original draft, Writing - review & editing), Ngoc Duyen Tanja Tu (Data Curation, Visualization, Software, Writing – original draft, Writing - review & editing), Lukas Weimer (Data curation, Writing - review & editing)

Bibliographie

Breslauer, Christine. 1996. "Formen der Redewiedergabe im Deutschen und Italienischen." *Sammlung Groos* 60. Heidelberg: Groos.

Brunner, Annelen, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu und Lukas Weimer. 2020a. "Corpus REDEWIEDERGABE". In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 803-12.

Brunner, Annelen, Lukas Weimer, Stefan Engelberg, Fotis Jannidis und Ngoc Duyen Tanja Tu. 2020b. "Annotationsrichtlinien des Projekts ‚Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse'". https://zenodo.org/record/3547594.

Genêt, Philippe, Tobias Gradl, Kilian Hensen, Christoph Kudella und Daniela Schulz. 2023. "F wie Registry. Die Text+ Registry als Hilfsmittel zur Auffindbarkeit von Ressourcen". In *FORGE23: Forschungsdaten in den Geisteswissenschaften – kritisch betrachtet*, 168-172.

Jäger, Siegfried. 1968. " *Die Einleitungen indirekter Reden in der Zeitungssprache und in anderen Texten der deutschen Gegenwartssprache. Ein Diskussionsbeitrag*". *Muttersprache* 78: 236-49.

McHale, Brian. 2014. "Speech Representation". In *The living handbook of narratology*, hg. Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert, 812-824. Hamburg: Hamburg University Press. http://hup.sub.uni-hamburg.de/lhn/index.php?title=Speech Representation&oldid=891.

Michel, Georg. 1966. " *Sprachliche Bedingungen der Wortwahl. Eine Untersuchung an Ausdrücken der Redeeinführung (Erster Teil)*". *Zeitschrift für Phonetik* 19: 103-129.

Tu, Ngoc Duyen Tanja. Erscheint 2024. "Eine korpuslinguistische Untersuchung zur lexikalischen Vielfalt von direkten und indirekten Redeeinleitern." *Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache*.

Deep Mapping der 'Merkwürdigkeiten' – Sightseeing im frühneuzeitlichen Wien

Rastinger, Nina C.
ninaclaudia.rastinger@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage (Österreichische Akademie der Wissenschaften), Österreich

Salzburger, Stefanie
stefanie.salzburger@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage (Österreichische Akademie der Wissenschaften), Österreich

Sogenannte ‚Sehenswürdigkeiten' spielen für heutige (Haupt-)Städte eine zentrale Rolle: Sowohl in Online-Formaten als auch in gedruckter Form werden Rankings sehenswerter Gegenstände, Bauten und Plätze erstellt, Routen für deren effiziente Besichtigung entworfen und festgeschrieben, was als ‚Must-See' unumgänglich gilt (vgl. z.B. Geiss, 2018; Eickhoff, 2021; Faulkner und Faulkner, 2023). Im Falle Wiens finden sich derartige subjektive Klassifizierungen spezifischer Orte bereits in der Frühen Neuzeit, nämlich in Form früher Reisehandbücher, die sowohl ausländischen Gästen als auch ansässigen Bewohner*innen ‚Merkwürdigkeiten' der damaligen Stadt empfahlen – also Orte, denen Raum im kulturellen Gedächtnis zugesprochen wurde. So verfasste beispielsweise Franz de Paula Gaheis eine „Beschreibung der auffallendsten Merkwürdigkeiten der Haupt- und Residenzstadt Wien" (1793, https://data.onb.ac.at/rep/10A11BA9), Joseph Edler von Kurzböck eine „Beschreibung aller Merkwürdigkeiten Wiens: Ein Handbuch f. Fremde u. Inländer" (1779) oder Jean Thedor Gontier de Faifve einen „Almanach von Wien zum Dienste der Fremden, oder historischer Begriff der anmerkungswürdigsten Gegenstände dieser Hauptstadt" (1774, https://resolver.obvsg.at/urn:nbn:at:AT-WBR-8423):

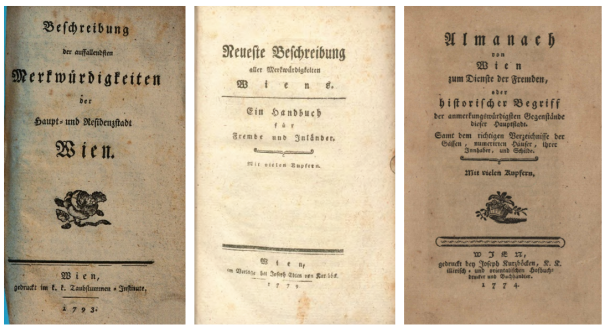


Abb.1: Titelseiten exemplarischer Reisehandbücher für das Wien des 18. Jahrhunderts

Mit ebensolchen historischen Wahrnehmungen Wiens setzt sich das im Poster vorgestellte, von der Stadt Wien geförderte Projekt „Wiens ‚Merkwürdigkeiten' – Deep Mapping frühneuzeitlicher Reisehandbücher" (2023–2024, PI: Nina C. Rastinger) auseinander. Hierfür werden Reisehandbücher des 18. und frühen 19. Jahrhunderts volltextdigitalisiert, annotiert und analysiert, um eine evidenzbasierte Deep Map (vgl. Bloom und Sacramento, 2017; Bodenhamer et al., 2021) der historischen Stadt zu erarbeiten und subjektive Ortswahrnehmungen sicht- und quantifizierbar zu machen. So stellen sich beispielsweise die Fragen, welche Orte Wiens zwischen 1700 und 1850 als ‚denk-' bzw.

‚merkwürdig' erachtet wurden und in welchen Stadtteilen sich Häufungen derartiger ‚Merkwürdigkeiten' finden. Außerdem gilt es, das Textmaterial auf synchrone und diachrone Muster zu untersuchen, indem nach Unterschieden und Parallelen zwischen diversen Reisehandbüchern und Zeitpunkten gefragt wird: Empfanden beispielsweise in Wien ansässige Autor*innen andere Orte als erinnerungswürdig als ausländische Verfasser*innen oder verschwanden gewisse (Arten von) Plätze(n) im Laufe der Zeit aus der untersuchten Textsorte, während andere neu in den Blick rückten?

Um Antworten auf diese und weitere Forschungsfragen zu finden, wird als erster Schritt, auf Grundlage einer umfänglichen Sichtung des überlieferten Textmaterials (vgl. https://www.zotero.org/groups/5128433), ein Analysekorpus von ungefähr zehn ausgewählten Reisehandbüchern erstellt. Jene Werke, für die bislang kein (verlässlicher) Volltext vorliegt, können hierfür mithilfe der Transkriptionsplattform Transkribus (https://readcoop.eu/de/transkribus) und des öffentlich verfügbaren HTR-Modells „German Fraktur 18th Century – WrDiarium_M9" (Resch und Kampkaspar, 2020) (semi-)automatisch transkribiert werden. Überdies erlaubt Transkribus seit Kurzem das Training von auf Textregionen ausgerichteten Layoutmodellen, wodurch auf der Strukturebene (z.B. über Zentrierung oder Sperrung) kenntlich gemachte ‚Merkwürdigkeiten' Wiens ebenfalls automatisch erkannt werden können. Abhängig von dem Aufbau der einzelnen Reisehandbücher setzt das Projektteam zudem zusätzlich auf (semi-)automatische Ansätze zur Named Entity Recognition sowie auf manuelle Annotationsprozesse, zum Beispiel über die Annotationsumgebung CATMA (Gius et al., 2023).

Im nächsten Schritt können die gesammelten Toponyme daraufhin geokodiert, d.h. auf Koordinaten zurückgeführt, und in das Open-Source-Geoinformationssystem QGIS (https://www.qgis.org) überführt werden. Auf diese Weise wird es möglich, die analysierten Reisehandbücher auf historischen Plänen Wiens (vgl. Opll 2004), wie dem bereits von der Stadt Wien georeferenzierten Steinhausenplan (1710), Nagelplan (1781) oder Behselplan (1825), abzubilden und die subjektive ‚Merkwürdigkeit' verschiedener Stadtteile – beispielsweise in Form von Heatmaps – zu visualisieren:

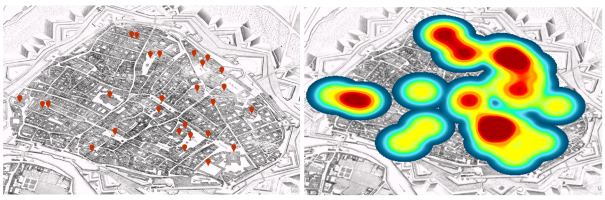


Abb. 2: Exemplarische Visualisierungen der ‚Merkwürdigkeiten' Wiens nach Franz Paula de Gaheis (1793) auf dem Nagelplan (1781)

Kombiniert bilden diese einzelnen Kartendarstellungen eine Deep Map des frühneuzeitlichen Wiens, innerhalb derer unterschiedliche Wahrnehmungen der Stadt miteinander kombiniert sowie dreifach semantisch angereichert werden können – nämlich einerseits über (1) Ortsbeschreibungen und (2) Bildmaterial aus den untersuchten Reisehandbüchern selbst und andererseits über (3) Verlinkungen zu weiterführenden Informationen (z.B. Einträgen der Wissensplattform „Wien Geschichte Wiki", www.geschichtewiki.wien.gv.at). Durch dieses multimodale und -dimensionale ‚Mapping der Merkwürdigkeiten' soll schlussendlich ein tieferes Verständnis des frühneuzeitlichen Wiens erreicht und ein weiterer Baustein zu einer projektübergreifenden „Vienna Time Machine"[1] beigetragen werden – denn wie Bloom und Sacramento (2017: 6) festhalten: "To get an understanding of an actual place, one must inhabit its multiple overlapping contradictory stories simultaneously."

Fußnoten

1. Vgl. die EU-Initiative „Time Machine" (https://www.timemachine.eu). Vorangehende Schritte in die Richtung einer lokalen „Vienna Time Machine" wurden u.a. bereits durch die von der Stadt Wien finanzierten Projekte „Vienna Time Machine: Corresponding digital data treasures and knowledge resources" (PI: Claudia Resch, 2020–2022, vgl. Resch, Rastinger, Kirchmair 2022) und „Visiting Vienna – digital approaches to the (semi-)automatic analysis of the arrival lists found in the *Wien[n]erisches Diarium* " (PI: Nina C. Rastinger, 2022–2023, vgl. Rastinger 2023) geleistet.

Bibliographie

Bloom, Brett und Nuno Sacramento. 2017. *Deep Mapping*. Auburn: BKDN BKDN Press.

Bodenhamer, David J., John Corrigan und Trevor M. Harris, Hrsg. 2021. *Making Deep Maps. Foundations, Approaches, and Methods*. Bloomington: Indiana University Press.

Eickhoff, Peter. 2021. *111 Orte in Wien, die man gesehen haben muss*. Köln: Emons.

Faulkner, Jennifer und Rosemary Faulkner. 2023. *Zu Fuß durch Wien. 12 Spaziergänge*. 2. Aufl. Düsseldorf: Droste.

Geiss, Heide Marie Karin. 2018. *Streifzüge Wien. Die besten Wege die Stadt und ihre Highlights zu erleben*. München: National Geographic Deutschland.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Mareike Schumacher und Dominik Gerstorfer. 2023. CATMA 7 (Version 7.0). 10.5281/zenodo.1470118

Opll, Ferdinand. 2004. *Wien im Bild historischer Karten. Die Entwicklung der Stadt bis in die Mitte des 19. Jahrhunderts*. Wien/Köln/Weimar: Böhlau.

Rastinger, Nina C. 2023. Visiting Vienna – digital approaches to the (semi-)automatic analysis and mapping of the arrival lists found in the *Wien[n]erisches Diarium*.

ADHO Digital Humanities Conference 2023 (DH2023), Graz. 10.5281/zenodo.8147909

Resch, Claudia und Dario Kampkaspar. 2020. *HTR-Modell „German Fraktur 18th Century – WrDiarium_M9“*. https://readcoop.eu/de/modelle/german-fraktur-18th-century (zugegriffen: 15. November 2023).

Resch, Claudia, Nina C. Rastinger und Thomas Kirchmair. 2022. Die historische *Wiener Zeitung* und ihre Sterbelisten als Fundament einer Vienna Time Machine. Digitale Ansätze zur automatischen Identifikation von Toponymen. *Wiener Digitale Revue* 4. 10.25365/wdr-04-03-04

Der radioaktive Spiegel

Schmitz, Jascha Merijn
smitzjak@hu-berlin.de
Humboldt-Universität Berlin, Deutschland

Schumacher, Mareike
mareike.schumacher@ur.de
Universität Regensburg, Deutschland

Geiger, Jonathan D.
jonathan.geiger@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0000-0002-0452-7075

Digital Humanities, Wissenschaftskommunikation und Podcasts

Die Jahreskonferenz der Digital Humanities (DH) im deutschsprachigen Raum, die DHd2024, fordert mit ihrem Motto "DH Quo Vadis?" zu einem kurzen Innehalten auf: Was hat sich getan in der deutschsprachigen DH-Landschaft der letzten 10 Jahre? Wo stehen wir gerade? Welche Prognosen gibt es für die Zukunft? Wir vom Wissenschaftspodcast RaDiHum20 stimmen in dieses Innehalten ein und werfen einen Blick auf einen der Trends in den DH: die Wissenschaftskommunikation.

In der Wissenschaftskommunikation haben sich, wie in allen Bereichen der DH im letzten Jahrzehnt, neue Methoden etabliert. Gerade durch den eingeschränkten physischen Radius, das Bedürfnis nach persönlichem Austausch aber auch nach wissenschaftlich fundierter Information während der Corona-Pandemie wurde der Wert alternativer Kommunikations- und Informationskanäle sowohl innerhalb der Community als auch in der wissenschaftlichen Außenkommunikation deutlich (vgl. Katzenberger, Keil und Wild, 2022). Neben anderen Formaten (allen voran Blogs und Social Media, vgl. Prinz, 2018) entwickelte sich Podcasting zu einer der digitalen Säulen nicht nur der science-to-public sondern auch der science-to-science-Kommunikation (vgl. Frick & Seltmann, 2023; Leander, 2022).

Das Radio Digital Humanities

RaDiHum20 ("Radio Digital Humanities", die "20" verweist auf das Gründungsjahr und das monatliche Erscheinungsdatum) ist als Wissenschaftspodcast speziell für die DH konzipiert. Er wurde in der ersten Jahreshälfte 2020 geboren und beteiligt waren und sind Jonathan D. Geiger, Lisa Kolodzie und Mareike Schumacher; Patrick Toschka verließ das Hostteam 2022, Nachfolger wurde Jascha Schmitz. Thematisch orientierte sich RaDiHum20 von Anfang an stark an den deutschsprachigen DH und suchte engen Kontakt mit dem DHd-Verband. Mittlerweile sind insgesamt 70 Folgen in vier Staffeln entstanden. Der Twitteraccount umfasst derzeit 738 Follower, der jüngere Mastodonaccount 167 Follower. Seit Einführung des Podcasts am 01.07.2020 wurden die Folgen insgesamt 13.329 Mal angehört. Im Durchschnitt wurde jede Folge 190 Mal angehört. Rund 100 Abonnent*innen folgen RaDiHum20 über den Hosting-Service Podcaster.de, weitere 118 Follower*innen verzeichnet Spotify und 56 hören regelmäßig über Apple Podcasts zu.

In unserem Posterbeitrag wird der DH-Podcast RaDiHum20, seine Hintergründe und Progammvielfalt ausführlich vorgestellt. Zudem wird der Versuch unternommen, den Podcast als Werkzeug für die DH-Forschung selbst einzusetzen.

Der radioaktive Spiegel – Wissenschaftspodcasts als Indikatoren für die Forschung

Die Erforschung von Podcasts erfuhr in den letzten Jahren einen massiven Aufschwung (vgl. beispielsweise Katzenberger, Keil und Wild (2022), MacKenzie (2019) oder Llinares, Fox und Berry (2018)). Doch Podcasts können nicht nur als Forschungsgegenstände (im Hinblick auf wissenschaftliche Kommunikation (vgl. Bonfadelli et al. (2017)), als Transformator für das wissenschaftliche Arbeiten (vgl. Howard-Sukhil, Wallace und Chakrabarti (2021)) oder als ein populäres Phänomen gedacht werden, sondern auch als Indikatoren für die metawissenschaftliche Forschung selbst. Hierzu liegen allerdings noch kaum Forschungsansätze (z. B. aus der Trendforschung) vor.

Durch seine enge Anbindung an den DHd-Verband und das Format der Interviews ist auch der RaDiHum20-Podcast nicht nur ein Wissenschaftskommunikationskanal, sondern eine Art Spiegel für die DH. Die einzelnen Aufnahmen lassen sich als Oral History begreifen, da sie den jeweils aktuellen Stand der DH auf vielen Ebenen und aus vielen Perspektiven widerspiegeln. Aus diesen Quel-

len lassen sich außerdem entsprechende Prognosen für die deutschsprachigen DH herauslesen, die sich auch auf das Konferenzthema der DHd2024 beziehen lassen.

Die Frage "DH Quo Vadis?" wurde von Beginn unseres Podcasts an immer wieder in Interviews gestellt. Mal geht es um Zukunftspläne bestimmter Arbeitsgruppen, z. B. darum, mehr Einstiegshilfen für DH-Neulinge zu schaffen (vgl. Geiger et al., 20.9.2020), alternative Publikationsformen zu entwickeln (vgl. Geiger et al., 20.10.2020) oder die Rolle von Research Software Engineers zu stärken (vgl. Geiger et al., 20.11.2020). Besonders relevant waren Fragen zur Zukunft der DH auch in den Folgen, die wir dem Thema des Studiengangsmanagements gewidmet haben. Hier gibt es einerseits den Wunsch nach konstanterer Koordination der einzelnen Studiengänge (vgl. Geiger et al., 20.12.2022) und auf der anderen Seite die These, dass es in absehbarer Zukunft gar keine DH-Studiengänge mehr bräuchte (vgl. Geiger et al., 20.01.2023). Auch beim Thema Community-Management kommen Zukunftsvisionen zur Sprache. Hier sind sich all unsere Interviewpartner*innen einig: Die DH-Community bringt ständig neue, eigene Ideen ein, sodass weder in der Vergangenheit noch derzeit viel aktives Community-Management betrieben werden muss (vgl. Geiger et al., 20.5.2022 und 20.6.2022). Zudem stellen die Gründungen und Entwicklungen der Arbeitsgruppen im DHd-Verband einen Marker für die aktuellen und dauerhaften Themen in der deutschsprachigen DH-Szene dar.

In unserem Poster stellen wir daher nicht nur den DH-Podcast RaDiHum20 vor, sondern extrahieren auch allgemeine gegenwärtige Trends in den DH aus den von uns geführten Interviews. Dem unterliegt die These, dass Wissenschaftspodcasts wie RaDiHum20 den DH nicht nur eine Kommunikationsplattform im Jetzt bieten, sondern auch als ein Mittel betrachtet werden können, sowohl die vergangenen als auch zukünftigen Trends der Community widerzuspiegeln und Antworten auf die im Call aufgeworfene Frage "DH Quo vadis?" zu wagen.

Bibliographie

Bonfadelli, Heinz et al. (Hrsg.). 2017. "Forschungsfeld Wissenschaftskommunikation." Springer: Wiesbaden.

Frick, Claudia und Melanie E.-H. Seltmann. 2023. "Referenzrahmen für eigenständige digitale Wissenschaftskommunikation durch Forschende." In Zeitschrift für digitale Geisteswissenschaften / Working Papers 3. Wolfenbüttel. DOI: 10.17175/wp_2023b.

Geiger, Jonathan D., Lisa Kolodzie, Mareike Schumacher und Patrick Toschka. "RaDiHum20 spricht mit der AG Digital Humanities Theorie." RaDiHum20 – Das Radio Digital Humanities. 20. August 2020. Podcast, MP3 audio, 33:03. https://radihum20.de/ag-digital-humanities-theorie/.

Geiger, Jonathan D., Lisa Kolodzie, Mareike Schumacher und Patrick Toschka. "RaDiHum20 spricht mit der AG Graphentechnologien." RaDiHum20 – Das Radio Digital Humanities. 20.9.2020. Podcast, MP3 audio, 49:22. https://radihum20.de/graphentechnologien/.

Geiger, Jonathan D., Lisa Kolodzie, Mareike Schumacher und Patrick Toschka. "RaDiHum20 spricht mit der AG Digitales Publizieren." RaDiHum20 – Das Radio Digital Humanities. 20.10.2020. Podcast, MP3 audio, 35:58. https://radihum20.de/ag-digitales-publizieren/.

Geiger, Jonathan D., Lisa Kolodzie, Mareike Schumacher und Patrick Toschka. "RaDiHum20 spricht mit der AG Research Software Engineering." RaDiHum20 – Das Radio Digital Humanities. 20.11.2020. Podcast, MP3 audio, 45:51. https://radihum20.de/research-software-engineering/.

Geiger, Jonathan D., Lisa Kolodzie, Jascha Schmitz und Mareike Schumacher. "RaDiHum20 spricht mit Melanie Seltmann und Torsten Roeder über Community-Management bei DHall." RaDiHum20 – Das Radio Digital Humanities. 20.5.2022. Podcast, MP3 audio, 27:39. https://radihum20.de/radihum20-community-management-bei-dhall/.

Geiger, Jonathan D., Lisa Kolodzie, Jascha Schmitz und Mareike Schumacher. "RaDiHum20 spricht mit dem DHd-Vorstand über Community-Management." RaDiHum20 – Das Radio Digital Humanities. 20.6.2022. Podcast, MP3 audio, 50:55. https://radihum20.de/community-management-dhd-vorstand/

Geiger, Jonathan D., Lisa Kolodzie, Jascha Schmitz und Mareike Schumacher. "RaDiHum20 spricht mit Jacqueline Klusik-Eckert über Studiengangsmanagement." RaDiHum20 – Das Radio Digital Humanities. 20.12.2022. Podcast, MP3 audio, 38:59. https://radihum20.de/radihum20-spricht-mit-jacqueline-klusik-eckert-ueber-studiengangsmanagement/

Geiger, Jonathan D., Lisa Kolodzie, Jascha Schmitz und Mareike Schumacher. "RaDiHum20 spricht mit Peter Niedermüller über den Studiengang in Mainz und dessen Studiengangsmanagement." RaDiHum20 – Das Radio Digital Humanities. 20.1.2023. Podcast, MP3 audio, 23:24. https://radihum20.de/radihum20-studiengangsmanagement2/

Howard-Sukhil, Christian, Samantha Wallace und Ankita Chakrabarti. 2021. "Developing Research through Podcasts: Circulating Spaces, A Case Study." In digital humanities quarterly 15: 3. http://www.digitalhumanities.org/dhq/vol/15/3/000554/000554.html.

Katzenberger, Vera, Jana Keil und Michael Wild. 2022. "Podcasts. Perspektiven und Potenziale eines digitalen Mediums." Springer: Wiesbaden.

Leander, Lisa. 2020. "Wissenschaft im Gespräch: Wissensvermittlung und -aushandlung in Podcasts." In kommunikation@gesellschaft 21: 2, 1–24. DOI: https://doi.org/10.15460/kommges.2020.

Llinares, Dario, Neil Fox und Richard Berry (Hrsg.). 2018. "Podcasting. New Aural Cultures and Digital Media." In Palgrave Macmillan Cham. DOI: https://doi.org/10.1007/978-3-319-90056-8.

MacKenzie, Lewis. 2019. “Science podcasts: analysis of global production and output from 2004 to 2018.” Royal Society Open Science 6: 180932. DOI: https://doi.org/10.1098/rsos.180932.

Prinz, Claudia. 2018. "Kommunikation im digitalen Raum." In Clio-Guide:Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften, 2. Auflage, S. A.4-1 – A.4-22.

Detection and Classification of Historic Watermarks using neural networks and nearest neighbor search

Pfaff, Sebastian
s.pfaff@tum.de
Technische Universität München, Deutschland

Beriozchin, Evghenii
evghenii.beriozchin@tum.de
Technische Universität München, Deutschland

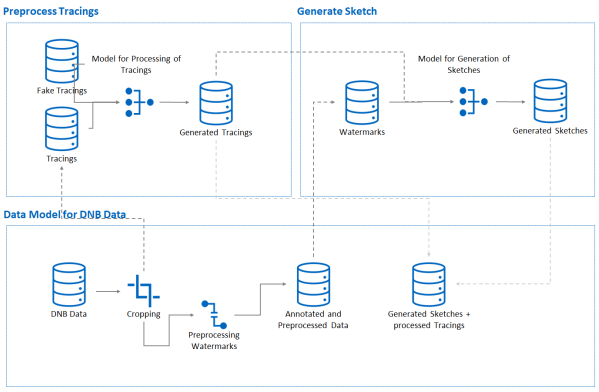
Weyh, Paulina
paulina.weyh@tum.de
Technische Universität München, Deutschland

The history of paper in Europe dates back to the 12th century when papermaking technology was imported from China through the Islamic world. The first paper mills were established in Spain and Italy, followed by France and Germany in the 13th century. Initially, paper was produced by scooping pulp from linen fibers with a sieve. These frames were often embroidered with a metal wire that formed some design or pattern, e.g., animals, or letters. The wire had a different thickness than the surrounding area. As a result, the pattern from the wire was imprinted on the paper. We call this imprint a watermark. (Hunter 1978) (Damberger 2006) (Vereinigung der Österreichischen Papierindustrie 2023)

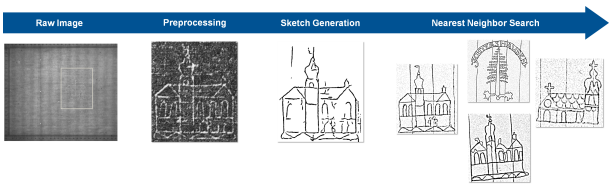
The study of watermarks holds significance for historical humanities. Watermarks offer valuable insights into the origins of paper, aiding in identifying papermakers, mills, and periods when a specific piece of paper was made. This information plays a crucial role in dating and verifying historical documents. Watermark designs evolved, enabling precise document dating. By identifying watermarks in various documents, one can establish connections between manuscripts, trace paper sources, and track trade routes and distribution networks. However, the identification and comparison of watermarks is currently difficult and time-consuming. (Barrett 2022) (Fuller 2002)

Historians ask the German National Library (DNB) to identify watermarks, typically by sending image attachments by mail. Then, a DNB expert physically goes into the archive to find the same or a similar watermark (e.g., from the same mill but a different period). Since this manual process requires highly trained and experienced people, it is both slow and expensive. Thus, there is a bottleneck for watermark recognition since the DNB lacks resources to quickly handle all requests.

We build a “human in the middle” model, to help historians efficiently search for watermarks on their own. We provide the user with a ranking containing the most similar images to her query image. Then the user must manually compare their watermark with the results from the nearest neighbor (NN) search. This simple comparison can easily be done by non-experts in a reasonable time.



This work presents a novel approach to automatically find similar watermarks based on the digitized collection of historical papers, watermarks, and traced watermarks from the German Museum of Books and Writing and the DNB (Deutsche Nationalbibliothek 2023). Previous approaches either used different image processing techniques or trained neural networks for classification (Stewart 1995) (Belov 1999). One disadvantage of the first one is its lack of robustness, while the second approach cannot deal with unseen groups of watermarks. Furthermore, existing approaches only include high-contrast tracings in the database, e.g., (Picard 2016) (Pondenkandath, et al. 2020) (Deng 2009) (Shen 2019). Limiting the database to tracings excludes much data in training and makes it impossible to find watermarks without corresponding tracings. As the DNB continuously adds new watermarks to the database, we aimed for an approach adaptive to new watermarks. Furthermore, as the watermarks are loosely labeled, the network needs to be independent of input-output pairs.



In our approach, we create an unpaired dataset of watermarks and preprocessed tracings (sketches). Using this dataset, we train a CycleGAN neural network that can generate a sketch of a watermark present in a scan of a historical paper (Zhu 2017). Using this model, we generate sketches for all watermarks from the dataset and combine them with the sketches produced by preprocessing tracings. We utilize a pre-trained ResNet18 neural network for extracting a feature vector from the sketch. Finally, we use the Spotify Annoy algorithm, for an efficient approximate NN search in the entire database. (Bernhardsson 2018).

To test the pipeline, we selected a test set with 22 classes of watermark-tracing groups ranging from 2 to 167 observations per class. Executing the pipeline on the watermark against a database of over 6200 digitized watermarks and traced watermarks, we achieve an accuracy of 50% of finding a corresponding tracing within 25 NNs, and over 68% within 50 NNs.

The pipeline shows promising results applicable in different scenarios. Non-experts can identify their watermark by examining fewer than 50 watermarks (~70% success). We anticipate even better results with 100-150 NNs. Watermark-experts can find similar watermarks based on the content of the image to find correlations of scientific importance. Moreover, the database easily integrates with DNB metadata for additional details on the watermark.

Training on a larger dataset or using a transformer-based model could enhance the pipeline and database, making it a primary resource for both experts and non-experts in historical watermark research.

Full report: https://www.mdsi.tum.de/en/di-lab/vergangene-projekte/ss2023-tu-delft-detection-and-classification-of-historic-watermarks/

Code: https://github.com/EvgheniiBeriozchin/watermark-detection

Bibliographie

Barrett, Timothy et al. 2022. "European Papermaking Techniques 1300-1800." University of Iowa.

Belov, V.V. and Esipova, V.A. and Kalaida, V.T. and Klimkin, V.M. 1999. "Physical and Mathematical Methods for the Visualization and Identification of Watermarks." *Solanus.*

Bernhardsson, Erik. 2018. "ANNOY library." Spotify.

Bounou, Oumayma, Tom Monnier, Ilaria Pastrolin, Xi SHEN, and Christine Benevent et al. 2020. "A Web Application for Watermark Recognition." *Journal of Data Mining & Digital Humanities*, 07 14.

Damberger, Joachim. 2006. "Geschichte der Papierherstellung." (LWF - aktuell).

Deng, Jia and Dong, Wei and Socher, Richard and Li-Jia Li, Kai Li and Li Fei-Fei. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." *2009 IEEE Conference on Computer Vision and Pattern Recognition.* Miami, Florida: Institute of Electrical and Electronics Engineers.

Deutsche Nationalbibliothek. 2023. *Historical Paper Collections* . 06. https://www.dnb.de/EN/Sammlungen/DBSM/PapierhistorischeSammlungen/papierhistorischeSammlungen.

Fuller, Neathery Batsell. 2002. "A Brief History Of Paper." St. Louis Community College.

Hunter, D. 1978. *Papermaking: The History and Technique of an Ancient Craft.* Dover Publications.

Picard, David and Henn, Thomas and Dietz, Georg. 2016. *Non-negative dictionary learning for paper watermark similarity.* Conference, Pacific Grove, United States: Asilomar Conference on Signals, Systems, and Computers.

Pondenkandath, V., M. Alberti, N. Eichenberger, R. Ingold, and M Liwicki. 2020. "Cross-Depicted Historical Motif Categorization and Retrieval with Deep Learning." *Journal of Imaging 6*, 71.

Shen, X., Pastrolin, I., Bounou, O., Gidaris, S., Smith, M., Poncet, O., & Aubry, M. 2019. "Large-Scale Historical Watermark Recognition: dataset and a new consistency-based approach." *2020 25th International Conference on Pattern Recognition (ICPR).* Milan: International Association of Pattern Recognition.

Stewart, D. and Scharf, R. A. and Arney, J. S. 1995. "niques for digital image capture of watermarks." *Journal of Imaging Science and Technology.*

Vereinigung der Österreichischen Papierindustrie. 2023. *Papier macht Schule - Geschichte der Papierproduktion.* Accessed Juni 2023. https://www.papiermachtschule.at/papierproduktion/geschichte/.

Zhu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A. 2017. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks." Computer Vision (ICCV).

Diagramme repräsentieren: Zu einer neuen Editionspraxis

Sutor, Nadine
sutor@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland
ORCID: 0000-0003-0577-4470

Einleitung und Problemstellung

In den letzten Jahrzehnten haben sich unsere Möglichkeiten zur Visualisierung von Information rasant weiterentwickelt. Für die Untersuchung bestimmter Fragestellungen sind digitale Methoden in Wissenschaft und Forschung längst angekommen – auch in der Erforschung frühmittelalterlicher Diagramme. Diese bieten einzigartige Einblicke in die Wissensdarstellung und -vermittlung dieser Epoche. Ihre Edition und Interpretation stehen jedoch vor drei besonderen Herausforderungen. 1.) setzt die Entschlüsselung der symbolischen und ikonografischen Bedeutung umfangreiche Kenntnisse der frühmittelalterlichen Kultur und des historischen Kontextes voraus. 2.) erfordert die Umwandlung von handschriftlichen Ausdrucksformen in digitale Repräsentationen (“kritische Wiedergabe”) eine sorgfältige Abwägung zwischen Quellentreue auf der einen und deutender Nutzernähe auf der anderen Seite. 3.) sind frühmittelalterliche Diagramme mit komplexen symbolischen Bezügen und Bedeutungsebenen versehen, die von zeitgenössischen Leser*innen verstanden wurden, für moderne Betrachter*innen jedoch nicht unbedingt offensichtlich sind. Die Editionspraxis muss daher Methoden und Werkzeuge entwickeln, um diese symbolische Sprache zu entschlüsseln und eine korrekte Interpretation ermöglichen.

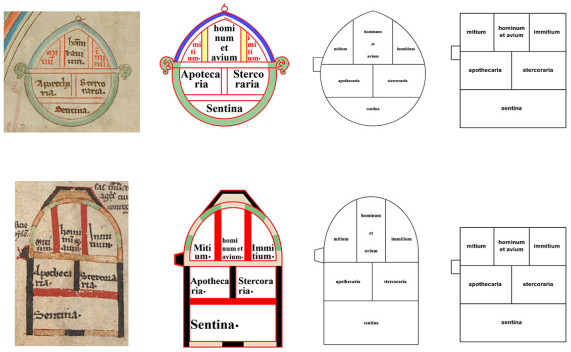
Durch mein Poster sollen insgesamt drei Wissenschaftsfelder näher betrachtet werden: 1) Semiotik 2) Editionswissenschaft 3) Sprachwissenschaft. Thematisch stehen dabei drei Aspekte im Fokus: 1) Entwicklung einer Beschreibungssprache für Diagramme 2) Strategien der Repräsentation im Skalenmodell 3) Dimensionen von Semantik in den Diagrammen. Die zentralen Aussagen sollen anhand eines Beispieldiagramms (Arche Noah) auf dem Poster präsentiert werden.

Eine Beschreibungssprache für Diagramme

In den 1960er Jahren entwickelte Jacques Bertin in seiner *Sémiologie Graphique* ein Konzept für die Untersuchung grafischer Elemente und visueller Strukturen von grafischen Darstellungen (Bertin, 1967). Inwieweit kann Bertins Modell bei der Entwicklung einer allgemeinen Beschreibungssprache für frühmittelalterliche Diagramme helfen? Es geht um den Versuch, Ausdruckskanäle von Informationsvisualisierung zu systematisieren und als Werkzeug zur Analyse und zur medialen Transformation zu nutzen. Anhand von sechs grafischen Variablen zeigt Bertin, dass eine grafische Darstellung aus Größe, Helligkeit, Muster, Farbe, Richtung und Form zusammengesetzt werden kann. Je nach Eigenschaft der darzustellenden Daten schlägt er eine bestimmte Nutzung dieser Variablen vor. Charles Peirce beschäftigte sich ebenfalls mit Zeichensystemen. Er konzentrierte sich jedoch primär auf abstrakte Zeichentheorien in der Philosophie und untersuchte Diagramme im Kontext von Mathematik und Logik (Bogen, 2005, 76). Bertins Schwerpunkt liegt auf der praktischen Anwendung der Semiotik auf die visuelle Darstellung von Information, insbesondere in der Kartografie und Datenvisualisierung. Vor diesem Hintergrund scheint Bertin ertragreicher, um die visuellen Elemente des Archen-Diagramms zu systematisieren und zu beschreiben, was zu einem tieferen Verständnis über dessen Funktion und Kommunikationsabsicht führt.

Abstraktion und Repräsentation: SVG als Verfahren der digitalen Bildcodierung

Eingebettet ist diese Diskussion in die Theorie der Repräsentation medialer Objekte wie Text oder Bild als Skala (Sahle, 2013), die Abstraktionsstufen bzw. Verarbeitungsschritte kartiert. Für Texte liefern Texttheorie und Editorik belastbare Ansätze: Was ist ein Text? Was sind Textwahrnehmungsschichten? Was sind Regeln der Repräsentation von Texten? Multiple Ansätze der Textwahrnehmung, Textverarbeitung und Textpräsentation lassen sich auf dem sogenannten Skalenmodell abbilden. Dieses positioniert eine quellennahe Darstellung (Faksimile) des Textes auf der einen Seite und eine benutzernahe Wiedergabe des Textes (konstituierter und bereinigter (Lese-)Text) auf der anderen Seite. Dieses Modell ist, so die These, auf Diagramme übertragbar. Das Ziel ist die explorative Realisierung unterschiedlicher Repräsentationsformen für Diagramme, die es für Texte schon gibt. Was bedeutet es, Diagramme auf dieser Skala zu verorten? Was ist der Mehrwert solcher Repräsentationsformen? Unterschiedliche Abstraktionsstufen geben Aufschlüsse über die Entstehung, über Schreiberspezifika oder offenbaren stemmatologische Abfolgen und Varianz. Es wurden bereits erste praktische Experimente durchgeführt, denen ein adaptiertes Skalenmodell zugrunde liegt. Anhand einer faksimilierten, mimetischen, intentionalen, kritischen, abstrakten und idealisierten Darstellung der Arche Noah zeigte sich, dass SVG (Scalable Vector Graphics) als Technologie für die Entwicklung skalierbarer Vektorgrafiken und somit für die Realisierung der Abstraktionsstufen gut geeignet ist.



Petrus von Poitiers, *Compendium historiae* (12. Jh.). Fortschreitende Abstraktion der Arche Noah.

Semantische Decodierung und Re-codierung

Die semantische Analyse soll Antworten auf die Frage liefern, wie “Bedeutung” in (historischen) Diagrammen ausgedrückt wird. Theologische oder wissenschaftliche Schaubilder enthalten oft Symbole, Bilder und Text, die auf komplexe Weise miteinander interagieren. Ihre Untersuchung erfordert ein interdisziplinäres Verständnis unter Einbeziehung der Geschichte, Kunstgeschichte und der Philologien. Welche Dimensionen von Semantik können in dem Diagramm der Arche Noah modelliert werden? Wie können diese auf praktischer Ebene, z.B. mit Methoden aus dem Bereich des Semantic Web formalisiert und (re)codiert werden? Auf welcher Granularitätsebene? Erste Überlegungen haben gezeigt, dass durch die Anwendung semantikbezogener Konzepte aus den Sprachwissenschaften (die Linguistik spricht von Morphologie, Syntax, Semantik und Pragmatik) historische Diagramme auf ähnliche Weise analysiert werden können wie Texte. Es scheint möglich, sprachwissenschaftliche Methoden auf Diagramme zu übertragen - insbesondere dann, wenn wir diese als sekundären Ausdruck von etwas, das ursprünglich textlich gefasst war, betrachten.

Bibliographie

Allemang, Dean, Jim Hendler und Fabien Gandon. 2020. Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL. New York: Association for Computing Machinery. https://doi.org/10.1145/3382097.

Barceló Aspeitia, A.A. 2022. "How to Visually Represent Structure." In *Diagrammatic Representation and Inference*, hg. von Valeria Giardino, Sven Linker, Richard Burns, Francesco Bellucci, Jean-Michel Boucheix und Petrucio Viana, 218-225. Cham: Springer. https://doi.org/10.1007/978-3-031-15146-0_18.

Bertin, Jacques. 1982. Graphische Darstellungen und die graphische Weiterverarbeitung der Information. Übersetzt und bearbeitet von Wolfgang Scharfe. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110871494.

Bertin, Jacques. 1974. Graphische Semiologie. Diagramme, Netze, Karten. Übersetzt von Georg Jensch, Dieter Schade und Wolfgang Scharfe. Berlin, New York: De Gruyter. https://doi.org/10.1515/9783110834901.

Bogen, Steffen. 2005. "Verbundene Materie, geordnete Bilder. Reflexion diagrammatischen Schauens in den Fenstern von Chartres." In *Bildwelten des Wissens. Diagramme und bildtextile Ordnungen*, hg. von Horst Bredekamp und Gabriele Werner. Kunsthistorisches Jahrbuch für Bildkritik 3: 72-85.

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 Denkwerkzeuge 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.

Hiippala, Tuomo und John A. Bateman. 2022. "Introducing the Diagrammatic Semiotic Mode." In *Diagrammatic Representation and Inference*, hg. von Valeria Giardino, Sven Linker, Richard Burns, Francesco Bellucci, Jean-Michel Boucheix und Petrucio Viana, 3-19. Cham: Springer. https://doi.org/10.1007/978-3-031-15146-0_1.

Kiryushchenko, Vitaly. 2015. "Maps, Diagrams, and Signs: Visual Experience in Peirce’s Semiotics." In *Handbook of Semiotics*, hg. von Peter P. Trifonas, 115-123. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9404-6_3.

Ljungberg, Christina. 2015. "Cartosemiotics." In *Handbook of Semiotics*, hg. von Peter P. Trifonas, 759-769. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9404-6_35.

Neuman, Yair. 2015. "Semiotics as an Interdisciplinary Science." In *Handbook of Semiotics*, hg. von Peter P. Trifonas, 125-134. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9404-6_4.

O’Halloran, Kay L. 2015. "Multimodal Digital Humanities." In *Handbook of Semiotics*, hg. von Peter P. Trifonas, 389-415. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9404-6_18.

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.

Sahle, Patrick. 2013. Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels, 3 Bände, Norderstedt: Books on Demand.

Salameh, Khouloud, Joe Tekli und Richard Chbeir. 2014. "SVG-to-RDF Image *Semantization*". In *Similarity Search and Applications*, hg. von Agma J. M. Traina, Caetano Traina Jr. und Robson Leonardo Ferreira Cordeiro, 214-228. Cham: Springer. https://doi.org/10.1007/978-3-319-11988-5_20.

Trifonas, Peter P. 2015. "Text and Images." In *Handbook of Semiotics*, hg. von Peter P. Trifonas, 1139-1152. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9404-6_53.

Die Memoiren der Gräfin Schwerin (1684-1732). Zur digitalen Edition eines einzigartigen Selbstzeugnisses.

Weis, Joëlle
weis@uni-trier.de
Universität Trier, Deutschland
ORCID: 0000-0002-0080-4362

Galka, Selina
selina.galka@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0003-4476-315X

Peper, Ines
ines.peper@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich
ORCID: 0000-0002-8676-3935

Pölzl, Michael
michael.poelzl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich
ORCID: 0000-0002-1291-9220

Petrolini, Chiara
chiara.petrolini@unibo.it
Universität Bologna, Italien
ORCID: 0000-0003-3355-6067

In den frühen 1720er-Jahren schrieb die Gräfin Luise Charlotte von Schwerin ihre Lebenserinnerungen nieder. Diese berichten in großem Detail von ihrem Leben und ihrem Übertritt vom reformierten zum katholischen Glauben im Jahr 1719 in Wien und geben einen für diese Zeit einzigartigen Einblick in die Lebenswelt einer Frau, die daraufhin aus Preußen verstoßen wurde und sich eine neue Existenz aufbauen musste. In ihren Memoiren zeichnet sie ein außergewöhnlich präzises Bild von Handlungsspielräumen und Netzwerken von Frauen des Hofadels im frühen 18. Jahrhundert. Insbesondere aus dem Raum der Habsburgermonarchie sind aus dieser Zeit keine vergleichbaren Selbstzeugnisse von Frauen erhalten.

Im Rahmen eines vom FWF finanzierten Forschungsprojekts (Fonds zur Förderung der wissenschaftlichen Forschung, Laufzeit: 2022-2025) wird eine digitale Edition entstehen, die den Text in französischer Sprache und deutscher Übersetzung für die Forschung zugänglich macht und durch Hintergrundinformationen erschließt. Die Memoiren der Gräfin sind in zwei Abschriften in Aix-en-Provence (A) und Wien (W) überliefert. Bereits 2013 erschien eine gedruckte Edition auf Basis von Handschrift A (Daumas et al. 2013); die Wiener Abschrift wurde erst 2016 entdeckt. Die Lesefassung der digitalen Edition basiert auf beiden Abschriften und weist auch die Abweichungen aus. Bei beiden Handschriften handelt es sich offensichtlich um Abschriften, bei denen jeweils mehrere Hände involviert waren. Manuskript A ist mit dem Enddatum Juni 1721 um ca. 24% länger als Manuskript W, das bereits Weihnachten 1720 abbricht (Ms A umfasst 1428 Seiten; Ms W 630 Seiten – wobei hier die Seiten größer und enger beschrieben sind). Darüber hinaus enthält Manuskript W aber zusätzliche Passagen, die in Manuskript A vermutlich absichtlich weggelassen wurden. Zudem ist die Schreibweise von Personen- und Ortsnamen in Manuskript W deutlich genauer, so dass davon ausgegangen werden kann, dass Manuskript W einen älteren Textzustand widerspiegelt und Manuskript A vermutlich absichtliche Kürzungen – eventuell zu Publikationszwecken – erfahren hat. Um die Textgenese und Überlieferungsgeschichte weiter zu klären, wird die digitale Edition eine synoptische Lesefassung sowie eine übersetzte Version, die im Zuge des Projekts erstellt wird, zugänglich machen. Die Lesefassung soll beschlagwortet werden; daneben sollen auch inhaltliche umfangreiche Kommentare zu Personen, Orten, Ereignissen und historischen Sachverhalten zur Verfügung gestellt werden. Der so gebotene flexible Zugang zur Quelle soll die Bearbeitung diverser Forschungsperspektiven zulassen und wird durch zusätzliche Explorationsangebote ergänzt (vgl. dazu auch das Konzept der “assertive edition” Vogeler 2019).

Im Zentrum der Forschungsfragen des Projekts steht das soziale Netzwerk der Gräfin in Wien. Dieses beruhte primär auf verwandtschaftlichen und informellen Beziehungen, aber auch Hofämter und kirchliche Institutionen spielten eine wichtige Rolle darin. Bei der Kodierung wird daher ein besonderes Augenmerk auf die Annotation von Personenbeziehungen gelegt - diese sollen mit <tei:relation> im Personenregister und mit der Kodierung der relevanten Stellen im Editionstext (<tei:seg> mit dem Attribut @ana, welches auf die @xml:id der <tei:relation> verweist) modelliert werden; außerdem werden sie auch in RDF abgebildet (vgl. Galka, Vogeler 2023). Im Sinne der

“assertive edition” wird somit eine mehrschichtige Repräsentation geschaffen – TEI/XML-Annotationen werden mit externen Informationsstrukturen verknüpft, um die im Text enthaltenen Fakten als Aussagen (assertions) zu modellieren (Vogeler 2019, 315). Im Projekt wird außerdem ein eigenes Datenmodell, basierend auf bestehenden Ontologien zu sozialen Beziehungen und kinship entworfen (vgl. Chui, Grüninger und Wong 2020; Herradi et al. 2015). Zusätzlich kann das Projekt auf die im prosopographischen Forschungsportal VieCPro hinterlegten Daten zu Personen am Wiener Hof zurückgreifen, die die Informationen aus den Memoiren ergänzen werden (vgl. Romberg et al. 2023).

Auch für die Selbstzeugnisforschung sind die Memoiren eine aufschlussreiche Quelle. Claudia Ulbrich und Gabriele Jancke folgend, wird das autobiographische Schreiben der Gräfin konsequent als soziales Handeln aufgefasst, dessen Vielschichtigkeit es in der Edition aufzudecken gilt (vgl. Jancke und Ulbrich 2005). Auf textlicher Ebene sind daher Fragen nach den Kommunikationsabsichten sowie dem intendierten Publikum besonders spannend. Bei der Auswertung sollen computergestützte Analyseverfahren dabei helfen, auktoriale Strategien und stilistische Vorbilder (etwa die französischen Übersetzungen von Augustinus’ Confessiones oder die Memoiren der Madame de Guyon) aufzudecken sowie Phänomene wie text reuse zu erkennen. Die Rückbindung der Edition an Voyant Tools wird auch den Nutzer*innen erlauben, den Text auf eigene Fragestellungen hin zu explorieren.

Im Projekt konnten bereits einige Dinge abgeschlossen werden (z.B. die Transkription von Ms W mitsamt Kodierung von Textphänomenen wie Streichungen etc., XSLT-Transformationen zur Transformation des Outputs aus Transkribus in das finale Datenmodell, Website-Prototyp), andere Arbeitsprozesse laufen parallel und kontinuierlich (Verfeinerung des Datenmodells, Kodierung der Lesefassung, Übersetzung etc.). Das Poster wird insbesondere auf die hier beschriebenen angewandten Methoden und erste Ergebnisse eingehen sowie das Datenmodell vorstellen, mit besonderem Fokus auf soziale Beziehungen. Darüber hinaus wird es den Workflow der digitalen Edition detailliert skizzieren und erste Designmockups und weitere geplante Funktionalitäten präsentieren.

Bibliographie

Chui, Carmen, Michael Grüninger und Janette Wong. 2020. “An Ontology for Formal Models of Kinship.” *Frontiers in Artificial Intelligence and Applications* 330: 92–106. https://doi.org/10.3233/FAIA200663.

Daumas, Maurice und Claudia Ulbrich (Hrsg.) 2013. *Mémoires de la comtesse de Schwerin. Une conversion au XVIIIe siècle*. Bordeaux: PUB 2013.

Galka, Selina und Georg Vogeler. 2023. “Relation^3: How to relate text describing relationships with structured encoding of the relationships?” *Encoding Cultures. Joint MEC TEI conference 2023*. https://teimec2023.uni-paderborn.de/.

Herradi, Noura, Fayçal Hamdi, Elisabeth Métais, Fatma Ghorbel und Assia Soukane. 2015. “PersonLink: An Ontology Representing Family Relationships for the CAPTAIN MEMO Memory Prosthesis.” *Advances in Conceptual Modeling. ER 2015. Lecture Notes in Computer Science* 9382: 3–13.Cham: Springer. https://doi.org/10.1007/978-3-319-25747-1_1.

Jancke, Gabriele und Claudia Ulbrich (Hrsg.) 2005. *Vom Individuum zur Person. Neue Konzepte im Spannungsfeld von Autobiographietheorie und Selbstzeugnisforschung*. Göttingen 2005.

Romberg, Marion, Maximilian Kaiser, Matthias Schlögl und Gregor Pirgie. 2023 (im Druck). “Von APIS zu VieCPro. Die Entwicklung einer multifunktionalen Prosopographiedatenbank. ” In *Historische Biographik und kritische Prosopographie als Instrumente in der Geschichtswissenschaft* hrsg. v. Bianka Trötschel-Daniels. Oldenburg: De Gruyter.

Vogeler, Georg. 2019.“The ‘assertive edition’: On the consequences of digital methods in scholarly editing for historians.” *International Journal of Digital Humanities* 1: 309–322.

Digitale Begriffsgeschichte: Zur historischen Semantik des Naturbegriffs in Spanien und Lateinamerika (18. Jh.)

Hillebrand, Philip
phillebrand@uos.de
Universität Osnabrück, Deutschland
ORCID: 0009-0003-7195-2682

Schlünder, Susanne
sschluen@uni-osnabrueck.de
Universität Osnabrück, Deutschland
ORCID: 0009-0004-4914-1864

Garita Figueiredo, Renato
rgaritafigue@uni-osnabrueck.de
Universität Osnabrück, Deutschland

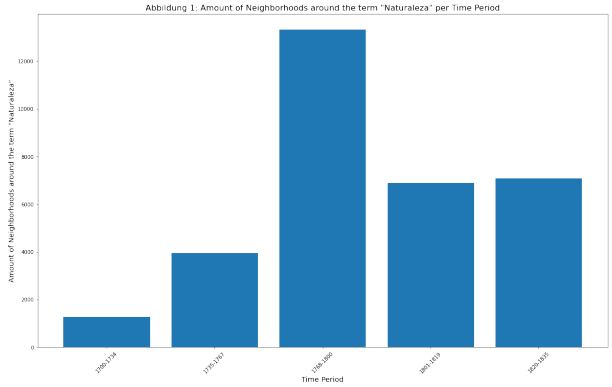
Rißler-Pipka, Nanette
rissler-pipka@maxweberstiftung.de
Max Weber Stiftung, Bonn, Deutschland
ORCID: 0000-0002-0719-9003

Der Begriff der Natur nimmt eine zentrale Stellung in den Sinn- und Handlungszusammenhängen von Gesellschaften ein. Umweltdebatten verweisen nicht zuletzt auf die konfliktive Pluralität, historische Wandelbarkeit und affektive Dimension von Naturkonzepten. Die im Zuge der Aufklärung deutlichen Umbrüche im Naturverhältnis zeichnen sich in den asymmetrischen Mensch-Umweltbeziehungen des spanischen Kolonialreichs im 18. und frühen 19. Jahrhundert anhand politischer, ökonomischer, epistemologischer und identitärer Debatten besonders deutlich ab (Carrasco und Schlünder, 2022). Trotzdem berücksichtigt die neueste Forschung zur Historischen Semantik der iberoamerikanischen Welt den Begriff ‘naturaleza’ nicht (Fernández Sebastián, 2009-2014). Das Projekt nimmt sich (1) dieser Lücke an und entwickelt (2) einen alternativen, digital begriffsgeschichtlichen Zugang, durch welchen bislang aus Quantitätsgründen nicht beforschte Korpora einbezogen werden. Dem Ansatz der digitalen Begriffsgeschichte entsprechend (Schwandt, 2018) verspricht die Verschränkung hermeneutisch-diskursanalytischer und computationeller Methoden neue Erkenntnisse als das bisherige, hermeneutisch geleitete Vorgehen klassisch qualitativer Analysen, das sich auf kanonisierte, vorwiegend ideengeschichtlich-philosophische Texte bezieht. Ziel des Projekt ist es, einen einschlägigen Beitrag zur Geschichte des Proto-Anthropozän in der iberischen Welt (Wendt, 2016) zu leisten und dabei die Verknüpfung von Digital Humanities-Ansätzen und Studien zum 18. Jahrhundert in Spanien und Lateinamerika (Apert-Abrams und McCarl, 2020) zu befördern.

Das Korpus besteht aus ca. 37.000 Dokumenten (ca. 187.830.500 Tokens) aus Spanien und den spanischen Vizekönigreichen Neugranada und Peru. Einbezogen werden insbesondere Zeitschriften, Reiseberichte, naturwissenschaftliche sowie politökonomische Abhandlungen – Textsorten, deren in ihnen eingeschriebene Diskurse Rückschlüsse auf die Naturbegriffe der Zeit zulassen. Der Großteil der Dokumente liegt in digitalisierter Form (PDF) in Bibliotheken und Archiven in Spanien, Lateinamerika und den USA vor. Um die Digitalisate weiter prozessieren und analysieren zu können, wurde ein OCR-Workflow verwendet. Mit Python-Skripten werden die Datenbanken durchsucht und die PDFs automatisch heruntergeladen. Für Dokumente ohne Textdaten wird Tesseract genutzt, wobei die OCR-Erkennungsrate maßgeblich durch die Bildqualität der Digitalisate, den Überlieferungszustand, idiosynkratische Schriftarten und orthographische Varianz beeinflusst wird. Die Performance der OCR Methode wird mittels CER und WER anhand einer Ground Truth auf einigen Dokumenten ermittelt und mit alternativen Methoden verglichen. In einem zweiten Schritt wird die Qualität jeder mittels OCR erstellten Transkription mithilfe eines Wörterbuchs, bestehend aus dem *Diccionario de Autoridades* (1726-1739) und dem Korpus *Projekt Gutenberg*, evaluiert. Für die Trennung automatisch zusammengezogener Wörter wird *Wordsegment* benutzt (Jenks, 2018). Anschließende Korrekturen des Textes werden mit *SymSpell* (Garbe, 2012), dem manuellen Beheben häufig auftretender Fehler sowie der Edition von Textpassagen vorgenommen.

Für die Analyse der Korpora werden Topic Modeling und Word Embeddings genutzt. Das Topic Modeling dient der thematischen Exploration des Korpus und der Konturierung von Teilkorpora für vergleichende Analysen. Bei Topic Models mit *LDA* (Blei, 2012) werden Dokumente als *bag of words* betrachtet. Die Festlegung der Parameter zur Bestimmung der optimalen Anzahl der Topics sowie deren Interpretation schaffen einen großen Spielraum. Dem versuchen neuere Ansätze (Egger und Yu, 2022) zu begegnen, indem sie für die Zusammenstellung von Topics Word-Embeddingrepräsentationen von *BERT* (Grootendorst, 2022) verwenden (Liimatta et al., 2023). Das Projekt testet und evaluiert diese Ansätze in Bezug auf die Anwendbarkeit und den Nutzen für die Korpusauswertung. Ziel der Word Embeddings ist es, in Kombination mit hermeneutischen Verfahren unterschiedliche Bedeutungen des Naturbegriffs zu identifizieren und Prozesse semantischen Wandels nachzuverfolgen. Dabei werden verschiedene Embedding-Modelle wie *word2vec* (Mikolov et. al., 2013), *glove* (Pennington et. al., 2014) und *fasttext* (Bojanowski et. al., 2017) erprobt, um zu evaluieren, wie sich semantische Informationen im Fall eines historischen spanischen Korpus am besten repräsentieren lassen (Hu, Amaral und Kübler, 2022). Für die diachrone Analyse sollen die Embeddings sukzessive in den einzelnen Zeitschichten des Korpus trainiert und ihre Räume anschließend angeglichen werden (Hamilton, Leskovec und Jurafsky, 2016; Kim et al., 2014). Da unser Korpus nicht groß genug ist, um Sprachmodelle vollständig zu trainieren und die Qualität der Daten zwischen den Zeiträumen variiert, werden wir versuchen, unser Korpus mit ähnlichen Quellen aus demselben Zeitraum zu kombinieren, beispielsweise mit entsprechenden Dokumenten des Gutenberg-Korpus.

In Bezug auf Textsorten, aber auch – aufgrund historischer Schreibpraktiken und Produktionsbedingungen – in Bezug auf die regionale und zeitliche Menge und Verteilung der Dokumente ist das Korpus heterogen. Spanische Schriftzeugnisse sind stärker vertreten als hispanoamerikanische, Zeitschriften häufiger als Reiseberichte, und die zweite Hälfte des 18. Jh. dominant gegenüber der ersten Hälfte. Bei der Auswertung stellt sich damit das Problem der Normalisierung und die Frage des Umgangs mit Über- und Unterrepräsentation bestimmter Daten. Abbildung 1 zeigt eine Visualisierung der Anzahl und Verteilung der Umgebungen des Lexems ‘naturaleza’ (+/- 50 Wörter) pro Zeitspanne in unseren Daten vor der Normalisierung.



Das Konferenzposter stellt die Workflows zur Korpuserstellung, -vorbereitung und -analyse zur Diskussion. Zudem werden erste Ergebnisse der Analyse präsentiert. Für die Nachnutzung und Reproduzierbarkeit der Ergebnisse werden möglichst alle Korpora (im bestmöglichen OCR-Reintext), Metadaten und Software frei zugänglich im DARIAH-DE Repository veröffentlicht. Dabei ist im Sinne der CARE Prinzipien mit den jeweiligen initialen Datenanbietern zu klären, wie angemessen auf Ursprung und Entstehung der originär lateinamerikanischen Kulturgüter verwiesen werden kann.

Bibliographie

"A Python wrapper for Google Tesseract.“ Python. https://github.com/madmaze/pytesseract (zugegriffen: 19. Juli 2023).

Albanese, Nicolo. 2022. “Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: A comparison between different topic modeling strategies including practical Python examples”. *Towards Data Science* (blog), September 19, 2022. https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5 (zugegriffen: 18. Juli 2023).

Alpert-Abrams, Hannah und Clayton McCarl. 2020. “Introduction: Digital Humanities & Colonial Latin American Studies.” *Digital Humanities Quarterly* 14, Nr. 4. http://www.digitalhumanities.org/dhq/vol/14/4/000531/000531.html (zugegriffen: 18. Juli 2023).

Blei, David M. 2012. “Probabilistic topic models.” *Communications of the ACM* 55, Nr. 4: 77-84. https://doi.org/10.1145/2133806.2133826 (zugegriffen: 18. Juli 2023).

Bojanowski, Piotr, Edouard Grave, Armand Joulin und Tomas Mikolov. (2017). “Enriching word vectors with subword information.” *Transactions of the association for computational linguistics* 5: 135-146. http://dx.doi.org/10.1162/tacl_a_00051 (zugegriffen: 18. November 2023).

Carrasco M., Rolando und Susanne Schlünder (Hg.). 2022. *Asymmetric Ecologies in Europe and South America around 1800*. Berlin/Boston: De Gruyter.

Egger, Roman und Joanne Yu. 2022. "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts." *Frontiers in sociology* 7: 886498. doi: 10.3389/fsoc.2022.886498 (zugegriffen: 18. Juli 2023).

Fernández Sebastián, Javier (Hg.). 2009-2014, *Diccionario político y social del mundo iberoamericano.* 2 Vol. Madrid: Centro de Estudios Políticos y Constitucionales.

Garbe, Wolfgang. 2012. *Symspell*. https://github.com/wolfgarbe/SymSpell (zugegriffen: 18. Juli 2023).

Grootendorst, Maarten. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." https://arxiv.org/abs/2203.05794v1 (zugegriffen: 19. Juli 2023).

Hu, Hai, Patrícia Amaral und Sandra Kübler. 2022. “Word embeddings and semantic shifts in historical Spanish: Methodological considerations”. *Digital Scholarships in the Humanities* 37, Nr. 2: 441-461. https://doi.org/10.1093/llc/fqab050 (zugegriffen: 18. Juli 2023).

Jenks, Grant. 2018. *Wordsegment*. https://github.com/grantjenks/python-wordsegment (zugegriffen: 19. Juli 2023).

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde und Slav Petrov. (2014). “Temporal Analysis of Language through Neural Language Models.” *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*: 61-65. http://dx.doi.org/10.3115/v1/W14-251726 (zugegriffen: 19. November 2023).

Liimatta, Aatu, Eetu Mäkelä, Filip Ginter, Iiro Rastas, Iiro Tihonen, Jinbin Zhang, Lidia Pivovarova et al. (2023). “Using ECCO-BERT and the Historical Thesaurus of English to Explore Concepts and Agency in Historical Writing Interpreting the Eighteenth-century Luxury Debate.” *Digital Humanities 2023. Collaboration as Opportunity (DH2023)*. https://doi.org/10.5281/zenodo.8108032. (zugegriffen: 19. November 2023).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. (2013). “Distributed representations of words and phrases and their compositionality.” *Advances in neural information processing systems* 26. https://arxiv.org/abs/1310.4546v1 (zugegriffen: 19. November 2023).

Pennington, Jeffrey, Richard Socher und Christopher Manning. (2014). “Glove: Global vectors for word representation.” *Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP): 1532-1543. http://dx.doi.org/10.3115/v1/D14-1162 (zugegriffen: 18. November 2023).

Project Gutenberg. https://www.gutenberg.org/browse/languages/es (zugegriffen: 19. Juli 2023).

Schippers, Heinrich. 1978. “Natur“. In *Geschichtliche Grundbegriffe*, hg. von Otto Brunner, Werner Conze und Reinhart Koselleck, Bd. 4, 215-244. Stuttgart: Klett-Cotta.

Schwandt, Silke. 2018. “Digitale Methoden für die Historische Semantik: Auf den Spuren von Begriffen in digitalen Korpora.” *Geschichte und Gesellschaft* 44: 107-134. https://vr-elibrary.de/doi/pdf/10.13109/gege.2018.44.1.107 (zugegriffen: 18. Juli 2023).

Wendt, Helge. 2016. “Epilogue: The Iberian Way into the Anthropocene“. In *The Globalization of Knowledge in the Iberian Colonial World*, hg. v. dems., 297-314. Berlin: Edition Open Access.

Wevers, Melvin und Marijn Koolen. 2020. “Digital begriffsgeschichte: Tracing semantic change using word embeddings.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53, Nr. 4: 226-243. https://doi.org/10.1080/01615440.2020.1760157 (zugegriffen: 19. Juli 2023).

Digitale Erschließung der Rechnungsbücher des Klosters Aldersbach Sozial- und wirtschaftshistorische Analyse eines prototypischen Großbetriebs mit digitalen Methoden

Klugseder, Robert
robert.klugseder@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich
ORCID: 0000-0002-0484-832X

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-1726-1712

Spoerer, Mark
mark.spoerer@ur.de
Universität Regensburg, Deutschland
ORCID: 0000-0002-5549-6512

Mittelalterliche und frühneuzeitliche Klosterrechnungen erlauben sehr vielfältige Einblicke, in den Alltag der Mönche ebenso wie in die sogenannte „große“ Geschichte. Und nicht zuletzt werden in diesen Quellen natürlich auch die wirtschaftlichen Angelegenheiten eines Klosters umfänglich dokumentiert. Die Aussagekraft und der Quellenwert dieser Texte sind somit sehr hoch anzusetzen. Rechnungsbücher erlauben vielfach Einblicke, die sonst verwehrt blieben. Oft wurden nämlich - unabsichtlich und daher zumeist umso wertvoller - Ereignisse und Vorkommnisse dokumentiert, die sonst in der übrigen schriftlichen Überlieferung fehlen. Aldersbach kann dabei als idealtypisches Beispiel für ein bayerisches Zisterzienserkloster während des 15. und 16. Jahrhunderts gelten. Zwar ist die Geschichte dieses Klosters selbst vergleichsweise schlecht erforscht, die gute Überlieferung gerade der Rechnungen im 15. Jahrhundert selbst sowie die Tatsache, dass Aldersbach eine vergleichsweise durchschnittliche religiöse Gemeinschaft war, die es in vergleichbarer Größe hundertfach in ganz Europa gab, lassen das niederbayerische Zisterzienserkloster als geradezu prototypisch erscheinen.

Die reichhaltige Erhaltung der Rechnungsbuchaufzeichnungen des Klosters, zusammen mit der Tatsache, dass Aldersbach in dieser Periode eine durchschnittlich große religiöse Gemeinschaft mit etwa 50 Angestellten war (ähnlich große Gemeinschaften waren in ganz Europa verbreitet), macht dieses niederbayerische Kloster zu einem veritablen Prototyp. Der Einsatz der automatisierten Layout- und Handschriftenerkennung zur Entzifferung der schwer lesbaren lateinischen und deutschen Rechnungsbücher und die automatisierte Datenverarbeitung durch die DH für eine auf dieser basierenden wirtschafts- und sozialhistorischen Klassifizierung kann in dieser Kombination als sehr effizient und innovativ angesehen werden.

Mittelalterliche Rechnungen sind eine Quelle, deren Edition sowohl paläographische und philologische Präzision als auch hoch strukturierte Datenrepräsentation in mindestens tabellarischer Form benötigt: Ohne „close reading“ der Quelle, um ihre Bearbeitungen zu dokumentieren (Streichungen und Einfügungen) kann die tabellarische Darstellung der Buchungen nicht verlässlich sein. Die unterschiedlichen Fragestellungen, die an die Quelle gestellt werden können, fordern zusätzlich, dass die Angaben in den Rechnungen leicht gefiltert und aggregiert werden können. Einfache Lösungen für diese Aufgabe sind digitale Editionen mit Standard-Tabellenkalkulationen. Die Forschungen in den DH haben jedoch gezeigt, dass die aus der digitalen Editorik geläufigen Verfahren einer Kodierung in XML/TEI die paläographischen und philologischen Befunde besser abbilden können und dass die strukturierten Daten flexibler in einem Datenbankmanagementsystem zugänglich gemacht werden. Das Projekt wird deshalb automatisch Transkriptionen mit Hilfe der Hand-Written-Text-Recognition-Software Transkribus erzeugen und dann manuell überprüfen. Sie werden aus Transkribus im Format XML/TEI exportiert. Mit Hilfe von lokalen Regelwerken (insbesondere regulären Ausdrücken) und für das Projekt spezifisch zu trainierenden automatischen Verfahren (kon-

kret mit Hilfe von Spacy) werden die Daten so annotiert, so dass die XML/TEI-Daten in eine den Standards des W3C Web of Data entsprechenden Datenbank überführt werden können. Das Projekt wird dabei die im DEPCHA-Projekt (https://gams.uni-graz.at/depcha) entwickelten Verfahren einsetzen, das an einer größeren Zahl von Rechnungen getestet wird, und auf die Vorarbeiten Georg Vogelers (https://gams.uni-graz.at/context:rem) zur digitalen Edition von mittelalterlichen Rechnungen aufbauen. Eine Forschungsaufgabe ist jedoch die Entwicklung von Regelwerken und Modellen zur automatischen Annotation der Buchungen, in denen die Kerninformationen einer Transaktion identifiziert werden müssen (das „Wer was wann/wo im Gegenzug für was“) und möglichst viele der Begrifflichkeit auf Taxonomien der für die sozial- und wirtschaftshistorische Auswertung wichtigen Kategorien abbilden sollte. Diese Kategorien können dann in einem kontrollierten Vokabular nach den Standards des W3C als Simple Knowledge Organisation System (SKOS) abgebildet werden.

Ziel einer geplanten fachwissenschaftlichen Dissertation mit dem Arbeitstitel „Leben und Arbeiten im spätmittelalterlichen Kloster. Eine Wirtschafts- und Sozialgeschichte des Klosters Aldersbach 1449-1567“ ist es, erstens die wirtschaftlichen Grundlagen des Klosters im Zeitablauf herauszuarbeiten. Dabei werden u.a. die Fragen im Vordergrund stehen, ob (und ggf. warum) die Klosterökonomie einem Strukturwandel unterlag und insbesondere, ob sich die Veränderungen des interkontinentalen Fernhandels bereits niederschlugen (neue Waren, „Preisrevolution“). Zweitens soll die Sozialstruktur der immerhin etwa 50 Beschäftigten herausgearbeitet werden. Idealerweise wird es auch möglich sein, Realeinkommen für bestimmte Berufsgruppen zu ermitteln und damit einen Beitrag zur seit gut zwanzig Jahren boomenden internationalen Forschung zum Lebensstandard in der Vormoderne beizusteuern. Grundlagen für die wirtschafts- und sozialhistorische Analysen sind die Rechnungsbücher des Klosters, die bis 1512 fast vollständig (auf Latein) und dann noch einmal von 1552 bis 1567 (auf Deutsch) vorliegen. Sie sollen digitalisiert, und, automatisch transkribiert und dann mit Hilfe neuer DH-Verfahren ebenfalls semiautomatisch analysiert werden.

Beteiligte:

PD Dr. Robert Klugseder, ÖAW-ACDH-CH Wien, Projektleiter, Transkribus Ambassador und Spezialist für die Aldersbacher Klostergeschichte.

Univ.-Prof. Dr. Georg Vogeler, Zentrum für Informationsmodellierung der Universität Graz. Historiker und Spezialist für Digital Humanities im Allgemeinen und im Besonderen für die digitale Edition von historischen Rechnungen.

Prof. Dr. Mark Spoerer, Institut für Geschichte der Universität Regensburg, Lehrstuhlinhaber Wirtschafts- und Sozialgeschichte.

Digital Humanities in Discuss Data: Aufbau eines Community Spaces

Kahlert, Torsten
kahlert@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Kurzawe, Daniel
kurzawe@sub.uni-goettingen.de
Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0001-5027-7313

Die Digitalisierung der Forschung hat eine tiefgreifende Veränderung der Forschungsparadigmen und -methoden bewirkt. Dies betrifft insbesondere auch Forschungsdaten und den Umgang mit diesen im Forschungsprozess. Indem Fachkommunikation, Begutachtung und Netzwerkbildung zunehmend in den digitalen Raum verlagert werden, ändern sich damit verbundene Prozesse. Etablierte Strukturen lösen sich auf und werden durch digitale Angebote abgelöst. Digitale Forschung und Methoden der Data Science finden Anwendung in den Geisteswissenschaften. (Rapp 2021) Aktuelle Standards und die Digitalisierung bestehender Prozesse und Angebote benötigen jedoch Strukturen, um nachhaltige Entwicklungsmodelle zu schaffen und zu ermöglichen. (Bingert, Buddenbohm, Engelhardt, Kurzawe 2017) Dies betrifft auch die Betrachtung der Datenqualität, welche in der offenen akademischen Diskussion von immer zentralerer Bedeutung wird. (Rat für Informationsinfrastrukturen 2019) Discuss Data bietet hierfür eine Plattform, die dem digitalen Forschungsdatenmanagement (FDM) eine weitere Ebene hinzufügt. Zur informationstechnischen Verwaltung, Archivierung und Bereitstellung von Daten kommt deren Kontextualisierung durch kuratierte Diskussion. So werden Diskurse über die Daten sowie deren Kontext und Methoden direkt am Objekt zugelassen und gefördert. (Herrmann und Kurzawe 2020) Die Plattform adressiert hierzu jeweils einzelne Communities und bietet diesen einen fachspezifischen Diskussionsraum und perspektivisch auch communityspezifische Werkzeuge. Communities sind dabei nicht Fächern oder Fachgruppen gleichzusetzen, sondern verstehen sich als Interessengruppen zu bestimmten Fragestellungen oder Datenmaterialien.

Mit diesem Poster wollen wir den neu aufzubauenden Digital Humanities Communityspace mit seinen Spezifika beschreiben und zeigen, wie Discuss Data diese Spezifika aufgreifen kann.

Erfahrungen im Aufbau von Discuss Data

Discuss Data startete 2016 als DFG gefördertes Projekt mit dem Ziel, eine Infrastruktur für Daten zu schaffen, in welcher diese im Kontext zum jeweiligen Diskurs stehen. Ziel des Vorhabens war es, das Bewusstsein für Forschungsdatenmanagement und Datentransparenz in der Forschungscommunity zur Forschung zu Osteuropa, Südkaukasus und Zentralasien zu stärken. Seit dem ersten Release im September 2020 wurden über Discuss Data 105 Datensätze publiziert und es haben sich 126 Nutzende registriert (Stand 12.07.23). Die von Discuss Data bereitgestellte Diskussionsfunktion wurde bisher vergleichsweise wenig verwendet. Dies ist insofern überraschend, als dies ein durch Forschende oft positiv erwähntes Alleinstellungsmerkmal ist. Es besteht eine gewisse Hemmung, Daten anderer Forschender öffentlich zu kommentieren und sich somit selbst zu exponieren. (Barlösius 2023) Diese auf Konferenzen und Reviews durchaus übliche und fachlich äußerst wichtige Diskussionskultur hat sich, trotz der positiven Haltung dazu, bisher nicht etabliert.

Aufbau eines Discuss Data Communityspaces für die Digital Humanities

Die DH Community ist als *Community of Practice* mit dem gegenseitigen Kommentieren und Begutachten im digitalen öffentlichen Raum vertrauter als viele andere Fachcommunities. Sie differenziert sich intern zunehmend aus, wofür die Gründung neuer Fachzeitschriften, wie das Journal of Digital History (seit 2021) oder das Journal of Computational Literary Studies (seit 2022), ein guter Indikator ist.

Bisher zielten digitale Methoden häufig darauf ab, Muster oder Trends in großen Text-, Bild- oder anderen Datenbeständen zu berechnen und sichtbar zu machen. Zukünftig werden die DH auch noch stärker dynamische Simulationen erzeugen und vermutlich wird auch KI an Bedeutung gewinnen. Bei all dem nutzen und produzieren die DH mehr als andere Disziplinen Datenbestände, bereiten diese teils aufwändig selbst auf oder verknüpfen sie miteinander, um Korpora und Datenbestände übergreifend maschinenlesbar zu machen und sie mittels digitaler Methoden weiterzuverarbeiten. Es wäre dennoch zu fragen, ob es sich bei der DH Community um eine oder ggf. auch mehrere transdisziplinäre Data Communities handelt. (Asef et al. 2022)

Während also Forschungsdaten in der DH Community eine herausragende Rolle spielen, fehlt es zugleich an communitygetragenen Möglichkeiten der kuratierten Diskussion von Forschungsdaten. Digitale Methoden- und Quellenkritik ist in den letzten Jahren eine der zentralen Herausforderungen der DH geworden. (Fickers 2020) Hierfür werden jedoch auch diskursive digitale Räume benötigt, in denen Datenkritik direkt an den Datenbeständen stattfinden kann. In der Regel werden Forschungsdaten auf institutionellen Repositorien oder Plattformen wie zenodo publiziert, jedoch ohne, dass hier eine Diskussion oder eine Qualitätskontrolle stattfinden würde, wie sie für Zeitschriftenaufsätze durch Begutachtung und redaktionelle Standards üblich ist. Dadurch bleiben Datenbestände für die Weiterverarbeitung oft ungenutzt, weil ungeklärt bleibt, welche Qualität die Forschungsdaten haben und wofür sie ggf. anschlussfähig wären.

Aus den Erfahrungen der ersten Förderphase von Discuss Data ist klar geworden, dass noch mehr Energie darauf verwendet werden muss, Datenkurator:innen und Redaktionsmitglieder zu gewinnen, um die Communityspaces auch langfristig von der Community tragen zu lassen. Dafür sind Positivbeispiele notwendig, die aufzeigen, welchen Mehrwert der Zeitaufwand individuell und für alle mit sich bringt. Die stärkere Einbindung von Diskussionen als Mikropublikationen könnte hierzu beitragen. Auch die Begutachtung und das Review von Forschungsdatensätzen wird als wichtiges Instrument der Qualitätssicherung im Forschungs- und Publikationsprozess an Bedeutung gewinnen.

Bibliographie

Asef, Esther Marie, Elisabeth Huber, Sabine Imeri, Eva Ommert, Michaela Rizzolli, and Cosima Wagner, ‘Bausteine Forschungsdatenmanagement: Data Communities: Datenmanagement jenseits von generischen und fachspezifischen Perspektiven’, in: Bausteine Forschungsdatenmanagement, 2 (2022) https://doi.org/10.17192/BFDM.2022.2.8434 .

Barlösius, Eva. 2023. „“We Share All Data with Each Other”: Data-Sharing in Peer-to-Peer Relationships“. *Minerva*, Februar. https://doi.org/10.1007/s11024-023-09487-y .

Bingert, Sven, Stefan Buddenbohm, Claudia Engelhardt und Daniel Kurzawe. „Herausforderungen und Perspektiven für ein geisteswissenschaftliches Forschungsdatenzentrum“. Bibliothek Forschung und Praxis, November 2017. https://doi.org/10.1515/bfp-2017-0036 .

Fickers, Andreas, Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?, 17 (2020), 1. https://doi.org/10.14765/zzf.dok-1765 .

Herrmann, Felix und Daniel Kurzawe. „Bausteine Forschungsdatenmanagement: 2020, 2 Discuss Data: Community-zentrierter Ansatz für das Forschungsdatenmanagement in den Geistes- und Sozialwissenschaften“. Application/pdf, 6. Oktober 2020. https://doi.org/10/gjs8hd .

Hughes, Lorna, Panos Constantopoulos, and Costis Dallas. “Digital Methods in the Humanities: Understanding and Describing Their Use across the Disciplines”, in: A New Companion to Digital

Humanities, hg. v. Susan Schreibman, Ray Siemens, and John Unsworth, Chichester 2015, S. 150–70. https://doi.org/10.1002/9781118680605.ch11 .

Rapp, Andrea. „Digitalisierung – Chancen für Überlieferung und geistes- und kulturwissenschaftliche Forschung“. Bibliothek Forschung und Praxis 45, Nr. 2 (2021): 255–61. https://doi.org/10/gm5x5z .

RfII – Rat für Informationsinfrastrukturen: Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage, Göttingen 2019, 172 S.

Digitalisierung von Sammlungssystematiken und Sammlungskatalogen am Beispiel der geowissenschaftlichen Systematiken von Abraham Gottlob Werner (1749–1817)

Rietdorf, Clemens
clemens.rietdorf@gmail.com
Universität Leipzig, Deutschland

Niekler, Andreas
andreas.niekler@uni-leipzig.de
Universität Leipzig, Deutschland
ORCID: 0000-0002-3036-3318

Heide, Gerhard
gerhard.heide@mineral.tu-freiberg.de
Technische Universität Bergakademie Freiberg, Deutschland

Burghardt, Manuel
manuel.burghardt@uni-leipzig.de
Universität Leipzig, Deutschland
ORCID: 0000-0003-1354-9089

Hintergrund

Abraham Gottlob Werner (1749–1817) widmete mehr als 40 Jahre (1775 bis 1817) seiner Lehrtätigkeit und Forschung an der Bergakademie Freiberg der geowissenschaftlichen Systematisierung von Mineralen, Erden und Gesteinen. Seine wissenschaftlichen Thesen, Erkenntnisse und Publikationen sowie sein Engagement als Lehrer und sein Einfluss auf eine Vielzahl von Schülern prägten maßgeblich die Entwicklung der Geowissenschaften, insbesondere in den Bereichen Mineralogie und Geologie („Abraham Gottlob Werner“ 2023). Schon während seiner Studienzeit in Leipzig entwarf Werner eine bahnbrechende Klassifikation für Mineralien, die außergewöhnlich erfolgreich war und ihm eine Professorenstelle in Freiberg einbrachte. Diese innovative Systematik umfasste nicht nur die heutige Definition von Mineralien, sondern auch Erden, verschiedene Arten von Gesteinen sowie organische Naturprodukte, die dem Reich der Mineralien zugeordnet wurden. Auf der Grundlage dieser Systematiken wurden etliche Sammlungen angelegt. Auch Johann Wolfgang von Goethe interessierte sich für die Mineralogie und war ein enger Freund von Werner. Es ist bekannt, dass Goethe während eines Besuchs in Freiberg die Wernersche Systematik intensiv studierte und Werners Mineraliensammlung bewunderte. Dies ist nur ein Beispiel für eine Sammlung in einem Netzwerk aus Sammlern, Händlern und entstehenden Systematiken, die zu dieser Zeit interagierten. Die Mineralsystematik Wernes und die auf ihr basierenden Sammlungskataloge zusammengenommen bieten eine einzigartige Möglichkeit für wissenschaftshistorische Forschung im Bereich der Geowissenschaften. Die Verknüpfung von Sammlungen durch Händler und Systematiken eröffnet auch Einblicke in die sozialen Netzwerke jener Zeit, da solche Sammlungen nicht für jedermann erschwinglich waren.

Fragestellung

Der Wert von historischen wissenschaftlichen Sammlungen wird oft in Frage gestellt, insbesondere wenn es sich um Sammlungen handelt, die in ihrem ursprünglichen Fachgebiet keinen direkten Nutzen mehr für Forschung oder Lehre haben (Heide und Heide 2021). Sie waren jedoch oft von großer Bedeutung für die Entstehung und Entwicklung von wissenschaftlichen Disziplinen und sind daher eine wichtige Primärquelle für die wissenschaftshistorische Forschung (Ludwing und Weber 2013). Solche Sammlungen bewahren und dokumentieren langfristige wissenschaftliche Entwicklungen, machen sie nachvollziehbar und erlauben auf einzigartige Weise die Beantwortung von wissenschaftshistorischen Forschungsfragen (Wissenschaftsrat 2011). Systematiken, wie die von Werner, erleichtern das Verständnis und die systematische Erschließung einer Sammlung, indem sie ein zugrunde liegendes Konzept bieten. Die Dokumentationen der Systematisierung von historischen wissenschaftlichen

Sammlungen bildet somit eine entscheidende Quelle für die wissenschaftshistorische Forschung. Sie digital zu erfassen, zu erschließen und aufzubereiten bietet die Möglichkeit, neue und bestehende Forschungsfragen zu beantworten.

Dies soll im Rahmen dieser Studie für die geowissenschaftlichen Sammlungen von Werner und seinem Umfeld geschehen. Die Grundlage für diese Studie bilden vier Sammlungskataloge zu systematischen Mineraliensammlungen und drei zu unterschiedlichen Zeitpunkten veröffentlichte Versionen der Mineralsystematik Werners (siehe Tabelle 1). Durch die Ergründung dieser Werke, die in der Umgebung Werners entstanden sind, eröffnet sich eine einzigartige Möglichkeit, die Geowissenschaften um 1800 und ihre Entwicklung zu erforschen. Insbesondere durch die herausragende Stellung Werners und den Einfluss seiner Lehrtätigkeit wird ein wertvolles Fenster geöffnet, um tiefgreifende Einblicke in die Entwicklung dieses Fachgebiet zu gewinnen. Durch die unterschiedlichen Ausführungen der Systematik lässt sich sowohl die Weiterentwicklung von Werners System als auch die praktische Umsetzung anhand von Sammlungskatalogen nachvollziehen.

Tabelle 1: Übersicht über die erfassten Sammlungskataloge und Systematiken.

Werk	Typ	Autor*in	Jahr
Mvsevm Leskeanvm. 2,1: Regnvm Minerale (https://mdz-nbn-resolving.de/details:bsb10706426)	Sammlungs-katalog	D.L.G. Karsten	1789
Ausführliches und systematisches Verzeichnis des Mineralien-Kabinets des weiland kurfürstlich sächsischen Berghauptmans Herrn Karl Eugen Papst von Ohain (https://mdz-nbn-resolving.de/details:bsb10284854)	Sammlungs-katalog	A.G. Werner	1791
Katalogs der Oryktognostische Sammlung von Abraham Gottlob Werner (Quelle: TU Freiberg)	Sammlungs-katalog	-	1823
Goethes Sammlungen: Zur Mineralogie, Geologie und Paläontologie (Quelle: TU Freiberg)	Sammlungs-katalog	H. Prescher	1978
Mineralsystem des Herrn Inspektor Werners mit dessen Erlaubnis herausgegeben (Quelle: TU Freiberg)	Systematik	C.A.S. Hoffmann	1789
Grundriß der Mineralogie, nach dem neuesten Wernerschen System : zum Gebrauch bey Vorlesungen auf Akademien und Schulen (https://digitale.bibliothek.uni-halle.de/vd18/content/titleinfo/1658913)	Systematik	Johann Georg Lenz	1793
Abraham Gottlob Werners letztes Mineral-System (Quelle: TU Freiberg)	Systematik	J.C. Freiesleben	1817

Methodische Umsetzung

Das Ziel unserer Arbeit ist die Erstellung einer Graph-basierten Datenbank in der die Sammlungen und Systematiken, sowie die Fundorte und Eigenschaften der Mineralien strukturiert erfasst und verknüpft sind. Durch den Einsatz von Natural Language Processing für die Extraktion von Informationen aus den historischen Quellen wird dies effektiv umgesetzt. Die Werke, die in dieser Studie als Datengrundlage dienen, sind in bereits digitalisierten Versionen vorhanden[12] oder liegen in einer OCR bearbeiteten beziehungsweise transkribierten Fassung vor[3]. Aber auch wenn die Daten bereits digital erfasst wurden, ist es eine Herausforderung, relevante und detaillierte Informationen einfach und automatisiert aus ihnen zu gewinnen. Etwas, das für viele historische Dokumente gilt (Quaresma und Finatto 2020). Dies liegt unter anderem daran, dass Textverarbeitungswerkzeuge meist nicht auf historische Texte angepasst sind und in der Regel weniger gute Resultate liefern, wenn sie auf diese angewandt werden. Die Unterschiede auf der syntaktischen Ebene und im Vokabular im Vergleich zu verwandten modernen Sprachen sind einer der Gründe dafür (Pettersson u. a. 2016). Es lohnt sich daher einen Blick in die Daten selbst zu werfen und ihre typografische Struktur und das Seitenlayout zur Extraktion von relevanten Informationen genauer zu studieren. Dadurch können Annahmen bezüglich der Grenzen von Abschnitten und Paragraphen getroffen und dann regelbasiert extrahiert werden (Vlachidis und Tudhope 2013). In dieser Studie werden aus den verwendeten Katalogen und Systematiken die einzelnen Einträge durch Ausnutzung ihres charakteristischen inhaltlichen und typografischen Aufbaus und aufgrund der Seitenstruktur extrahiert. Dabei werden durch die Ausnutzung struktureller Muster automatisch Informationen wie die Kategorisierung des Fundstücks, seine Beschreibung und der Vermerk des Fundorts abgeleitet (siehe Beispieleintrag in Abbildung 1). Aus den Versionen der Systematik werden die Ordnungen von Arten, Gattungen, Geschlechtern und Klassen der Mineralien extrahiert.

Abbildung 1: Eintrag aus der Sammlung des Karl Eugen Papst von Ohain. Die Eingeschalten können extrahiert werden wie z.B. der Fundort Färör Inseln.

Alle Informationen werden anschließend verknüpft und in einer Graph-Datenbank gespeichert. Weiterhin werden die aus den Versionen von Werners Systematik extrahierten Kategorisierungen hierarchisch verknüpft, so dass die Systematik als Ganzes und deren zeitliche Varietäten abgebildet werden. In Abbildung 2 zeigen wir einen Ausschnitt der Graph-Datenbank und der entstandenen strukturierten Modellierung.

Anwendung

Über die Graph-Datenbank und die verknüpften Informationen lassen sich nun verschiedene für die wissenschaftshistorische Forschung relevante Fragen beantworten. Mithilfe dieser Methode können wir beispielsweise untersuchen, wie sich im Laufe der Zeit die Systematik verändert hat und welche Mineralnamen hinzugefügt oder entfernt wurden. Darüber hinaus kann analysiert werden, welchen Eigenschaften vermehrt Mineralien zugeordnet

wurden. Durch die Datenstruktur lässt sich auch leicht erkennen, wie sich die Zusammensetzung von Sammlungen sowohl hinsichtlich der Herkunft als auch der Mineraltypen darstellt. Des Weiteren können wir die räumliche Überschneidung der Sammlungen in Bezug auf Fundorte untersuchen und feststellen, wie stark die Sammlungen sich überlappen, was möglicherweise Rückschlüsse auf involvierte Händler oder Kontakte zwischen den Sammlern zulässt.

Die Graph-Datenbank soll nach ihrer Fertigstellung frei verfügbar sein und wird in den nächsten Monaten aber bis spätestens zur Konferenz veröffentlicht.

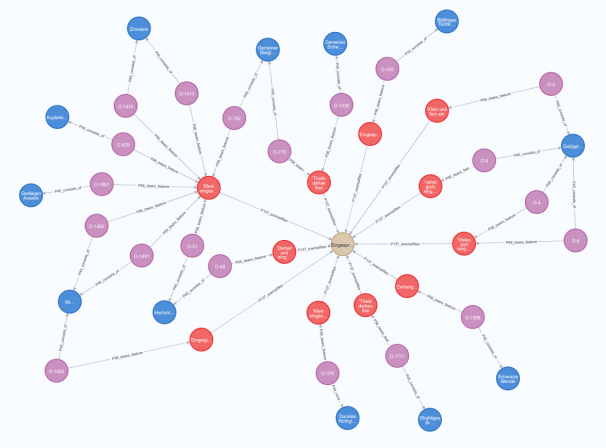


Abbildung 2: Ausschnitt des Graph-basierten Datenmodells mit eingeblendeten Mineraliennamen und Eigenschaften.

Beitrag des Posters

Die Beiträge des Posters umfassen die Präsentation von automatischen Informationsextraktionsprozessen, die für die Extraktion aus den digitalisierten Dokumenten entwickelt wurden, einschließlich der NLP-Methoden. Des Weiteren werden Beispieldaten vorgestellt und die daraus extrahierten Informationen werden anschaulich dargestellt. Ein Beispiel der Verknüpfungen im Graphen sowie eine exemplarische Illustration eines konkreten Minerals sind ebenfalls Teil des Posters. Zusätzlich wird auf dem Poster eine spezifische Fragestellung hervorgehoben, die sich mit der räumlichen Verteilung der Fundorte von Mineralien in der Sammlung von Karl Eugen Papst von Ohain auseinandersetzt. Die Fragestellung kann anhand der Daten beantwortet werden und das Poster demonstriert mithilfe von Illustrationen wie diese mit dem Graph operationalisiert und beantwortet wird.

Fußnoten

1. Die von Werner und Karsten erstellten Sammlungskataloge sowie die von Hoffmann herausgegebene Systematik sind über das Münchener Digitalisierungszentrum verfügbar.
2. Die von Lenz herausgegebene Systematik ist über die Universitäts- und Landesbibliothek Sachsen-Anhalt verfügbar.
3. Dies betrifft die Kataloge zu den Sammlungen Goethes (OCR bearbeitet) und Werners sowie die von Freiesleben herausgegebene Systematik (beide in transkribierter Form).

Bibliographie

„Abraham Gottlob Werner“. 2023. In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Abraham_Gottlob_Werner&oldid=233844512.

Heide, Gerhard, und Beata Heide. 2021. „Oryktognostische Sammlung von Abraham Gottlob Werner online“. *ACAMONTA* 28.

Ludwing, David, und Cornelia Weber. 2013. „University collections as archives of scientific practice“. *Revista Electrónica de Fuentes y Archivos* 4 (4): 85–94.

Pettersson, Eva, Jonas Lindström, Benny Jacobsson, und Rosemarie Fiebranz. 2016. „HistSearch-Implementation and Evaluation of a Web-based Tool for Automatic Information Extraction from Historical Text.“ In *HistoInformatics@ DH*, 25–36.

Quaresma, Paulo, und Maria José Bocorny Finatto. 2020. „Information Extraction from Historical Texts: a Case Study.“ In *DHandNLP@ PROPOR*, 49–56.

Vlachidis, Andreas, und Douglas Tudhope. 2013. „Classical Art Semantics Information Extraction: CASIE Pilot Project“. In . ISKO UK.

Wissenschaftsrat. 2011. *Empfehlungen zu wissenschaftlichen Sammlungen als Forschungsinfrastrukturen*. Berlin.

Digitalität in der germanistischen Literaturwissenschaft, quo vadis? Ein Bericht aus der Praxis

Boucher, Marie-Christine
marie-christine.boucher@uni-bielefeld.de
Universität Bielefeld, Deutschland
ORCID: 0000-0002-2475-3184

Gold, Julia
julia.gold@uni-bielefeld.de
Universität Bielefeld, Deutschland

Menke, Fabian
fabian.menke@uni-bielefeld.de
Universität Bielefeld, Deutschland

Preis, Matthias
matthias.preis@uni-bielefeld.de
Universität Bielefeld, Deutschland

Benz, Maximilian
maximilian.benz@uni-bielefeld.de
Universität Bielefeld, Deutschland
ORCID: 0000-0001-7262-7679

Buschmeier, Matthias
matthias.buschmeier@uni-bielefeld.de
Universität Bielefeld, Deutschland
ORCID: 0000-0002-5264-2545

Kababgi, Daniel
daniel.kababgi@uni-bielefeld.de
Universität Bielefeld, Deutschland
ORCID: 0009-0002-0990-6418

Kauffmann, Kai
kai.kauffmann@uni-bielefeld.de
Universität Bielefeld, Deutschland
ORCID: 0000-0003-3029-3097

Erhart, Walter
walter.erhart@uni-bielefeld.de
Universität Bielefeld, Deutschland

Herrmann, Berenike
berenike.herrmann@uni-bielefeld.de
Universität Bielefeld, Deutschland
ORCID: 0000-0002-5256-0566

Dieser Beitrag berichtet aus der Praxis der Implementierung von Digitalität in der germanistischen Lehre an der Universität Bielefeld. Ausgehend von den Herausforderungen der digitalen Transformation von Gesellschaft und Geisteswissenschaften wird geschildert, welche digitalen und kollaborativen Möglichkeiten im Bereich der Literaturgeschichte entwickelt werden, wie Studierende in einem Bachelor-Studiengang der Germanistik Kompetenzen der Digital Humanities und Fachwissen über emergente Phänomene der Digitalität in Gegenwartsliteratur und Medialität erwerben, wie Fragen der Alterität in der mediävistischen Lehre begegnet wird und welche fachdidaktischen Ansätze sich zu medienkritischen Diskursen, Algorithmizität und digitaler Ästhetik bewähren.

Dabei wird kritisch beleuchtet, vor welchen Herausforderungen die Teilfächer, die Dozierenden und Studierenden gerade einer stark auf Lehramtsstudiengänge ausgerichteten Germanistik standen und stehen. Im Kontext des Projekts Curriculum 4.0.NRW (Gestaltung von Hochschulcurricula für die digitale Welt) werden das ‚Fachportal', aber auch das ‚Basismodul Literaturwissenschaft' des Studiengangs BA Germanistik an der Universität Bielefeld weiterentwickelt und zudem ein Profilmodul mit dem Schwerpunkt ‚Literatur in der Gegenwart: Kultur, Medien, Digitalität' definiert, um damit das Qualifikationsprofil des Studiengangs mit Blick auf einen weiten Begriff von Digitalität zu modifizieren. Dies erfolgt z.B. durch die Einbindung von Open Educational Resources (OERs) zu Data Literacy (BiLinked), und insbesondere zur Literaturgeschichte (LiGeDi).

LiGeDi (‚Literaturgeschichte(n) erarbeiten – Gemeinsam im Digitalen') ist ein universitäres Lehr-Lern-Projekt mit dem Anspruch, Literaturgeschichte auf kollaborative und interaktive Weise im digitalen Raum zu vermitteln, um Literaturgeschichte als ein gemeinsames Gespräch über Texte, Fachdiskurse und ihre Kontexte zu verstehen. In den digitalen Lerneinheiten werden eine Vielzahl von digitalen Formaten integriert, darunter Videos, Podcasts und interaktive H5P-Inhalte wie Quizzes, Lernkarteien oder eEssays. Die Kurse enthalten über ihre literaturgeschichtlichen Inhalte hinaus zahlreiche Verweise auf essentielle Hilfsmittel, welche für die gemeinsame Teamarbeit, und den literaturwissenschaftlichen Umgang mit Texten verwendet werden können.

In der Mediävistik wurde im Sommersemester 2022 und Wintersemester 2022/23 im Rahmen eines vom Qualitätsfonds der Universität Bielefeld geförderten Projekts ein virtueller Escape Room entwickelt und implementiert, der die fachliche Einführung begleitet und ergänzt. Virtuelle, immersive Escape Room-Erfahrungen stellen eine zukunftsorientierte Form des Lehrens und Lernens dar; die Studierenden sind spielerisch aufgefordert, die Wissensinhalte, Konzepte und Materialien anwenden, überprüfen, bewerten und in Beziehung setzen zu können, die sie parallel in einer traditionellen Lernumgebung (Präsenz- oder Hybrid-Seminar) kennenlernen. Der Escape Room ist als *branching scenario* in H5P realisiert und in Moodle eingebunden. Es besteht aus einem Abenteuer, das die Studierenden durch fünf virtuelle Räume führt, die mit Rätseln, Hinweisen und Aktivitäten gefüllt sind, die ihnen helfen, sich mit den Kursinhalten tiefergehend zu beschäftigen. Gamification-Ansätze, die im Bereich der Forschung zur digitalen germanistischen Mediävistik bislang ein Desiderat darstellen (vgl. Lienert u.a., 2022), ermöglichen eine niederschwellige, intensive Auseinandersetzung mit den fachlichen Inhalten sowie den interaktiven Wissenstransfer.

Fachdidaktische Veranstaltungen legen einen besonderen Fokus auf den rezeptiven und produktiven Umgang mit diversen digital-medialen Texten unter Vermittlungsaspekten (analoges vs. digitales Lesen, Medienverbund, Social Media/Reading, Fanfiction etc.). Ferner führen sie ein in digitale Lehr- und Lernmethoden und thematisieren medienkritische Diskurse u.a. zu Fragen der digitalen Kommunikation (z.B. Echokammer; Identitätskonstruktion; konzep-

tionelle Mündlichkeit/Schriftlichkeit), der Algorithmizität (z.B. TikTok, ChatGPT, Midjourney) und der digitalen Ästhetik (z.B. poetische KI, postartifizielle Texte). Der jährlich von der germanistischen Fachdidaktik organisierte Medienbildungstag (MeBiT) widmet sich wechselnden Themenstellungen zu Fragen des Lehrens und Lernens im (post)digitalen Zeitalter, etwa Kreativitätsentwicklung, Gamification oder Algorithmen im Unterricht.

Teildisziplinübergreifend wird ein digitales Literatur-Korpus (KOLIMO+) aufbereitet, um über Services der Nationalen Forschungsdateninfrastruktur Text+ auch für die Lehre langfristig zur Verfügung zu stehen. Anhand des skalierbaren Korpus soll die Einübung und Auseinandersetzung mit digitalen Verfahren der Philologie erprobt werden, insbesondere quantitative Verfahren des Distant Reading am Beispiel von Frequenzanalysen, Stilometrie, aber auch Annotationspraktiken und ein digitales Close Reading. Diese werden durch entsprechende Tools wie R/Python-Bibliotheken, aber auch Voyant und CATMA (‚Computer Assisted Textual Markup & Interface') praktisch eingeführt. Studierende sollen so durch die Nutzung von *state of the art*-Anwendungen nicht nur literaturhistorisches Domänenwissen erlangen, sondern auch Data Literacy-Kompetenzen, die u.a. in gesellschaftlichen (*citizenship*) und Erwerbskontexten (*employability*) maßgeblich sind.

Die Rückführung der praktischen Erfahrungen in die Konzeptualisierung und Operationalisierung einer domänenspezifischen Data Literacy sowie Aspekte der fachspezifischen Hochschuldidaktik in den Studiengang erfolgt u.a. im Rahmen eines Curriculum 4.0.NRW-Projekts zur Digitalisierung der Hochschullehre, das fachrelevante Praktiken ab dem Beginn des Studiums gezielt vermittelt. Zudem fragt die BiLinked Community of Pratice ‚Data Literacy': Worin besteht eine genuin literaturwissenschaftliche Data Literacy? Wie ist sie von Medienkompetenz abzugrenzen? Welche Tools bewähren sich für welche Zwecke? Und welche generalisierbaren Methoden-Kompetenzen jenseits des einzelnen Tools sind zentral? Darf man anderswo Abstriche machen, um zentrale Daten- und Medien-Kompetenzen zu fördern? Wenn überhaupt, wo?

Auf unterschiedlichen Ebenen wird bei uns Digitalität – teils auch in Kollaboration mit Computerlinguistik, Geschichts- und Erziehungswissenschaft sowie Gender Studies – zunehmend im germanistischen Curriculum verankert. Wir berichten von diesbezüglichen Herausforderungen und Best Practices, von Einsichten und Perspektiven.

Bibliographie

Albrecht, Christian, Matthias Preis und Peter Schildhauer. 2020. *Verstetigung im Wandel. Antinomien als Konstanten digitaler Transformation?* In *Digitale Innovationen und Kompetenzen in der Lehramtsausbildung,* hg. von Michael Beißwenger et al., 15-41. https://doi.org/10.17185/duepublico/73330 .

Beiträge Zur Mediävistischen Erzählforschung. 2022. Themenheft 12: Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik. https://doi.org/10.25619/BME_H20223 .

Universität Bielefeld. BiLinked - Universität Bielefeld. https://www.uni-bielefeld.de/lehre/innovative-lehrprojekte/bilinked/ (zugegriffen: 28. November 2023).

Ciecior, Jens, Tanja A. Kunz, Stephanie Wollmann, Karima Lanius und Matthias Buschmeier. 2023. „Literary History in Digital Teaching and Learning: The KoLidi-Project—Collaborative and Interactive Approaches for German Studies". In: *Innovative Approaches to Technology-Enhanced Learning for the Workplace and Higher Education*, hg. von David Guralnick, Michael E. Auer, und Antonella Poce, 581: 114–124. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-21569-8_11.

Curriculum 4.0.nrw | Stifterverband. https://stifterverband.org/curriculum_4_0_nrw (zugegriffen: 28. November 2023).

Cwielong, Ilona, Sophie Sossong, Malte Persike, Philipp Weyers und Alina Vogelgesang. 2021. „Daten und Data Literacy im Kontext der Wissenschaft". *Medienimpulse* , September, 35 Seiten Seiten. https://doi.org/10.21243/MI-03-21-14 .

Frederking, Volker und Axel Krommer. 2019. *Digitale Textkompetenz. Ein theoretisches wie empirisches Forschungsdesiderat im deutschdidaktischen Fokus.* https://www.deutschdidaktik.phil.fau.de/files/2020/05/frederking-krommer-2019-digitale-textkompetenzpdf.pdf .

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Mareike Schumacher und Dominik Gerstorfer. 2023. CATMA 7 (Version 7.0). Zenodo. https://doi.org/ 10.5281/zenodo.1470118 .

Herrmann, J. Berenike. 2017. „In a test bed with Kafka. Introducing a mixed-method approach to digital stylistics". Herausgegeben von Joris J. van Zundert, Sally Chambers, Marijn Koolen, Mike Kestemont, und Catherine Jones. *DHQ: Digital Humanities Quarterly* 11 (4). http://digitalhumanities.org/dhq/vol/11/4/000341/000341.html .

Herrmann, J. Berenike, und Gerhard Lauer. 2018. „Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne". *Osnabrücker Beiträge zur Sprachtheorie (OBST)* 92: 127–56.

Primorac, Antonija, Rosario Arias, Roxana Patras, Eva Eglāja-Kristsone, Karina van Dalen-Oskam, Berenike Herrmann, Christof Schöch, und Pieter François. 2023. „Distant Reading Two Decades On: Reflections on the Digital Turn in the Study of Literature". *Digital Studies / Le Champ Numérique* . https://doi.org/10.16995/dscn.8855 .

Kunz, Tanja Angela, Matthias Buschmeier, Jens Ciecior, Karima Lanius und Stephanie Wollmann. 2023. „Digital Open Education Im Bachelor-Studium: Lesen Und Schreiben in LMS-Basierten Selbstlernkursen Im Bereich Der Deutschsprachigen Literaturgeschichte".

MiDU - Medien Im Deutschunterricht 5 (1): 1-16. https://doi.org/10.18716/ojs/midu/2023.1.8 .

Lauer, Gerhard. 2020. *Lesen im digitalen Zeitalter* . Darmstadt: wbg Academic.

Literaturgeschichten.de. https://literaturgeschichte-kolidi.de (zugegriffen: 28. November 2023).

Sinclair, Stéfan und Geoffrey Rockwell. 2016. *Voyant Tools* . http://voyant-tools.org/ (zugegriffen: 28. November 2023).

Distant Reading Textual AI Art Prompts

Efer, Thomas

efer@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig, Deutschland
ORCID: 0000-0002-8376-3884

Niekler, Andreas

aniekler@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig, Deutschland
ORCID: 0000-0002-3036-3318

The past months have seen a Cambrian explosion of high-quality neural image generation systems. They share the quite novel idea that users can enter textual prompts to inform the generation process according to their desired visual output. While early tools required a lot of technical expertise and the use of dedicated hardware, online services like DALL·E, Midjourney and Adobe Firefly have brought this cutting-edge technology to everyone, unlocking novel creative opportunities and allowing all kinds of people to express themselves artistically in an unparalleled visual quality.

But what and how did they express exactly in the advent of this new technology? Which subjects are popular, which artists' styles do users try to imitate, are they informed users of art-historical categories and vocabulary? How much of the prompts do genuinely originate from the user's own imagination and prompt engineering skills and how much is likely just copied from others?

Generative models for image synthesis have already been proposed in the DH as valuable means to "facilitate the exploration of the semantic structure of large image corpora" (Offert, 2020). The analysis of their widespread use "in the wild" through large collections of natural language prompts adds further layers to the understanding of many concepts in art.

For this poster we collected prompts from various sources[1] covering different diffusion models and prompting environments. Our dataset incorporates and thus surpasses in size recently published collections like the one Wang et al. (2023) provide. To anchor our analyses in the world of traditional art, we extracted artist's names from the comprehensive Getty's Union List of Artist Names as well as Wikidata and several web-published "AI Art Style Guides", shared by power users. From these sources we generate a structured representation of interrelated artists, styles, periods, and techniques allowing us to formally describe and compare the analyzed prompts. The goal is to uncover, document and evaluate presence, exchangeability and perceived stylistic surroundings of mentioned artists.

The prompts constitute a form of "natural language" expression but at the same time have the quality of simple "commands" to a computer system – seemingly unordered lists of short repetitive phrases. They are neither grammatically correct nor strictly governed by a closed set of words or phrases which complicates NLP and generally makes a deeper understanding of the semantics of prompt parts quite challenging. While short text topic models like (Yin and Wang, 2014) can be employed to capture the domain specific distributed semantics of individual words, there are specific measures to be taken in order to segment the prompts into coherent workable phrases and to extract boilerplate sections. Entity and feature recognition as well as phrase recognition must be customized, as detailed in the poster.

A graph-based analysis is then used to correlate the identified phrases with each other and to reveal clusters and interrelations. We consider co-occurrence and second-order co-occurrence (occurrence in similar contexts but not with each other) in order to highlight certain aspects of narrower and broader semantic relatedness. Ultimately, we aim to reliably identify similar prompts in the dataset that were part of a prompt engineering exploration process. From this we gain insights into the genesis of complex and increasingly sophisticated prompts and can infer the role of certain phrases for achieving specific visual goals. The image below shows a small portion of a cluster of interrelated prompts with their diverging passages:

highly detailed long-shot photo of a long hair princess in the middle of a baroque dreamy room full of renaissance furniture, cinematic lighting, intricate, 4k resolution, elegant

highly detailed long-shot photo of a long hair princess walking in a baroque dreamy room full of renaissance furniture, cinematic lighting, intricate, 4k resolution, elegant

highly detailed long-shot photo of a unique long hair princess in a baroque dreamy room full of renaissance furniture, cinematic lighting, intricate, 4k resolution, elegant

highly detailed medium shot photo of a a long flowery hair princess walking in a baroque dreamy room full of renaissance furniture, cinematic lighting, intricate, 4k resolution, elegant

highly detailed photo of an humanoid android walking in a baroque dreamy room full of renaissance furniture, cinematic lighting, intricate, 4k resolution, elegant, gold and purple

This in-depth analysis is finally enriched by overlaying the graph with our list of artist's names. For the artists we ask several high-level questions: Are they mere "synonyms" for a certain style? Are they part of a boilerplate text, copied to enhance the overall artistic impression of the piece indiscriminately? Do they have the "correct" seman-

tic surroundings regarding style, genre and motives, as art history would put them?

This contribution looks specifically into the emerging cultural technique of creating textual prompts, focusing mainly on text and network analysis. However, the readily available diffusion models such as Stable Diffusion (Rombach et al. 2021) lend themselves to experiments on image creation, therefore we also generate images in order to assess and communicate our findings.

This figure shows a collage from 12 images generated with the same seeds and parameters, only varying in one word. It shows a result of systematic experiments to validate the existence of (sometimes subtle) yet systematic stylistic effects of cohyponyms from identifiable clusters on the overall visual appearance.

Fußnoten

1. These include data sets from Kaggle, Github repositories as well as crawled data from pages like https://lexica.art and https://krea.ai – A full list will be given in the poster's fine print.

Bibliographie

Offert, Fabian and Peter Bell. 2020. *"Generative Digital Humanities" In CEUR Procedings of the CHR 2020: Workshop on Computational Humanities Research.* 202–212

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Omme. 2021. *"High-Resolution Image Synthesis with Latent Diffusion Models"*. arXiv 2112.10752

Wang, Zijie J., Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, Duen Horng Chau. *"DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models" In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.* 893–911

Yin, Jianhua and Jianyong Wang. 2014. *"A dirichlet multinomial mixture model-based approach for short text clustering"* In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 233–242

Edition historischer Patiententexte mit Präsenz im Deutschen Textarchiv und DWDS

Brolich, Nina
nina.brolich@fau.de
Friedrich-Alexander Universität Erlangen-Nürnberg, Deutschland

Schiegg, Markus
markus.schiegg@fau.de
Friedrich-Alexander Universität Erlangen-Nürnberg, Deutschland; Universität Leipzig, Deutschland
ORCID: 0000-0003-0357-3421

Wiegand, Frank
wiegand@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0002-1096-3957

Quellen

Historische Krankenakten des 19. und frühen 20. Jahrhunderts beinhalten oft selbst geschriebene Texte der damaligen Patientinnen und Patienten, vor allem nicht abgeschickte Briefe, aber auch Lebensläufe, Notizen, sowie Briefe ihrer Angehörigen. Über 4.500 dieser Texte aus unterschiedlichen Regionen Deutschlands und auch Großbritanniens wurden im Rahmen der vom Elitenetzwerk Bayern finanzierten Nachwuchsforschungsgruppe „Flexible Schreiber in der Sprachgeschichte" (2017–23) am Lehrstuhl für Germanistische Sprachwissenschaft der FAU Erlangen-Nürnberg erstmals für das *Corpus of Patient Documents* (CoPaDocs; ca. 1850–1940) erschlossen und im XML/TEI-Format aufbereitet (vgl. Schiegg 2022). Die Patiententexte erlauben es, den historischen Sprachgebrauch aller damaligen sozialen Schichten zu rekonstruieren, sodass der Quellenfund nicht nur ein Schatz für die sogenannte Sprachgeschichte ‚von unten' (vgl. Elspaß 2005), sondern auch für die Sozial- und Medizingeschichte ist. Die genaue, manuelle Transkription zahlreicher verschiedener

Handschriften könnte zudem als ‘Ground Truth’ für Handwritten-Text-Recognition-Modelle (HTR) genutzt werden.

Bislang waren Texte von Personen der allgemeinen Bevölkerung wie Handwerkern, Dienstleuten oder Bauern nicht als Korpora digital der Öffentlichkeit zugänglich. Die Textgattungen des Deutschen Textarchivs (DTA-Kernkorpus) etwa beschränken sich auf Belletristik, Gebrauchsliteratur, Wissenschaft und Zeitung. Erst 2022/23 wurde es um einige Texte weniger routinierter Schreiber erweitert – Soldatenbriefe des 18. und 19. Jahrhunderts (vgl. Hug & Neumann 2023).[3]

In den Archiven schlummern allerdings noch tausende solcher Quellen, deren enormes Potential, vor allem auch wegen ihrer schweren Zugänglichkeit, erst bruchstückhaft zunutze gemacht werden konnte. Über mehrere Jahre angelegte Erschließungsprojekte waren bisher die Voraussetzung, um sinnvoll mit diesen oft schwer lesbaren, handschriftlichen Texten arbeiten zu können. Der zweite Schritt einer Veröffentlichung der Quellen unterblieb in der Regel, vor allem auch wegen der damit verbundenen technischen Herausforderungen. Die Aufbereitung der Patiententexte im XML/TEI-Format ist allerdings eine sehr gute Voraussetzung für die hier präsentierte Online-Edition.

Workflows und technische Implementation

Die technische Implementation des CoPaDocs verfolgt – sowohl für den Workflow zur Genese der digitalen Edition als auch für die Erarbeitung der Präsentation der Forschungsdaten – einen Ansatz, der als *KISS-Prinzip* [4] bezeichnet wird. Für die dynamische Komponente der Volltextsuche über die Editionsinhalte werden die offenen Schnittstellen der Korpusinfrastruktur des Digitalen Wörterbuchs der deutschen Sprache (DWDS)[5] in Verbindung mit der Software zur Normalisierung historischer Schreibweisen des Deutschen Textarchivs (DTA)[6] genutzt.

Zur Erstellung der Online-Edition wurde das Framework TEI## entwickelt, das in seiner generischen Anlage 2023 im Projektkontext entstanden ist. Auch die Digital Humanities müssen sich im Prozess der Klimakrise stetig Fragen stellen: Stehen Technologien, die wir im Dienste der Wissenschaft betreiben, in ihrem Nutzen im Einklang mit ökologischer Nachhaltigkeit? Wie können Forschende die Kosten ressourcenintensiver Werkzeuge vorbildhaft auf das Notwendigste reduzieren?[7] Unser Set von Tools verzichtet bewusst auf einen komplexen Technologiestack: Für das Bereitstellen und die Nutzung der Forschungsdaten über eine Webseite kommen nur statische Dateien zum Einsatz. Das minimiert auch den Aufwand für einen dauerhaften Betrieb, v.a. nach Ende von Projektlaufzeiten und -finanzierungen. Diese Herangehensweise folgt dem Prinzip des Minimal Computing mit seinem Ansatz, ausschließlich Technologien zu adaptieren, die für die Entwicklung und den Betrieb eines DH-Projektes unbedingt notwendig sind. Sie reduziert zudem die Menge und Komplexität der benötigten Tools, um die Daten für die Textpräsentation und auch verschiedene Zugänge, z.B. über Listen und die Projektdokumentation, zu erzeugen und nutzt dabei etablierte, zuverlässige und in der Community breit adaptierte Plattformen wie z.B. GitHub, wo die CoPaDocs-Daten zugänglich sind.[8] Diese Dienste bieten außerdem eine Versionskontrolle, Ticketsysteme, können Workflows abbilden und ermöglichen es, generierte Inhalte selbst bereitzustellen. Zudem sind sie an Forschungsdatenrepositorien wie Zenodo[9] angebunden.[10]

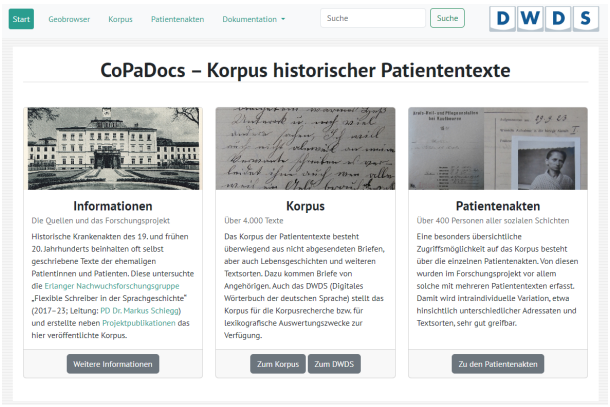


Abb. 1: Startseite CoPaDocs, http://copadocs.de.

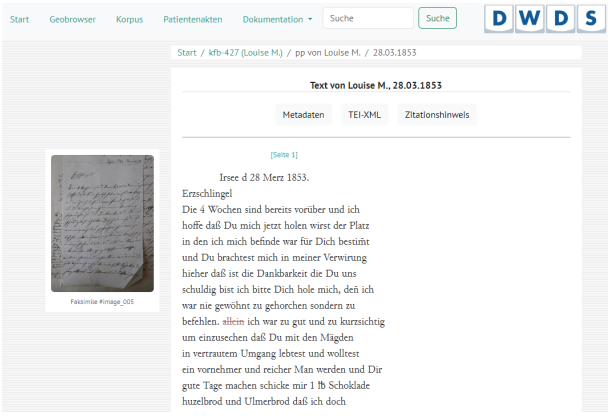


Abb. 2: Edition eines Patientenbriefs (Louise M., kfb-427-pp-1853-03-28)

Visualisierung: DARIAH-DE Geo-Browser

Zur Visualisierung kommt auf der Webseite ein etabliertes Tool der DH-Community zum Einsatz: der DARIAH-DE Geo-Browser.[11] Mittels des Python-Moduls Geopy[12] wurden hierfür zunächst den (ehemaligen) Wohnorten der Patienten semiautomatisch im TEI-XML Koordinaten zugewiesen. Mit dem Geobrowser erfolgt eine georeferenzierte Visualisierung dieser Orte, farblich differenziert nach den psychiatrischen Einrichtungen sowie den Zeiten der jeweiligen Anstaltsaufenthalte. Abb. 3 zeigt,

dass der Großteil der Patienten von CoPaDocs aus Bayerisch-Schwaben (orange) stammt und von den 1850er- bis 1930er-Jahren in der dortigen Einrichtung in Kaufbeuren-Irsee untergebracht war.

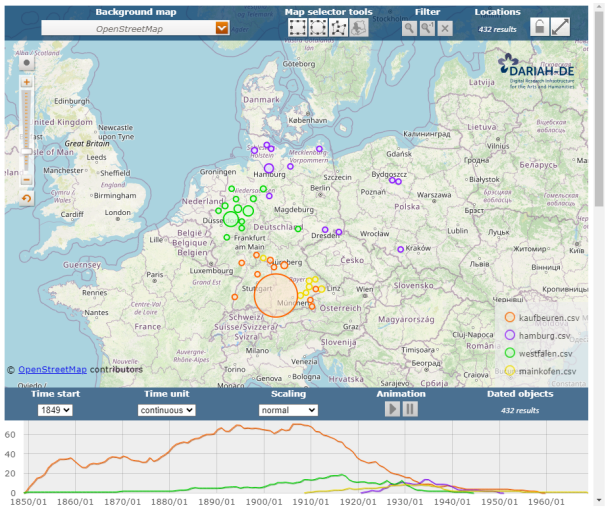


Abb. 3: DARIAH-DE Geo-Browser zur regionalen und zeitlichen Visualisierung der Daten

Poster

Das Poster illustriert anhand der *Edition historischer Patiententexte – CoPaDocs*, wie Editionsprojekte mit konsequenter Nutzung etablierter Tools und vorhandener technischer Infrastrukturen auch ohne eigenen Technologiestack über Institutionen verteilt arbeiten und ihre Ergebnisse nachnutzbar präsentieren können. Der technisch minimalistische Ansatz unterstützt dabei die Bestrebungen der DH-Community nach einer umweltfreundlichen Forschungspraxis.[13] Durch das Bereitstellen von Textbasis, Stylesheets und Tools zur Generierung der Online-Edition – als Open Source und über eine Creative-Commons-Lizenz – lassen sich alle Bausteine des CoPaDocs für Fragen der Reproduzierbarkeit und auch für eigene Weiterentwicklungen und im Kontext der NFDI im Konsortium Text+[14] nachnutzen.

Fußnoten

1. Rollen der Beitragenden nach CRediT-Taxonomie: N. Brolich (Visualization), M. Schiegg (Conceptualization, Investigation), F. Wiegand (Data Curation, Software).
2. Auf alle referenzierten URLs in diesem Dokument wurde zuletzt am 06.12.2023 zugegriffen.
3. https://www.dwds.de/d/korpora/soldatenbriefe
4. KISS = *Keep it simple (and) stupid*, vgl. https://de.wikipedia.org/wiki/KISS-Prinzip
5. Dokumentation – Export von Korpusergebnissen aus DWDS-Korpora: https://www.dwds.de/d/api#export
6. DTA::CAB – “Cascaded Analysis Broker” for error-tolerant linguistic analysis: https://kaskade.dwds.de/~moocow/software/DTA-CAB/
7. Die Arbeitsgruppe Greening DH erarbeitet Best-Practice-Richtlinien, evaluiert Methoden und Tools und fördert den Austausch zum sog. ökologischen Fußabdruck innerhalb der DH-Community: https://dig-hum.de/ag-greening-dh.
8. https://github.com/deutschestextarchiv/copadocs
9. Zenodo kann Datensätze von GitHub automatisch in sein Forschungsdatenrepositorium übernehmen und damit Daten dauerhaft verfügbar und mit persistenten Identifikatoren (DOI) nach guter wissenschaftlicher Praxis referenzierbar machen. Dies geschieht komfortabel automatisch; das CoPaDocs-Projekt ist unter dem DOI 10.5281/zenodo.10276396 verfügbar.
10. GitHub-Dokumentation: Inhalte referenzieren und zitieren https://docs.github.com/de/repositories/archiving-a-github-repository/referencing-and-citing-content
11. https://de.dariah.eu/geobrowser
12. https://github.com/geopy/geopy
13. The Digital Humanities Climate Coalition: https://www.cdcs.ed.ac.uk/digital-humanities-climate-coalition
14. NFDI-Konsortium Text+: https://text-plus.org/

Bibliographie

DHCC Measurement and Practice Action Group (Anne Alexander, Sarah Ames, James Baker, James Cummings, Racelar Ho, Leif Isaksen, Barbara McGillivray, Anna Vignoles, Jo Lindsay Walton, Jane Winters). 2022. “‘A Researcher Guide to Writing a Climate Justice-Oriented Data Management Plan’. Digital Humanities Climate Coalition Information.” DOI: 10.DHCC Measurement and Practice Action Group5281/zenodo.6451499.

Elspaß, Stephan. 2005. „ Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert.“ Tübingen: Niemeyer.

Hug, Marius und Marko Neumann. 2023. „Ressourcen-Reigen, #5: Soldatenbriefe. Integration von Forschungsdaten aus einem Promotionsprojekt“, in: „Text + Blog, 25.04.2023“. https://textplus.hypotheses.org/4815 (zugegriffen: 06.12.2023).

Risam, Roopika und Alex Gil. 2022. „ Introduction: The Questions of Minimal Computing.“ In: Digital Humanities Quarterly, 2022, vol. 6, no. 2.

Schiegg, Markus. 2022. „ Flexible Schreiber in der Sprachgeschichte. Intraindividuelle Variation in Patientenbriefen (1850–1936)“. Heidelberg: Winter. https://www.winter-verlag.de/de/detail/978-3-8253-4955-4/Schiegg_Flexible_Schreiber/ (zugegriffen: 06.12.2023).

Erkennen historischer Datierungen in den Reichstagsakten

Reinert, Matthias
matthias.reinert@hk.badw.de
Historische Kommission München, Deutschland

Historische Datierungen anhand der Heiligentage

Die im Heiligen Römischen Reich bis zum Beginn des 16. Jahrhundert vorherrschende Datierungspraxis orientierte sich am christlichen „Kirchenjahr“, an den Sonntagen, den Heiligentagen und Festen. In den digital verfügbaren Bänden der Reichstagsakten finden sie sich - anhand der unten beschriebenen Erkennungsverfahren - vor allem bis 1541, weniger in 1543 und selten in 1556/57 und 1575. Nach der Gregorianischen Kalenderreform 1582 bildeten sich je nach konfessioneller Ausrichtung der Territorien und deren individueller Übernahme der Reform parallele Datierungen alten und neuen Stils.

Reichstagsakten

Die Reichstagsakten sind ein Editionsvorhaben der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften, das seit ihrer Gründung 1858 in mittlerweile 4 Reihen vorangetrieben wurde (vgl. Annas 2021, Wolgast 2008).

Seit 2012 werden die Bände der Reihen durch ein Hybrides Editionsverfahren sowohl für den PDF-Satz als auch für eine XML-basierte Online-Fassung vorbereitet (RTA digital).

Der Arbeitskorpus besteht aus den XML-Fassungen von

- Heil, Dietmar (Bearb.) 2014. Der Reichstag zu Konstanz 1507 (Reichstagsakten Mittlere Reihe. Reichstagsakten unter Maximilian I. Band 9), München.
- Heil, Dietmar (Bearb.) 2017. Der Reichstag zu Worms 1509 (Reichstagsakten Mittlere Reihe. Reichstagsakten unter Maximilian I. Band 10), München.
- Seyboth, Reinhard (Bearb.) 2017. Die Reichstage zu Augsburg 1510 und Trier/Köln 1512 (Reichstagsakten Mittlere Reihe. Reichstagsakten unter Maximilian I. Band 11), München.
- Seyboth, Reinhard (Bearb.) 2022. Die Reichstage zu Worms 1513 und Mainz 1517 (Reichstagsakten Mittlere Reihe. Reichstagsakten unter Maximilian I. Band 12), München.
- Luttenberger, Albrecht P. (Bearb.) und Christiane Neerfeld (Druckvorbereitung) 2018. Der Reichstag zu Regensburg 1541 (Deutsche Reichstagsakten, Jüngere Reihe. Reichstagsakten unter Kaiser Karl V., XI. Band.), München.
- Schweinzer-Burian, Silvia (Bearb.) und Friedrich Edelmayer (Vorarbeiten) 2021. Der Reichstag zu Nürnberg 1543 (Deutsche Reichstagsakten, Jüngere Reihe. Reichstagsakten unter Kaiser Karl V., XIV. Band.), München.
- Leeb, Josef (Bearb.) 2013. Der Reichstag zu Regensburg 1556/57 (Deutsche Reichstagsakten. Reichsversammlungen 1556-1662), München.
- Neerfeld, Christiane (Bearb.) 2016. Der Kurfürstentag zu Regensburg 1575 (Deutsche Reichstagsakten. Reichsversammlungen 1556-1662), München.

Lokale Grammatiken

Der Korpusprozessor Unitex (Paumier 2021) erlaubt die Anwendung von speziellen Wörterbüchern in linguistische Strukturen beschreibenden Graphen. Diese können zudem in Reihe geschaltet werden (vgl. Kap. 12 CasSys in Paumier 2021, Maurel/Friburger 2011), wodurch eine Verarbeitung bestehender XML-Strukturen, dynamische Erkennung von Wörterbuch-Einträgen (Transducer) also auch serielle Anwendungen von Graphen auf einen Korpustext (Kaskaden) möglich sind.

Bereits auf der DHd 2014 wurde ein ähnlicher Ansatz (Stotz 2014) für die Erkennung von Entitäten rund um das Verb „emigrieren“ vorgestellt.

Vorgehen

- Bootstrapping der Entitäten (Tage, Heilige) = typische Struktur der Ausdrücke
- Aufbau der Wörterbücher, u.a. Extraktion Grotefend.digital
- Realtagging der Konstrukte
- Auswertung bezogen auf das Korpus
- Rewrite als Cascade: – internes Kodieren der Strukturleerstellen (mind. 2 Fassungen der Bootstrap-Graphen) – Ausgabegraph bezieht intern kodierte und extern hinterlegte Entitäten mit ein
- Entity Linking to Grotefend.digital

1. Bootstrapping Heiligentage & Bezeichner

Relevante Entitäten der Datierungen sind offenbar der Wochentag bzw. die Anzahl der Tage (N) vor oder nach („post“/„ante“[pa]) einem Fixtag. Für die Fixtage kommen Heiligen(namens)tage (SAINT) oder Feiertage (FERIA) in

Frage. Diese alle können in verschiedenen Schreibweisen und Sprachen (Latein, Deutsch) vorkommen.

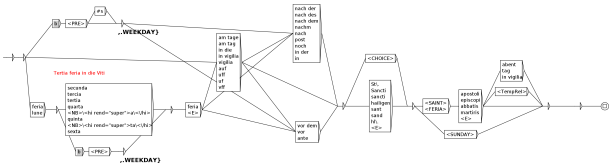


Abbildung 1: Graph zur Bestimmung von Wochentagen. Ein sog. Bootstrapping versucht zunächst eine grobe Struktur der Art „N pa SAINT“ bzw. „N pa FERIA“ abzubilden und der Graph wird auf die *Ausgabe* der auf die Leerstellen passenden Ausdrücke ausgelegt.

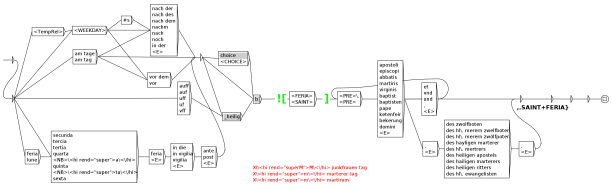


Abbildung 2: Graph zur Bestimmung von Heiligentagen. Die geschweiften Klammern kodieren die gefundenen Matches wiederum als Wörterbucheinträge für die folgenden Transducer-Schritte.

2. Aufbau der Wörterbücher

Für die im Bootstrap-Graphen gefunden Entitäten-Bezeichner ist ein manueller Kontrollgang nötig um ggf. Fehlerkennungen aus dem Wörterbuch zu entfernen. U.U. wird der Strukturgraph erweitert.

Eine weitere Gruppe Wörterbücher für Entitäten läßt sich aus dem Angebot Grotefend.digital (Raunig 2021) sowie (Grotefend 2004) extrahieren: die Bezeichner der Heiligen(-Tage) und Festtage. Diese können ebenso in einem Graph eingebunden werden, um passende Bezeichner an den strukturell vorgegebenen Stellen zu erlauben.

Anhand des Korpus der digialisierten RTA wurden Wörterbücher für die Heiligen und Feste, die temporalen Relationen (nächster, kommender etc.) sowie Wochentage in deren unterschiedlichen Schreibformen unter besonderer Kodierung der Formen für Sonntag angelegt.

3. Auswertung

Der bisherige Ansatz hat mithilfe von Bootstrapping eine korpusbezogene Entitätensammlung in den Wörterbüchern hinterlegt. Eine Auswertung der Erkennungsfähigkeit des Graphen bezieht sich auf den Recall (Auffinden der entsprechenden Datierungen) und die Precision (korrektes Tagging der Entitäten).

Zum gegenwärtigen Zeitpunkt – ohne Vergleichskorpus – kann lediglich angegeben werden, dass die Entitäten-erkennenden Graphen 935 Wochentage (eine Sonntag-Form) und 721 Heiligennamen fanden; der Ausgabegraph 4438 Datumsangaben mit 1523 Heiligennamen erkannte. Die erkannten Entitäten können in einem iterativen Verfahren wiederum den Wörterbüchern hinzugefügt werden, was den Recall theoretisch erhöhen würde.

4. Verallgemeinerung des Strukturgraphen

Der Korpusprozessor Unitex bietet die Möglichkeit in einer „Cascade“ mehrere Graphen hintereinander in unterschiedlichen Modi anzuwenden. Der Bootstrap-Graph wird so modifiziert, dass er keine Ausgabe sondern eine interne, wörterbuchanaloge Kodierung vornimmt. Die strukturellen Leerstellen können auch vereinzelt und dementsprechend mehrere Fassungen des Graphen erzeugt werden.

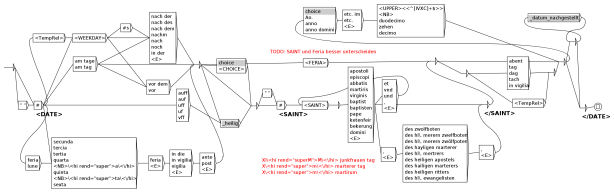


Abbildung 3: Ausgabegraph (Transducer). Die Ergebnisse können unter www.reichstagsakten.de/feriafestum überprüft und die Erkennungsgraphen ausprobiert werden.

5. Ausblick: Erkannte Entitäten sinnvoll verlinken

Naheliegend ist einerseits die Umrechnung der erkannten Zahlangaben in einen numerischen Wert. Interessanter erscheint es die besonderes Tagesangaben zunächst auf eine stabile URL zu leiten. Mithilfe der Wörterbücher läßt sich rekonstruieren, welches Lemma einer Variante zugrundeliegt.

Eine erste Sichtung ergab, dass Datierungen mit Bezug auf „Jacobi apostoli“ häufigste Heiligennennung darstellten. Beim Versuch, dies auf die Grotefend.digital - Einträge zu beziehen, zeigte sich, dass die dazugehörige Ontologie (Ontologie 2021) Heiligennamen (Jacobus / -i) und Funktion (apostoli) über die Werte in den Aussagen prefLabel (bevorzugte Bennenung) und hasFunction (hat Funktion) abbilden. Für die Bennenung prefLabel = “Jacobi” und den Wert hasFunction = “ap.” gibt es eine beinahe vollständig übereinstimmende Eintrag. Bereits bei Datierungen auf „Viti“ ergibt sich, dass die Ontologie den exakt übereinstimmenden, aber irreführenden Wert prefLabel = “Viti” (Konzept #viti_epcf_viennen) vorhält. Der inhaltlich passende Wert für prefLabel = “Viti, Modesti et Crescentie” wäre aus dem Konzept #vitimodestietcrescentie_m für eine Verlinkung herzunehmen.

Die Ableitung eines Datumwerts aus dem Grotefend muss die geografische Verortung der Datierung berücksichtigen:

gleiche Heiligennamen können in verschiedenen Orte an unterschiedlichen Tagen gefeiert worden sein. Hier könnte die geografische Angabe des edierten Stückes herangezogen werden, jedoch muss der Ausstellungsort, z.B. auf dem Reichstag, nicht mit dem konfessionellen Bezugsort des Ausstellers übereinstimmen. Dies würde das automatische Berechnen des tatsächlichen Datums noch erschweren, wenn nicht unmöglich machen.

Dank

Ich danke Dr. Dietmar Heil für Korrekturen und Hinweise zur Datierungspraxis, die sich in den Reichstagsakten findet.

Bibliographie

Annas, Gabriele. [2021]. Die „Deutschen Reichstagsakten“: Einleitung, in: Mathias Kluge (Hg.), Mittelalterliche Geschichte. Eine digitale Einführung (2021). https://mittelalterliche-geschichte.de/annas-gabriele-1

Grotefend, Hermann. 1898. Zeitrechnung des Deutschen Mittelalters und der Neuzeit. 2 Bde., Hannover 1891-1898. nach der HTML-Version von Horst Ruth. 2004. http://bilder.manuscripta-mediaevalia.de/gaeste//grotefend/grotefend.htm.

Maurel, Denis, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol-Taravella, und Damien Nouvel. 2011. „Cascades de transducteurs autour de la reconnaissance des entités nommées [CasEN: a transducer cascade to recognize French Named Entities]“. Traitement Automatique des Langues 52 (1): 69–96.

Ontologie zu „Grotefend.digital. Kalender, Feste, Heilige“. [2021] https://gams.uni-graz.at/o:grotefend.ontology.

Paumier, Sébastien (with the collaboration of Wolfgang Flury, Franz Guenthner, Eric Laporte, Friederike Malchok, Clemens Marschner, Claude Martineau, Cristian Martínez, Denis Maurel, Sebastian Nagel, Alexis Neme, Maxime Petit, Johannes Stiehler, Gilles Vollant). 2021. “Unitex 3.3. User manual”, Marne-la-Vallée. https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf.

Raunig, Elisabeth. 2021. Grotefend - digital : eine digitale Kalenderanwendung für liturgische Kalender auf Basis einer RDF-Repräsentation von Grotefends Heiligenverzeichnis und Kalendersammlung. https://unipub.uni-graz.at/urn:nbn:at:at-ubg:1-161584.

Reichstagsakten digital. 2012ff. https://www.reichstagsakten.de.

Stotz, Sophia, Valentina Stuß. 2014. “Relationsextraktion mit lokalen Grammatiken am Beispiel der Relation „emigrieren“”. Vortrag bei *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)*, Universität Passau, 25.-28. März 2014.

Wolgast, Eike. 2008. Deutsche Reichstagsakten, in: Lothar Gall (Hg.), „…für deutsche Geschichts- und Quellenforschung“. 150 Jahre Historische Kommission bei der Bayerischen Akademie der Wissenschaften, München 2008, S. 79-120. https://www.historischekommission-muenchen.de/fileadmin/user_upload/dateien/abteilungen/Deutsche_Reichstagsakten/RTA_Mittlere_Reihe/E._Wolgast_Dt._RTA_S.79_120.pdf

EU-CONEXUS Joint Master in Digital Humanities

Alvares Freire, Fernanda
fernanda.freire@uni-rostock.de
Universität Rostock, Deutschland
ORCID: 0000-0002-6414-5212

Laut dem Digital Humanities Course Registry gibt es derzeit 28 Studienangebote in den Digital Humanities in Deutschland und 168 in Europa insgesamt. Diese Zahlen umfassen sowohl Studiengänge in den Digital Humanities als auch in angrenzenden Bereichen. Obwohl es sich um ein eigenständiges Feld handelt, zeichnen sich die Digital Humanities dadurch aus, dass sie ihrem Wesen nach interdisziplinär zwischen den Geisteswissenschaften und der Informatik angesiedelt sind. Zugleich gibt es auch spezifischere Ausrichtungen der DH, die sich an einzelnen geisteswissenschaftlichen Disziplinen orientieren (Sahle 2015). So gehören auch die Digital History, die Archäoinformatik, das Literary Computing und die Computerlinguistik zum übergreifenden Feld der Digital Humanities. Bei einem so breiten Feld konzentrieren sich die Studiengänge in der Regel auf bestimmte Bereiche, nämlich auf solche, an denen die jeweilige Institution ein stärkeres Forschungsinteresse hat.

Angesichts der Interdisziplinarität und des breiten Spektrums der Digital Humanities planen wir einen internationalen Joint Master in Digital Humanities zwischen den Universitäten Rostock (Deutschland), La Rochelle (Frankreich), Zadar (Kroatien), Klaipeda (Litauen) und Valencia (Spanien). Der „Joint Master in Digital Humanities“ (JMDH) wird im Rahmen der „European University for Smart Urban Coastal Sustainability“ (EU-CONEXUS) entwickelt, deren Ziel es ist, gemeinsame europäische Module und Abschlüsse anzubieten. Alle am Joint DH Master beteiligten Universitäten sind Partner im EU-CONEXUS Projekt, das ein „Exzellenzzentrum für Smart Urban Coastal Sustainability“ (SmUCS) Forschung aufbauen will. Dabei werden auch sozial- und geisteswissenschaftliche Aspekte adressiert.

In Europa gab es bisher nur einen gemeinsamen Masterstudiengang für Digital Humanities, EuroMACHS. Mit dem EU-CONEXUS JMDH wollen wir die Expertise der beteiligten Institutionen in einer gemeinsamen Initiative

bündeln, um die interdisziplinäre Zusammenarbeit zu fördern.

Das Programm wurde unter Berücksichtigung der Überlegungen zu DH-Curricula der letzten zwei Jahrzehnte konzipiert (Brier 2012; Spiro 2012; Sahle 2015; Locke 2017). Ziel ist es, die Forschungsschwerpunkte aller Universitäten zu nutzen, um ein breites interdisziplinäres Spektrum an Kursen für Studierende mit Interesse an DH anzubieten. Das Programm ist daher an der Schnittstelle von Philosophie, Informationswissenschaft, Informatik, Literaturwissenschaft und Digital Humanities angesiedelt. Diese Ausrichtung orientiert sich an der Spezialisierung der Partnerinstitutionen.

Im DH-Umfeld gibt es bereits mehrere Studiengänge, die sowohl auf BA- als auch auf MA-Ebene weiterführende DH-Programme anbieten. Der JMDH bietet jedoch Studierenden, die bereits über einen Bachelor-Abschluss verfügen, die Möglichkeit, sich auf MA-Ebene in einem internationalen Umfeld weiterzubilden. Das Programm bedient sowohl die Interessen von Studierenden mit einem geisteswissenschaftlichen Hintergrund als auch solche mit einem Hintergrund in Informatik oder Informationswissenschaft (Smedt 2002). Es besteht aus drei Tracks: einem DH-Track, einem geisteswissenschaftlichen Track und einem IT-Track. Der IT-Track und der geisteswissenschaftliche Track zielen darauf ab, die Kompetenzen der Studierenden in ihrem jeweiligen Fachgebiet zu stärken und sie gleichzeitig in das jeweils andere Fachgebiet einzuführen. So können IT-Studierende beispielsweise ihre Fähigkeiten im Bereich Datenbanken oder Data Mining ausbauen, während sie gleichzeitig in die Methoden und Konzepte der Computerlinguistik eingeführt werden (siehe Abb. 1). Es ist vorgesehen, dass alle Studierenden am DH-Track teilnehmen, in dem die Methoden, Werkzeuge und der neueste Stand der Forschung in DH vermittelt werden. Darüber hinaus legt der Studiengang großen Wert auf die Fähigkeit zur Kooperation und Teamarbeit (Mahony und Pierazzo 2012). In diesem Sinne sollen die Studierenden am Ende des 4. Semesters an einem DH Project Lab teilnehmen, in dem sie lernen, einen Workflow in einem interdisziplinären Forschungsprojekt im Team umzusetzen (Locke 2017).

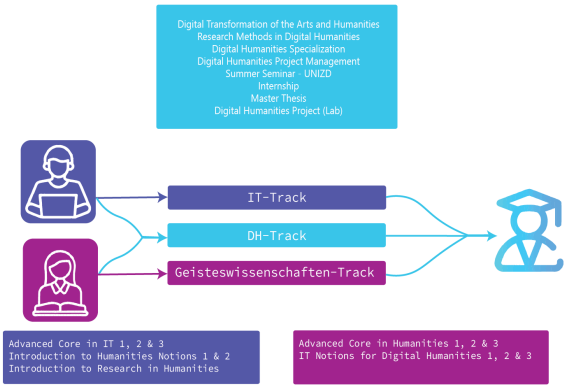


Abb. 1: visuelle Darstellung der drei Tracks des Programms.

Die Zugänglichkeit von Studienangeboten ist für uns ebenfalls ein sehr wichtiger Aspekt, und die digitale Welt bietet viele Möglichkeiten, Studierende aus der ganzen Welt mit unterschiedlichen Voraussetzungen in unser Programm zu integrieren. Aus diesem Grund wurde unser Programm von Anfang an als Hybrid konzipiert. Auf diese Weise können wir die Möglichkeiten des Zugangs zu Wissen erweitern und gleichzeitig allen, die es wünschen, die Möglichkeit einer Präsenzteilnahme bieten. Um das Gemeinschaftsgefühl und die Zugehörigkeit zu stärken, planen wir auch halbjährliche Treffen (Sommer- und Winterschulen), bei denen sich die Studierenden untereinander und mit den Dozent:innen treffen und an intensiven DH-Kursen teilnehmen können.

Der Masterstudiengang befindet sich derzeit in der Vorbereitungsphase für den Akkreditierungsprozess. In dieser Phase erstellen wir eine Übersicht über die Module, Kurse, Kosten und Dozent:innen, aus denen sich der Studiengang zusammensetzen wird. Diese Dokumente werden später verwendet, um die europäische Akkreditierung des Studiengangs zu beantragen. Der Start des Masterstudiengangs ist für das Wintersemester 2025 geplant, abhängig von der endgültigen Genehmigung und dem Abschluss des Akkreditierungsverfahrens. Die Absolvent:innen des Studiengangs erhalten einen Master of Arts, unabhängig von ihrer Vorbildung oder der von ihnen belegten Studienrichtung.

Bibliographie

Brier, Steven. 2012. "*Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities*". In *Debates in the Digital Humanities*, hg. von Matthew K. Gold. Minneapolis: University of Minnesota Press.

Locke, Brandon T. 2017. "Digital Humanities Pedagogy as Essential Liberal Education: A Framework for Curriculum Development." *Digital Humanities Quarterly* 11 (3).

Mahony, Simon, und Elena Pierazzo. 2012. "*Teaching Skills or Teaching Methodology?*" In *Digital humanities pedagogy: Practices, principles and politics* hg. von. Brett D. Hirsch, 215–226. Vereinigtes Königreich: Open Book Publishers.

Sahle, Patrick. 2015. *"Digital Humanities? Gibt's doch gar nicht!"* In *Grenzen und Möglichkeiten der Digital Humanities*, hg. von Constanze Baum und Thomas Stäcker, Herzog August Bibliothek.

Smedt, K. de. 2002. "Some Reflections on Studies in Humanities Computing". *Literary and Linguistic Computing* 10.1093/llc/17.1.89.

Spiro, Lisa. 2012. "*Opening up Digital Humanities Education*". In *Digital humanities pedagogy: Practices, principles and politics* hg. von Brett D. Hirsch, 331–364. Vereinigtes Königreich: Open Book Publishers.

„Finde den … in buddhistischen Höhlenmalereien!“ - Ein digitales Suchspiel Ein Fallbeispiel, wie Spiele dazu dienen können, die Reichweite wissenschaftlicher Projekte zu erhöhen.

Radisch, Erik
radisch@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0002-0089-9082

Konczak-Nagel, Ines
konczaknagel@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) enthaltenbeeindruckende Wandmalereien aus dem 5. bis 10. Jahrhundert. Bereits Anfangdes 20. Jahrhunderts wurden erste Hinweise auf eine frühere buddhistische Kulturentdeckt, wasmehrere Länder dazu veranlasste, Expeditionen in die Region zu entsenden, um die einst vorherrschende Religion zu erforschen. Die Entdeckung der zahlreichenbuddhistischen-Höhlenkomplexe war eine Sensation. Zu dieser Zeitwurden auch die ersten Fotografien vom Zustand der Höhlen angefertigt und Teile der Malereien aus den Höhlen entfernt, um siein die jeweiligen Nationalmuseen zu bringen. Heute sind die Fragmente der Wandmalereien auf derganzenWelt verstreut, was eine Zuordnung zu den einzelnen Ursprungshöhlen äußerstschwierig macht (weitere Informationen: Yaldiz 1987; Popova 2008; Dreyer 2015; Zhao 2021).

Eine Aufgabedeshier vorgestelltenProjektes besteht darin, die Wandmalereien in situ und die weltweit vorhandenen Einzelstücke zu dokumentieren, zu beschreiben und mithilfe historischerFotografien wieder in ihren ursprünglichen Kontext zu bringen.[1] Erleichtert wird dies durch das Vorhandensein von Digitalisaten in den verschiedenen Mus een . Um die Arbeit der Wissenschaftler zu unterstützen, soll hier bei der Bildbeschreibung auf moderne Möglichkeiten der Digital Humanities zurückgegriffen werden. Es erfolgt nicht mehr nur eine textliche Beschreibung der einzelnen Szenen, sondern ebenfalls eine Verschlagwortung innerhalb einer eigens für die Beschreibung der Kucha-Malereien entwickelten umfangreichen Taxonomiemit etwa 1.000 Einträgen, die im Unterschied zu bestehenden Kunst-Taxonomien Begriffe der buddhistischen narrativen Ikonographie enthält und somit auch für die Beschreibung anderer narrativer Darstellungen im Buddhismus verwendet werden kann. Die einzelnen Schlagworte werden mit Hilfe des Annotations-Tools Annotorious [2] direkt in den Bildern einer Segmentation zugeordnet. Annotorious ermöglicht das sehr genaue Zeichnen von Polygonen in den Bildern, um die exakten Grenzen der einzelnen taxonomischen Elemente zu definieren (Siehe Abbildung 1). Da die Bilder in diesem Projekt oft fragmentarisch sind, wurde das Tool um die Möglichkeit erweitert, ebenfalls Multipolygone in Bildern zu annotieren. Diese Praxis führte zu einer deutlichen Präzisierung der Arbeit. Denn eine detaillierte Annotation mittels Polygonen erfordert eine intensive Auseinandersetzung damit, was zu einem Objekt gehört und was nicht.

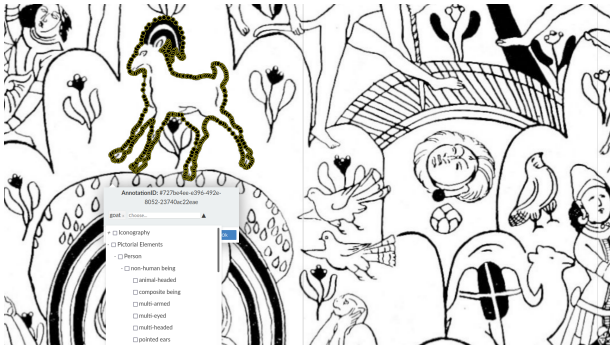


Abbildung 1: Beispiel für eine Segmentation mit Annotorious mit Hilfe unserer Taxonomie.

Bislang hat das Projekt ein beachtliches Korpus von über 13.500 Annotationen aufgebaut und leistet einen wichtigen Beitrag zur wissenschaftlichen Erschließung der buddhistischen Malereien an der nördlichen Seidenstraße. Allerdings bleiben die Erkenntnisse – ähnlich wie in vielen anderen wissenschaftlichen Projekten – auf einen kleinen akademischen Kreis beschränkt.

Um dies zu ändern und die Erkenntnisse unseres Projektes einem größeren Publikum zugänglich zu machen, wurde im Rahmen der Langen Nacht der Wissenschaften 2023 ein einfaches Spiel entwickelt, das dazu dient, die begrenzte Reichweite zu überwinden. Auf Grundlage unserer Annotationen wurde mithilfe des bereits verwendeten Annotorious und dem Component Framework Vue[3] ein Suchspiel ähnlich einer Suche nach versteckten Gegenständen in Wimmelbildern erstellt, das es interessierten Nutzern ermöglicht, sich intensiv mit dem Forschungsgegenstand auseinanderzusetzen.[4] In jedem Level wird zunächst eine kurze Einführung zu den zu findenden Objekt gegeben, was auch eine ideale Möglichkeit zur Wissensvermittlung

bietet. Der Spieler muss dann die Objekte im Bild suchen und durch ein Klicken auf dem Bereich des Objektes verifizieren. Als Feedback wird das Objekt grün hervorgehoben und der dazugehörige Zähler erhöht sich entsprechend. Fehlklicks werden im Sinne eines "Graceful Failures" nur durch ein Wackeln des Bildes sowie das Hochzählen der entsprechenden Anzeige verdeutlicht, um die Motivation während des Spielens hoch zu halten. (Ćosović/Brkić 2020) Wurden alle Objekte im Bild identifiziert, wird ein Konfet-tiregen ausgelöst. Das Design der Oberfläche ist sehr mi-nimalistisch gehalten. Es wird durch den Bildbereich do-miniert, wo die Malerei präsentiert wird, in welchen die Objekte zu finden sind. Des Weiteren gibt es über dem Bildbereich Menüpunkte, die den Zugang zur Levelwahl und zur Beschreibung garantieren. Auch findet sich hier ein statistischer Überblick über die Levelabschlüsse und eine Möglichkeit, den Spielfortschritt zurückzusetzen. Jede Le-velbeschreibung bietet außerdem die Möglichkeit einer Hil-festellung, wo die zu suchenden Bereiche leicht farblich hervorgehoben werden. Infos zum Levelfortschritt und die Anzahl der Fehlklicks werden auch in einem kleinen Over-lay rechts oben im Bildbereich zur Verfügung gestellt.

Die Eingaben des Spielers beschränken sich also im Wesentlichen auf das Finden der Objekte im Bild und folgt so-mit der Maxime, die benötigten Eingaben möglichst zu be-schränken. Auch die Navigation ist so simpel wie möglich gehalten und funktioniert – solange keine Userintervention stattfindet – automatisch. (DaCosta/ Kinsell 2023, S. 11)

Das Spiel soll ein möglichst breites Publikum (vom Grundschüler bis zum interessierten Wissenschaftler mit Hintergrundwissen in dem Bereich) ansprechen. Allen Nut-zern soll über die Einführung zu den Levels Wissen zu ver-schiedenen Aspekten der Malerei vermittelt werden. Au-ßerdem soll das Spiel sensibilisieren, wie herausfordernd die Arbeit mit den Malereien sein kann. Dazu sind die Le-vels nach Schwierigkeitsgrad gestaffelt. Die ersten Level sind eher für Grundschüler konzipiert und darauf ausgerich-tet, einfache Objekte wie Tiere in den Malereien zu iden-tifizieren. Dabei setzen sich aber auch Kinder intensiv mit den Malereien auseinander und lernen mit Abstraktionen in den Bildern umzugehen. Auch ihnen wird ein einfacher Kontext zu den gesuchten Objekten geliefert.

Höhere Levels sind für ein interessiertes adultes Publikum gedacht. Hier werden zunächst verschiedene Motive oder Charaktere der Malereien vorgestellt. Diese müssen dann auf den Wänden erkannt werden. Da gerade bei den höheren Levels auch Wende in eher schlechtem Erhaltungszustand vorkommen, ist dies teils sehr schwierig, was hilft, für die Herausforderungen zu sensibilisieren, mit denen un-sere Mitarbeiter tagtäglich konfrontiert sind.

Serious Games (also Spiele, deren Hauptziel nicht nur die reine Unterhaltung ist) als Element der Außendarstellung sind in den Digital Humanities sicherlich nicht neu. Sie bie-ten nicht nur eine hervorragende Möglichkeit zur Vermitt-lung von Wissen und stellen eine zusätzliche Chance dar, wissenschaftliche Projekte einem breiteren Publikum zu-gänglich zu machen; sie können auch der Verdeutlichung der Schwierigkeit und Komplexität der Arbeit der Projektmitarbeiter dienen. Viele Realisierungen von Spielen setzen aber auf Immersion und komplexe Video-Engines, was die Umsetzung langwierig und komplex macht.(Cesaria et al. 2020; Bontchev 2015; DaCosta 2023; Connolly et al. 2012; Backlund/Hendrix 2013)

Das hier vorgestellte Spiel ist ein perfektes Beispiel dafür, wie für ähnliche Projekte mit vielen Bildannotationen mit relativ wenig Aufwand, ein Game-Setting entwickelt werden kann, um die Forschungsdaten einem breiteren Publikum zu präsentieren.

Das Poster wird das Game-Setting und dessen Umsetzung näher erläutern sowie das fertige Spiel vorstellen (siehe Abbildungen 2–5).

Abbildung 2: Präsentation des Spiels auf der Langen Nacht der Wissenschaften.

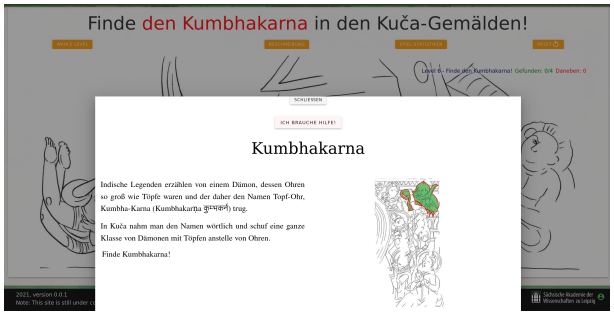


Abbildung 3: Beispiel für eine Level-Einführung. Hier wird dem Spieler anhand des Levels die indische Legendengestalt des Kumbhakarna erläutert.

Abbildung 4: Beispiel für den Spielverlauf. Die gefundenen Affen wurden grün hervorgehoben.

Abbildung 5: Beispiel für den Konfettiregen bei erfolgreichem Levelabschluss.

Fußnoten

1. www.saw-leipzig.de/hoehlenmalereien. Das Projekt hat eine eigene Publikationsreihe«Leipzig Kucha Studies», in der bereits vier Bände erschienen sind, siehe https://www.saw-leipzig.de/de/projekte/wissenschaftliche-bearbeitung-der-buddhistischen-hoehlenmalereien-in-der-kucha-region-der-noerdlichen-seidenstrasse/publications.
2. https://recogito.github.io/annotorious/. Zur Dastellung der Bilder wird OpenSeadragon verwendet: https://openseadragon.github.io/
3. https://vuejs.org/
4. https://kucha.saw-leipzig.de/game

Bibliographie

Backlund, P. and Hendrix, M. (2013). "Educational games - Are they worth the effort? A literature survey of the effectiveness of serious games" 5th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), Poole, UK, 1-8, doi: 10.1109/VS-GAMES.2013.6624226.

Bontchev, Boyan. (2015). "Serious Games for and as Cultural Heritage" In Digital Presentation and Preservation of Cultural and Scientific Heritage V: 43-58.

Cesaria, F., Cucinelli, A.M., De Prezzo, G., Spada, I. (2020). Gamification in Cultural Heritage: A Tangible User Interface Game for Learning About Local Heritage. In Digital Cultural Heritage. hg. von Kremers, H., Cham, Springer. https://doi.org/10.1007/978-3-030-15200-0_28 (zugegriffen: 05. Oktober 2023)

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T. und Boyle, J. M. (2012). “A systematic literature review of empirical evidence on computer games and serious games” In Computers & Education 59, 2: 661–686.

Ćosović, M, Brkić, Br. (2020). "Game-Based Learning in Museums—Cultural Heritage Applications" In Information; 11(1):22. https://doi.org/10.3390/info11010022.

DaCosta, B, Kinsell, C. (2023). "Serious Games in Cultural Heritage: A Review of Practices and Considerations in the Design of Location-Based Games" In Education Sciences 13(1):47. https://doi.org/10.3390/educsci13010047 (zugegriffen: 05. Oktober 2023)

Dreyer, C.(2015). *Abenteuer Seidenstraße: Die Berliner Turfan-Expeditionen 1902–1914*. Leipzig: Seemann.

Popova, I. F. (ed.) (2008). *Russian Expeditions to Central Asia at the Turn of the 20th Century: Collected Articles*. St Petersburg: Slavia Publishers.

Yaldiz, M.(1987). *Archäologie und Kunstgeschichte Chinesisch-Zentralasiens (Xinjiang)*. Leiden: Brill, Handbuch der Orientalistik, Abteilung 7, Kunst und Archäologie, Band 3, Innerasien.

Zhao, L. 赵莉 (2021) *Kezi‘er shiku bihua fuyuan yanjiu* 克孜尔石窟壁画复原研究 / *A Study on the Restoration of the Kizil Grotto Murals*. Shanghai: Shanghai shuhua chubanshe.

forTEXT-Hefte: Eine Open-Access-Plattform für den Wissensaustausch in den digitalen Literaturwissenschaften

Gerstorfer, Dominik
dominik.gerstorfer@tu-darmstadt.de
TU-Darmstadt, Deutschland
ORCID: 0000-0002-8095-2540

Akazawa, Mari
mari.akazawa@tu-darmstadt.de
TU-Darmstadt, Deutschland
ORCID: 0009-0007-2653-6275

Einführung

Die Digitalisierung hat die Geisteswissenschaften in den letzten Jahren maßgeblich beeinflusst und neue Möglichkeiten für Forschung und Lehre eröffnet. Gleichzeitig hat die Einführung des Open-Access-Modells im wissenschaft-lichen Publikationssystem nicht nur eine Neudefinition des Zugangs zu wissenschaftlichen Erkenntnissen und eine Neuausrichtung der Finanzierung des Publizierens bewirkt, sondern auch zu neuen institutionellen Organisationsfor-men und technischen Gestaltungsmöglichkeiten geführt (vgl. Wissenschaftsrat 2022). In diesem Zusammenhang geben wir im Folgenden einen Einblick in die Erweiterung des forTEXT-Portals durch die *forTEXT-Hefte* und gehen insbesondere auf den technisch-redaktionellen Workflow ein, durch den die Portalinhalte in ein Open-Access-Journal überführt werden.

Das forTEXT-Portal: Vermittlung digitaler Methoden in den digitalen Literaturwissenschaften

Das forTEXT-Projekt ist ein Disseminationsprojekt, das seit 2017 die Verbreitung digitaler Methoden in den Geisteswissenschaften vorantreibt. Es umfasst bisher drei Kernbestandteile zur Vermittlung digitaler Methodik: das Portal fortext.net, das Annotationstool CATMA (Gius et al. 2023) sowie die Dissemination in Form von Durchführung von Vorträgen und Workshops.

Im Mittelpunkt steht das zentrale Portal forTEXT.net, das systematisch lehrbezogenes Wissen, aktuelle Methodenbeschreibungen, Lern- und Lehrmaterialien sowie Informationen über Ressourcen und Tools bereitstellt. Dabei liegt der Fokus des Portals auf der Vermittlung digitaler Kompetenzen in den digitalen Literaturwissenschaften – und den textbasierten Geisteswissenschaften generell – und es dient als Anlaufstelle für Forschende, Lehrende und Interessierte mit und ohne technische Vorkenntnisse. Zusätzlich wird das forTEXT-Projekt durch die Förderung der Stiftung Innovation Hochschullehre[1] in einer weiteren Projektphase durch zwei Schwerpunkte ergänzt. Einerseits wird ein stärkerer Fokus auf die Lehre gelegt, andererseits sollen die nachhaltige Wissensbereitstellung und der Austausch der DH-Community insbesondere durch die forTEXT-Hefte vorangetrieben werden.

Die *forTEXT-Hefte*:Eine Plattform für Veröffentlichungen in den digitalen Literaturwissenschaften

Einführung der *forTEXT-Hefte* als Erweiterung des Portals

Die ab dem Dezember 2023 erscheinenden forTEXT-Hefte das Portal erweitern und eine Plattform für Veröffentlichungen von Forschungsbeiträgen im Bereich der digitalen Literaturwissenschaften bieten.

Im Portal wurden bereits über 80 Artikel veröffentlicht, die sich unter anderem aus Tool- und Methodenbeschreibungen, Lerneinheiten für autodidaktische Methodenkompetenzvermittlung, Beschreibungen von Textsammlungen und Lehrmodulen zur Vorbereitung universitärer Lehre zusammensetzen. Diese Beiträge werden nun zusätzlich in dem neuen Open-Access-Journal in thematisch gebündelten Themenheften publiziert und bereitgestellt. Durch die Erweiterung der Themenhefte um externe Einreichungen, sollen außerdem perspektivisch die Portalinhalte ergänzt werden und die Zeitschrift zu einer dynamischen Plattform für den Wissensaustausch heranwachsen. Forschende und Lehrende erhalten durch das Journal die Möglichkeit, ihre Erfahrungen, Erkenntnisse und bewährten Methoden mit der Community zu teilen.

forTEXT-Hefte: Verlagsunabhängige Veröffentlichungen und nachhaltige Verfügbarkeit

In Kooperation mit der Universitäts- und Landesbibliothek Darmstadt (ULB) als Infrastrukturpartner werden die *forTEXT-Hefte* als verlagsunabhängige Zeitschriftbetrieben. Die ULB hält die publizierten Artikel in verschiedenen Formaten wie PDF, XML, HTML und LaTeX nachhaltig vor. Über eine Kooperation mit der gemeinnützigen Open Library of Humanities (OLH)[2] stellt sie das Python-basierte Redaktionsmanagement- und Publikationssystem Janeway[3] zur Verfügung, das von der OLH als Open-Source-Software entwickelt wird (Eve und Byers, 2018).

Die Übertragung der Portalinhalte in das Janeway-System stellt eine zentrale Herausforderung dar. Die Einteilung von Lernabschnitten, die Darstellung eines Workflows, die Integration von Fallstudien oder visuellen Darstellungen, die Auffindbarkeit von externen Tools, genutzten Korpora oder Tagsets sind dabei nur einige Kernelemente, die bei der Planung der *forTEXT-Hefte* eine wesentliche Rolle spielen, um eine erfolgreiche technische Transformation sicherzustellen (vgl. Wissenschaftsrat 2022).

Technisch-Redaktioneller Workflow der *forTEXT-Hefte*

Zur Formatierung und Übertragung der Portalinhalte in das Janeway-System wird ein teilautomatisierter Ansatz umgesetzt, um die HTML-Artikel der Webseite in die Formate XML-JATS (Journal Article Tag Suite)[4] und LaTeX[5] zu übertragen (siehe Abb. 1). Diese Methode des Single-Source-Publishing soll eine konsistente Formatierung und die Nachhaltigkeit der Artikel sicherstellen. Dazu werden die Artikel im HTML-Format zur Strukturierung zunächst in Markdown überführt. Eine anschließende Umwandlung in das XML-JATS-Format gewährleistet die Kompatibilität mit anderen Systemen. Parallel dazu erfolgt die Konvertierung von Markdown in das LaTeX-PDF-Format, woraufhin eine Übertragung der formatierten Artikel und entsprechender Metadaten in das Janeway-System[6] stattfindet. Dabei werden DOIs (Digital Object Identifiers)[7] generiert, die die eindeutige und nachhaltige Identifizierbarkeit und Zitierbarkeit der Artikel gewährleisten. Ein weiterer wesentlicher Bestandteil des Workflows ist die Anbindung von Literaturverwaltungsprogrammen wie Zotero[8], die über Git[9] versioniert und in Repositorien verfügbar gemacht werden sollen. Diese Anbindung stellt eine kontrollierte Quellenbasis dar, die den Zugriff auf aktualisierte Literatur ermöglicht. Durch die Versionierung der Bibliographien über Git und die Verwendung des Janeway-Systems wird eine nachhaltige Verstetigung und langfristige Verfügbarkeit der Artikel gewährleistet. Das Projekt *forTEXT-Hefte* soll somit dazu

beitragen, Wissen und Methoden in den digitalen Literaturwissenschaften breit und nachhaltig zugänglich zu machen.

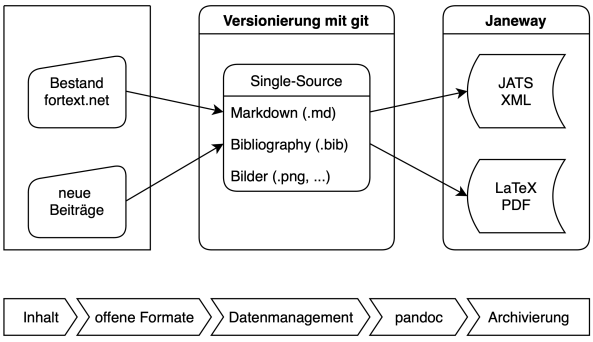


Abb. 1: forTEXT-Hefte Publikations-Workflow

Fußnoten

1. https://stiftung-hochschullehre.de/
2. https://www.openlibhums.org/
3. https://janeway.systems/
4. JATS (Journal Article Tag Suite) ist ein XML-Standard für die strukturierte Darstellung von wissenschaftlichen Artikeln, siehe https://jats.nlm.nih.gov
5. LaTeX ist ein von Leslie Lamport entwickeltes Makro-Paket für das Satzsystem TeX, siehe https://ctan.org/pkg/latex
6. Janeway ist eine open-source Journal-Plattform , siehe https://janeway.systems
7. https://www.doi.org
8. https://www.zotero.org
9. https://git-scm.com

Bibliographie

Eve, Martin Paul und Andy Byers. 2018. „Start-up Story“. Insights the UKSG Journal 31 (Mai): 15. https://doi.org/10.1629/uksg.396 .

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Mareike Schumacher, und Dominik Gerstorfer. 2023. „CATMA“. Zenodo. https://doi.org/10.5281/ZENODO.1470118 .

Wissenschaftsrat. 2022. „Empfehlungen zur Transformation des wissenschaftlichen Publizierens zu Open Access“. https://doi.org/10.57674/FYRC-VB61.

Geosemantische Kontextualisierung im Spannungsfeld domänenspezifischer Anforderungen – Methoden(kritik) der Integration von GIS und Semantic Web-Technologien

Schumacher, Anna-Lena
a.schumacher@uni-muenster.de
Institut für vergleichende Städtegeschichte, Münster, Deutschland

Runkel, Tobias
tobias.runkel@uni-muenster.de
Institut für vergleichende Städtegeschichte, Münster, Deutschland

Normann, Immanuel
immanuel.normann@uni-muenster.de
Service Center for Digital Humanities, Universität Münster, Deutschland
ORCID: 0000-0003-4702-1282

Einleitung

Geodaten und historischen Ortsdaten in den digitalen Geschichtswissenschaften werden bisher oftmals alleinig unter Verwendung von Geoinformationssystemen (GIS), zumeist v.a. zur Visualisierung der Daten, bearbeitet (Bol, 2011). Im Rahmen des Projekts HisMaComp (Historical survey maps and the comparative study of the functionality and morphology of urban space. Standardisation – Digital processing – Research) wird die historische und räumlich-topographische Entwicklung von Städten mittels eines datengetriebenen Ansatzes komparatistisch analysiert. Die Forschungsfrage dahinter lautet: Wie lassen sich Topographie, Funktionalität und Morphologie historischer Stadträume erfassen und vergleichend analysieren? Für die gemeinsame Analyse all dieser Aspekte sind GIS nicht ausreichend. Durch die Kombination von GIS und Semantic Web-Technologien lässt sich die Komplexität der zu verarbeitenden Informationen jedoch erhöhen, sodass tieferge-

hende und multidimensionale komparatistische Analysen möglich werden.

Projektkontext

Das kooperative deutsch-polnische Forschungsprojekt HiSMaComp wird von zwei Forschungsteams, am Institut für vergleichende Städtegeschichte (IStG) in Münster und an der UniwersytetMikołaja Kopernika in Toruń,, bearbeitet. Gefördert wird es im Rahmen des BEETHOVEN 4-Programms von DFG und NCN seit Juli 2022.

Sechs Städte in Polen und Deutschland dienen als Fallstudien, die auf drei Stadttypen basieren; Warschau und Magdeburg als „vormoderne Metropole", Olsztyn/Allenstein und Ochsenfurt als „mittelalterliche Stiftsstadt" sowie Inowrocław/Hohensalza und Bad Pyrmont als „moderne Kurstadt" (Szende und Szilágyi, 2019)

Kern der vergleichenden Analyse der Fallstudien sind die von Historiker:innen erhobenen Forschungsdaten, die auf schriftlichen und kartographischen Quellen basieren und mit Normdaten angereichert werden.[1] Zur Auswertung dieser Quellengrundlage kommt allerdings das Potential des GIS nicht ohne Weiteres zur Geltung, da sich basierend auf Geodaten nur Topographie und Morphologie, nicht jedoch die Funktionalität und zeitliche Entwicklung historischer Stadträume analysieren lassen (Kuhn et al., 2014; Purves et al., 2019). Um dies zu realisieren, wird eine Ontologie gebraucht, mit Hilfe derer die Daten semantisiert und kontextualisiert werden. Durch diese *geosemantische Kontextualisierung* können Geodaten auf der einen Seite und komplexe historische Kontextinformationen (semantisierte historische Ortsdaten) auf der anderen gemeinsam abgefragt und analysiert werden. Hierfür wird die „Historical Ontology of Urban Space" (HOUSe) in Teilen nachgenutzt, umstrukturiert und grundlegend erweitert. Dafür kommt die *ontology engineering*-Methode des *modular ontology modelling* zum Einsatz (Shimizu et al., 2020; Hitzler und Krisnadhi, 2018).

Anforderungsanalyse

Obwohl die HOUSe-Ontologie eine solide Grundlage darstellt, wurde sie ursprünglich für einen anderen Anwendungskontext entwickelt. Daher erfordert ihre Verwendung in unserem Projekt spezifische Anpassungen und Erweiterungen. Um diese zu bestimmen, wird eine umfassende Anforderungsanalyse durchgeführt. Diese elementare Komponente des *modular ontology modelling*, basiert auf *user stories* und der Formulierung von *competency questions*. Wir zielen dabei jedoch nicht auf die Neuentwicklung einer Ontologie, sondern auf die kritische Bewertung der bestehenden Ontologie, um so notwendige Erweiterungen und Umstrukturierungen identifizieren zu können. Hierdurch stellen wir sicher, dass die angepasste Ontologie optimal auf die spezifischen Anforderungen und Ziele unseres Projekts abgestimmt ist. Dies gewährleistet eine effektive und aussagekräftige Anwendung der HOUSe-Ontologie für unsere vergleichende Analyse der historischen Stadträume.

Ontologie und GIS

Auf dieser Grundlage werden die einzelnen Module der Ontologie iterativ entwickelt. Folgende Module sind geplant:

- topographic types (Einordnung der topographischen Objekte in Typen; z.B. Synagoge, Parkanlage)
- functions (Funktion bzw. Nutzung der Entitäten im urbanen Raum; z.B. Freizeitfunktion, Gesundheitsfunktion)
- geometries (Dient zur Einordnung des Geometrietyps; z.B. Polygon, Linie etc.)
- events (Modellierung von Veränderungen von geometries und functions über die Zeit)

Im Rahmen der o. g. geosemantischen Kontextualisierung werden die so modellierten Entitäten mit Geometrien im GIS integriert, um diese gemeinsam analysieren zu können.

Methodenkritik

Die Verbindung von GIS und Ontologien birgt sowohl Chancen als auch Herausforderungen, die es, wie auch bei anderen digitalen Methoden, kritisch zu reflektieren gilt (Dobson, 2019; Hiltmann et al., 2021). Wichtig ist es, die epistemologischen Beschränkungen unserer Forschungsdaten anzuerkennen. Historische Daten, ob Texte oder Karten, bilden nicht die Realität ab, sondern sind Konstrukte, die sich durch die subjektiven Perspektiven derer, die sie erstellt haben, und durch die dabei verwendeten Methoden verzerren können (Drucker, 2011) Jedoch arbeiten unsere beiden Zugänge – GIS und Ontologien – hier unter unterschiedlichen Voraussetzungen. So muss etwa *Unsicherheit* in beiden Ansätzen auf völlig unterschiedliche Weise modelliert werden. Zugleich bieten beide Zugänge unterschiedliche, eigentlich voneinander getrennte, Analyseperspektiven ab. All diese Aspekte müssen bei der Formulierung von *user Stories* und *competency questions* entsprechend berücksichtigt werden. Unser Ziel ist es, ein Gleichgewicht zwischen der Expansion unseres methodischen Spektrums und der Bewältigung der daraus resultierenden Komplexität zu finden

Ausblick

Auf dem geplanten Poster werden anhand eines illustrativen Fallbeispiels, einer konkreten Fallstudie, unsere Forschungsfrage, der Prozess der Anforderungsanalyse und des *ontology engineerings* sowie die damit einhergehende und nötige Methodenkritik dargestellt. Es wird beispielhaft das Forschungsinteresse einer Historiker:in von der *user story*, über die *competency questions*, die dafür nötigen Abfragen bis hin zu einem Ausschnitt aus einem Graphen repräsentativ dafür genutzt, um den Workflow und vor allem

die Modellierungsanforderungen und die Umsetzung dieser aufzuzeigen.

Fußnoten

1. Aufgrund der Granularität unseres Ansatzes berücksichtigen wir alle kartographisch darstellbaren Objekte innerhalb des urbanen Raums. Dabei konzentrieren wir uns geometrisch ausschließlich auf die räumliche Position dieser Objekte – wir 'betreten' diese nicht.

Bibliographie

Bol, Peter K. 2011. „What do humanists want? What do humanists need? What might humanists get?“ In *GeoHumanities: Art, History, Text at the Edge of Place*. London und New York: Routledge.

Dobson, James E. 2019. Critical Digital Humanities: The Search for a Methodology. 1. Aufl. Topics in the Digital Humanities Illinois Scholarship Online. Urbana: University of Illinois Press.

Drucker, Johanna. „Humanities Approaches to Graphical Display“. *Digital Humanities Quarterly* 5, Nr. 1 (2011). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Hiltmann, Torsten, Jan Keupp, Melanie Althage, und Philipp Schneider. „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* 47.1 (2021): 122--156. https://doi.org/10.13109/gege.2021.47.1.122.

Hitzler, Pascal, Adila Alfa Krisnadhi. „A Tutorial on Modular Ontology Modeling with Ontology Design Patterns: The Cooking Recipes Ontology“. *ArXiv*, 2018, 1–22.

Kuhn, Werner, Tomi Kauppinen, und Krzysztof Janowicz. 2014. „Linked Data – A Paradigm Shift for Geographic Information Science“. In *Geographic Information Science*, herausgegeben von Matt Duckham, Edzer Pebesma, Kathleen Stewart, und Andrew U. Frank, 173–86. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-11593-1_12.

Purves, Ross S., Stephan Winter, und Werner Kuhn. 2019. „Places in Information Science“. *Journal of the Association for Information Science and Technology* 70 (11): 1173–82. https://doi.org/10.1002/asi.24194.

Shimizu, Cogan, Karl Hammar, Pascal Hitzler. 2023. “Modular Ontology Modelling”. Semantic Web, vol. 14, no. 3: 459-489.

Szende, Katalin, Magdolna Szilágyi. 2019. “Town Typology in the Context of Historic Towns Atlases: a Target or a Tool?”. In Political functions of urban spaces and town types through the ages, hg, von Roman Czaja, Scheutz, Martin, Opll, Ferdinand, Noga, Zdzislaw, 267-302. Wien: Böhlau.

Glockengussdaten als Indikator für die regionale Wirtschaftsentwicklung seit dem Spätmittelalter? Eine explorative Analyse

Spoerer, Mark
mark.spoerer@ur.de
Universität Regensburg, Deutschland
ORCID: 0000-0002-5549-6512

Pößniker, Sebastian
sebastian.poessniker@ur.de
Universität Regensburg, Deutschland

Fragestellung

Seit der bahnbrechenden Untersuchung von Robert Allen (2001) hat die Frage, weshalb die Industrialisierung und in ihrer Folge der Übergang zu einem anhaltenden Wirtschaftswachstum ihren Ausgang in England nahm und nicht etwa in Kontinentaleuropa (Little Divergence Debate) oder (Süd-)Ostasien (Great Divergence Debate) weltweit neue Forschungen zur wirtschaftlichen Entwicklung angeregt.

In diesem Zusammenhang ist Deutschland ein ungewöhnlicher Fall. Dort begann trotz eines im internationalen Vergleich ausgesprochen hohen Bildungsstandards (Lindert 2004; Reis 2005) die Industrialisierung erst zaghaft in den 1830er Jahren, entwickelte sich aber vor allem ab ca. 1880 so dynamisch, dass kurz vor dem Ersten Weltkrieg das deutsche Volkseinkommen dasjenige Großbritanniens überholte (Tilly und Kopsidis 2020). Für den Zeitraum vor 1880 und insbesondere vor 1850 sind Schätzungen der wirtschaftlichen Entwicklung für Deutschland sehr schwierig und nur mit Interpolationen zu bewerkstelligen (Pfister 2020, 2022).

Ein wesentliches Hindernis für eine retrospektive Rekonstruktion des Wirtschaftswachstums, wie sie mittlerweile für andere europäische und auch einige außereuropäische Länder vorgenommen worden ist, liegt in der politischen Zersplitterung Deutschlands. Statistische Erhebungen vor der Gründung des Zollvereins 1833, wenn überhaupt durchgeführt, sind über Zeit und Raum verstreut und methodisch kaum kompatibel. Die Forschung versucht daher eine sol-

che Rekonstruktion mit der Analyse von Preisen und Löhnen, um Reallöhne errechnen zu können (Pfister 2017), oder Proxy-Variablen, wie sie etwa dem Deutschen Städtebuch entnommen werden können (Cantoni 2019-2021).

Jede dieser Variablen hat ihre Vor- und Nachteile. Zu den letzteren gehört die Tatsache, dass die vorgenannten Quellen auf in bzw. für Städte(n) erhobene(n) Daten beruhen – in einer vormodernen, dezidiert agrarisch geprägten Gesellschaft. Angesichts der vielen methodischen Schwierigkeiten wagen wir es daher hier, eine andere, völlig ungewohnte Variable auszuwerten, nämlich Glockengussdaten.

Daten

Diese haben wir Hermann Göring zu verdanken. Im Zweiten Weltkrieg führte die Produktion von Kriegsmaterial zu einem enormen Anstieg des Verbrauchs von Kupfer, der die Ressourcen des Dritten Reichs deutlich überstieg. Die Rüstungsplaner des nationalsozialistischen Regimes griffen daher auf ein geradezu klassisches und bewährtes Mittel zurück: Sie beschlossen 1940, Glocken und Denkmäler aus Bronze (76% Kupfergehalt) einzuschmelzen. Interessanterweise ließen die Behörden jedoch die kunsthistorisch oder musikalisch wertvolleren Glocken von Kunsthistoriker/inne/n erfassen. Dadurch entstanden im Rahmen einer deutschlandweiten Totalerhebung zwei Quellenbestände; zum einen eine über 9.000 Einträge umfassende Liste der sehr wertvollen, nicht für die Einschmelzung bestimmten „D-Glocken“ (knapp 10% des Gesamtbestands an Glocken) und eine über 15.000 Karteikarten umfassende Kartei der nicht ganz so wertvollen und daher für die Einschmelzung bestimmten B- und C-Glocken. Die A-Glocken wurden unseres Wissens nach nicht zentral erfasst. In diesem Paper beschäftigen wir uns mit der D-Glocken-Liste die vermutlich 1942 (oder 1943) in wenigen Exemplaren gedruckt wurde.

Sozusagen der Clou aus Sicht der Geschichtswissenschaft ist, dass in beiden Quellen (B-C-Karteikarten und D-Liste) für die meisten Glocken das Gussjahr (oder ein aufgrund kunsthistorischer Expertise geschätzter Gusszeitraum) und der Standort (im Jahre 1940) verzeichnet wurde. Beide Quellen zusammengenommen sind nicht weniger als eine Totalbestandsaufnahme aller wertvolleren Kirchenglocken im „Großdeutschen Reich“ des Jahres 1940, also bevor ein Teil der B- und C-Glocken eingeschmolzen wurde bzw. alliierte Luftangriffe viele Kirchen (und ihre auf den Türmen verbliebenen D-Glocken) zerstörten.

Wir interpretieren Glockengussdaten als Proxy-Variablen dritten Grades für die eigentlich interessierende wirtschaftliche Entwicklung. Es gilt als kaum bestritten, dass in einer vormodernen Wirtschaft, die sich in dem von Malthus beschriebenen demographischen Regime befand, ein relativ enger Zusammenhang zwischen Bevölkerungs- und Wirtschaftswachstum unterstellt werden darf. Sieht man Wirtschaftswachstum als das eigentliche Explanandum an, ist also Bevölkerungswachstum eine Ersatz- oder Proxy-Variable. Für diese wiederum, mit vielleicht noch weniger starkem Zusammenhang, wird man die Veränderung der Anzahl von Kirchen als Proxy nehmen können, und für diese wiederum den Guss von Kirchenglocken. Mit anderen Worten: regionale Glockengussaktivität wird hier als Indikator für regionales Wirtschaftswachstum interpretiert.

In der Studie beschreiben wir die Charakteristika der D-Liste und argumentieren, dass die Glockengussdaten eine wertvolle Quelle darstellen, mit denen quasi ein Fenster in die Vergangenheit geöffnet wird. Während das grundsätzlich in gewissem Sinne für jede historische Quelle gilt, liegt der besondere Wert der Glockengussdaten darin, dass sie mit geringfügigen Ausnahmen für ganz Deutschland (in Grenzen von 1940) vorhanden sind und bis ins 11. Jahrhundert zurückreichen. Der Guss einer Glocke war eine kostspielige Angelegenheit. Eine bestehende Glocke vom Kirchturm abzunehmen, sie zu transportieren und anderswo wieder aufzuhängen, war kostspielig und nicht ungefährlich. Wenn man sich vergegenwärtigt, dass es wohl kaum ein anderes Artefakt gibt, das aufgrund seiner materiellen Beschaffenheit so standorttreu und haltbar ist wie Glocken, so deutet sich der Wert dieser Quelle schon an.

Aufgrund der Einzigartigkeit der Quelle gibt es unseres Wissens weltweit kein direkt vergleichbares Projekt. Cermeño und Enflo (2018) und Buringh et al. (2020) werten Daten zu Kirchenbauten aus, haben jedoch weitaus weniger Beobachtungen. Immerhin erweist sich bei ihnen der methodische Ansatz, materielle Artefakte als Proxy für Wirtschaftswachstum in der vorstatistischen Zeit zu nutzen, als fruchtbar.

Gleichwohl unterliegt auch die Interpretation von Glockengussdaten methodischen Problemen, die im nächsten Abschnitt diskutiert werden. Im vierten und fünften Abschnitt werden die beiden DH-Komponenten dieses Projekts beschrieben. Erstens wird die D-Liste genauer beschrieben, und erklärt, wie sie mit Hilfe von kommerzieller Texterkennungssoftware semiautomatisiert in eine Excel-Tabelle übertragen worden ist. Zweitens werden die Standorte dieser Glocken anschließend mit QGIS visualisiert. In der Schlussbetrachtung diskutieren wir den Stand des Projekts und eine mögliche Ausweitung auf die Kartei der B- und C-Glocken.

Potenzial und Grenzen der Daten

Im Gegensatz zu den recht detaillierten Karteikarten der B- und C-Glocken (s. Abbildung unten) beschränkte sich die über 9.000 Glocken umfassende D-Liste auf Standort, Durchmesser, Gewicht und Gussjahr bzw. -zeitraum. Sie enthält zudem einige wenige umklassifizierte Glocken der unteren Kategorien A bis C, die mit einem Asterisk markiert wurden, und außerdem Glocken von Rathäusern und anderen nichtkirchlichen Gebäuden.

Im Paper gehen wir der Frage nach, ob Glockengussdaten mangels anderer Daten Aufschluss über die wirtschaftliche Entwicklung ergeben können. Aufgrund vorangegangener Einschmelzaktionen, insbesondere im Ersten Weltkrieg, können wir nur für den Zeitraum bis 1800 va-

lide Aussagen treffen (Glocken mit Gussjahr vor 1800 wurden im Ersten Weltkrieg grundsätzlich verschont). Aus der zeitlich-räumlichen Verteilung von Glockengussdaten vom Spätmittelalter bis 1800 lassen sich im Prinzip konjunkturhistorische Fragen aus zwei Dimensionen beantworten. Aus der Längsschnittperspektive kann man sich vorstellen, dass die Häufigkeit etwas über die Wirtschaftsdynamik einer bestimmten Region im Zeitablauf aussagt, und aus der Querschnittsperspektive lässt sich dies für den Vergleich verschiedener Regionen in einem gegebenen Zeitfenster betrachten.

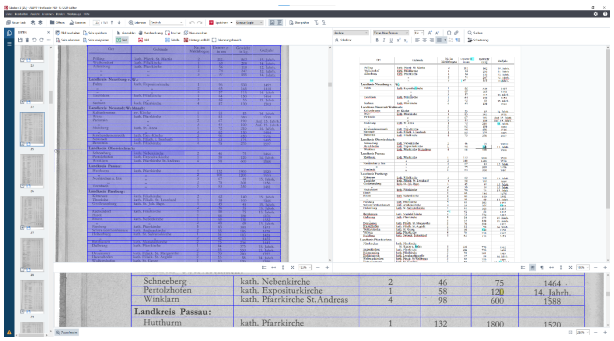
In Hinsicht auf die Längsschnittperspektive sind wir aus im Paper erläuterten Gründen, z.B. der regional ungleichmäßigen Zerstörung im 30jährigen Krieg, skeptisch oder zumindest vorsichtig. Das bescheidenere Ziel, aus dem Vergleich von Veränderungsraten unterschiedlicher Regionen Aussagen über deren relative Veränderung der Wirtschaftskraft zu machen, erscheint uns hingegen realistischer und einen Versuch wert.

Transkription und Extraktion der Daten in eine Excel-Datei

Inhaltlich stellen wir zunächst lessons learned eines an Ergebnissen mit geringem Geld- und Zeitbudget ausgestatteten Projekts beim Umgang mit gedruckten Tabellen vor. Die D-Liste ist in nur wenigen Exemplaren erhalten. Die im Bischöflichen Zentralarchiv vorhandene Ausgabe ist in konservatorisch bedenklichem Zustand, was die Digitalisierung mit konventionellen Methoden anbelangt. Jedoch konnte in der UB Passau ein Exemplar gefunden werden, das für diese Studie herangezogenen wurde. Es handelt sich hierbei um ein eng bedruckt vorliegendes Buch, das so gescannt werden musste, dass durch einen pragmatisch orientierten Workflow die benötigten Daten aus der Liste extrahiert werden konnten. Dieser im Folgenden geschilderte Workflow war zwar semiautomatisiert, jedoch lieferte er aus den erstellten Digitalisaten nahezu fehlerfreie Ergebnisse für das OCR.

Vorrangig für die Digitalisierung waren Verfüg- und Finanzierbarkeit der Geräte unter konservatorischem Vorbehalt des Bischöflichen Zentralarchivs. Dies sollte zunächst also mit einer fixierten Digitalkamera geschehen. Dieser Prozess scheiterte schließlich, da sich die doppelseitigen Aufnahmen aufgrund der Wellung in der starren Buchbindung nicht für das OCR entzerren ließen (failed dewobble), brachte aber für die anschließende OCR-Software die besseren Kontrastpunkte. Auf elf Seiten brachte das OCR eine CER von 24,6 % bei 28.606 Zeichen insbesondere dominant nahe der Buchfalz, also in der Spalte der Datierung oder rechtsseitig in der Spalte der Orte. Da insbesondere der Faktor Zeit bis hin zu weiternutzbaren Daten eine entscheidende Rolle für uns spielte, brachte schließlich erst der Scan des Exemplars der UB Passau ein qualitativ hochwertiges Digitalisat für effiziente subtasks im OCR-Workflow.

Hier nutzten wir nun einen Bookeye 5 der UB Regensburg und ließen Scans mit 300 dpi Auflösung machen. Über das kommerzielle Programm ABBYY Finereader 15, was sich finanziell tragbar gegenüber Adobe Acrobat oder OpenSource-Programmen, wie Calamary, anbot, wurde die Tabellenerkennung vorgenommen. Zudem brachte dieses Programm eine niedrigschwellige Hürde am UI, welches die Tabellenerkennung justieren, Zeilen begradigen und eine Maskenansicht mit sich, die Fehler vor dem Export (nach Excel) gut sichtbar korrigieren ließ (siehe Bild). Nach der automatischen Auftrennung der einzelnen Doppelseiten brachte ein Umweg über die Bildbearbeitung mit marginalen Eingriffen (Kontrasterhöhung; Senkung von highlights) ein gut OCR-lesbares PDF. Größte Probleme hierbei war schließlich nur, neben Druckartefakten und nicht erkannte Zellentrennungen, die auf wenigen Seiten unbefriedigende Begradigung der Textzeilen, so dass auf diese manuell zwei bis drei Tabellenmasken gezogen werden mussten. Diese mussten anschließend in Excel wieder zum ursprünglichen Layout zusammengeführt werden, was jedoch zeitlich vertretbar war. Der Weg über Abbyy Fineredear 15 brachte so bei einer CER von unter 0,3% auf 141 Seiten mit 281.000 Zeichen ein Excelsheet für die Weiterverwendung in QGIS.



Abbyy Finreader 15 bietet, im Gegensatz zu seinem Nachfolger, noch eine blau gefärbte Tabellenerkennung mit gelb angezeigten Vorschlägen für die Zellenerkennung. Als problematisch identifizierte Erkennversuche werden mit türkis auf der rechten Seite in der Exceltabellenmaske hervorgehoben.

Visualisierung mit QGIS und erste Hypothesen

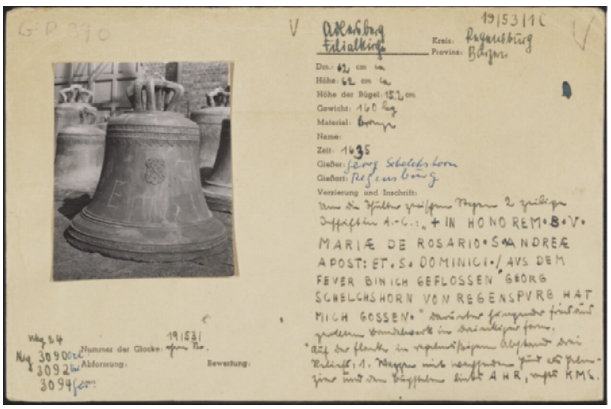
Nach Ausscheidung von Glocken niederer Kategorien bleiben etwa 7.000 valide Glockengussdaten übrig, die zeitlich und räumlich zugeordnet werden können. Um erste inhaltliche Hypothesen ableiten zu können, werden anhand der Datenkategorien Applikationskarten erstellt, für die die Glockengussdaten in größere räumliche Einheiten zusammengefasst werden. Wir haben dafür die Kreiseinteilung des Deutschen Reichs im Jahr 1940 verwendet, für die man sich das shapefile in Berlekey beim Mosaic-Projekt herunterladen kann (https://censusmosaic.demog.berkeley.edu/data/

historical-gis-files). Wenn man nun die Glockengussdaten einer Region für ein Viertel- oder halbes Jahrhundert aggregiert, erhält man genügend Beobachtungen einer Raum-Zeit-Einheit, die im Vergleich mit anderen eine Impression der regionalen wirtschaftlichen Entwicklung über die Zeit bietet. Für die Visualisierung haben wir uns ebenfalls aus Kostengründen für QGIS entschieden.

Zusammenfassung und Ausblick

Zum jetzigen Zeitpunkt können noch keine inhaltlichen Aussagen zur fachwissenschaftlichen Kernfrage des Projekts getroffen werden. Für uns sehr wichtig ist, dass wir eine zuverlässige Transkription der maschinenschriftlichen D-Liste in eine Tabelle bewerkstelligen konnten – dies wäre vor noch sehr wenigen Jahren nicht erreichbar gewesen. Auch die mit verhältnismäßig geringem Aufwand verbundene kartographische Visualisierung der Glockengussdaten ist erst seit einigen Jahren in dieser Form möglich.

Mittelfristig planen wir, die Karteikarten der für die Einschmelzung bestimmten B- und C-Glocken semiautomatisiert zu erfassen. Dies ist eine ungleich höhere Herausforderung und wird nicht ohne den Einsatz speziell zugeschnittener KI möglich sein, weil die Karteikarten nicht alle nach demselben Schema aufgebaut wurden und die Einträge von Dutzenden unterschiedlicher Personen handschriftlich vorgenommen wurden. Die nachfolgende Karteikarte gehört noch zu den übersichtlicheren. Neben der Rubrik „Zeit“ sind für uns vor allem die den Raum betreffenden Variablen interessant. Die Nummer rechts oben wurde nach einheitlichen Kriterien vergeben, so steht die 19 für Bayern und die 53 für den Stadtkreis Regensburg.



Die Einbeziehung der B- und C-Kartei wird den Datensatz in etwa verdreifachen.

Bibliographie

Allen, Robert C. 2001. “The Great Divergence in European Wages and Prices from the Middle Ages to the First World War”. In *Explorations in Economic History* 38: 411–447.

Buringh, Eltko, Bruce Campbell, Auke Rijpma; Jan Luiten van Zanden. 2020. „Church Building and the Economy during Europe’s ‘Age of the Cathedrals’, 700-1500”. In *Explorations in Economic History* 76. https://doi.org/10.1016/j.eeh.2019.101316,1 (zugegriffen: 01. Dezember 2023).

Cantoni, Davide et al. 2019-2021. „ *Princes and Townspeople. „A Collection of Historical Statistics on German Territories and Cities”*. 6 Bände, o.O.: Harvard Datavers.

Cermeño, Alexandra L., Kerstin Enflo. 2018. “Building up Faith: the Relationship between Local Wealth and Church Investments in Medieval Sweden”. University of Lund: Working Paper.

Lindert, Peter H. 2004. „ *Growing Public: Social Spending and Economic Growth Since the Eighteenth Century”*. Cambridge: Cambridge University Press.

Pfister, Ulrich. 2017. „The timing and Pattern of Real Wage Divergence in Pre-Industrial Europe: Evidence from Germany, c. 1500–1850. In *Economic History Review* 70/3: 701-729.

Pfister, Ulrich. 2020. „The Crafts–Harley view of German industrialization. An independent estimate of the income side of net national product, 1851-1913. In *European Review of Economic History* 24/3: 502–521. https://doi.org/10.1093/ereh/hez009 (zugegriffen: 01. Dezember 2023).

Pfister, Ulrich. 2022. „Economic growth in Germany, 1500-1850” In. *Journal of Economic History* 82/4: 1071–1107.

Reis, Jaime. 2005. „Economic growth, human capital formation and consumption in western Europe before 1800. In Robert C. Allen, Tommy Bengtsson, Martin Dribe (Hg.). 2005. „ *Living Standards in the Past*. Oxford: Oxford University Press: 195–225.

Tilly, Richard H., Michael Kopsidis. 2020. „ *From old regime to industrial state. A history of German industrialization from the eighteenth century to World War I.*” Chicago: University of Chicago Press Markets and governments in economic history.

GND4C@ThULB – Möglichkeiten und Grenzen der Normverdatung in Thüringer Kultureinrichtungen

Markert, Michael
michael.markert@uni-jena.de
Thüringer Universitäts- und Landesbibliothek, Deutschland

Jüngst feierte die Gemeinsame Normdatei (GND) mit Identifiern und Informationen zu den für den gesamten GLAM(Galleries, Libraries, Archives,Museums)-Sektor wichtigen Entitätstypen Personen, Körperschaften, Geografika und Sachbegriffe ihr 10-jähriges Jubiläum. Etwa 1.000 vorrangig bibliothekarische Einrichtungen pflegen die derzeit ca. 9.6 Millionen GND-Datensätze – darunter 6 Millionen Personen und über 200.000 Schlagworte (vgl. etwa Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) 2023). Umfang und Datenqualität machen die GND für viele Kultureinrichtungen zu einem zentralen Knotenpunkt für die Identifizierung von Entitäten und deren Vernetzung.

Der Einsatz der GND durch nichtbibliothekarische Stakeholder wie Forschungseinrichtungen, Museen, Universitätssammlungen oder Kommunalarchive wird durch die GND-Kooperative und vor allem die Deutsche Nationalbibliothek (DNB) seit mehreren Jahren forciert. Während dabei eine Referenzierung auf bestehende Datensätze leicht möglich ist, gibt es bei einer Einspielung eigener Daten in die GND systembedingte Hürden. So haben viele GLAM-Einrichtungen keine wissenschaftliche Bibliothek mit der entsprechenden Expertise im Haus und damit auch keinen Zugang zu einem Client für die Software WinIBW (vgl. etwa Verbundzentrale des GBV (VZG) 2023) zur manuellen Bearbeitung und Neuanlage von Datensätzen. Zudem sind die Erfassungsregeln ausgesprochen anspruchsvoll in der Anwendung und nicht auf objektbezogene Datenmodelle sowie archivalische oder museale Anforderungen ausgerichtet.

Mit dem seit 2018 laufenden DFG-Projekt GND für Kulturdaten (GND4C) (vgl. Deutsche Nationalbibliothek 2023) wird unter Leitung der DNB von Institutionen wie dem Deutschen Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg (2023) und der digiCULT-Verbund e.G. (2023) der Öffnungsprozess der GND auf verschiedenen Ebenen unterstützt, unter anderem durch die Modellierung und Erprobung von Prozessen zur Aufnahme neuer Institutionen in die GND-Community, etwa als sogenannte GND-Agenturen, die beispielsweise am Regelwerk mitarbeiten, die Bearbeitung von Datensets aus Museen, Landesdenkmalämtern und Archiven in Hinblick auf die Möglichkeiten und Grenzen des GND-Regelwerks, technische Hilfsmittel wie das DataPreparationTool des Landesarchivs Baden-Württemberg zur Generierung von MARCXML-Daten im GND-‚Flavor' (vgl. Deutsche Digitale Bibliothek 2023), die Beratung und Unterstützung von GLAM-Einrichtungen bei der Bereinigung ihrer Daten und der Anreicherung mit GND-Identifiern etwa durch Schulungen zur Arbeit mit OpenRefine (vgl. OpenRefine 2023).

Die Thüringer Universitäts- und Landesbibliothek ist in der zweiten Projektphase von Januar 2022 bis Juni 2024 an diesem Projekt beteiligt. In ihrer Doppelfunktion als Hochschul- und Landeseinrichtung unterstützt sie seit vielen Jahren sowohl kleine Regionalmuseen bei der Erschließung und Digitalisierung ihrer Bestände als auch wissenschaftliche Forschungsprojekte bei der Erzeugung, Anreicherung und Zugänglichmachung von Kulturdaten. Sie unterhält zudem das Landesportal kulthura mit derzeit ca. 1,1 Millionen digitalen Objekten aus Erfassungssystemen wie digiCULT.web (digiCULT-Verbund e.G. 2023) oder FAUST (Land Software Entwicklung 2023) und unterschiedlichsten Institutionentypen.

In GND4C betreut die ThULB zwei Arbeitspakete:

Erstens wird unter dem Namen data4kulthura (vgl. Thüringer Universitäts- und Landesbibliothek 2023a) eine regionale GND-Agentur aufgebaut. Im Einklang mit dem Landesauftrag zur Förderung der Kulturgutdigitalisierung in Thüringen (vgl. ThürBibG §3 Abs. 3) soll sie die GLAM-Institutionen des Bundeslandes in ihren Bemühungen um Standardisierung und Normverdatung unterstützen. Dies betrifft insbesondere die Bereitstellung von technischer Infrastruktur, Beratung und Support zur technischen wie auch intellektuellen Prozessierung von Datensets, GND-Redaktionsarbeit sowie Community-Vertretung in den Gremien der GND-Kooperative.

Zweitens gilt es, mit der sogenannten GND4C-Toolbox (s. Thüringer Universitäts- und Landesbibliothek 2023b) ein Software-Werkzeug für den leichteren, genaueren und effektiveren Abgleich von Kulturdaten mit der GND zu entwickeln. Implementiert sind darin neben Abfragen verschiedener Vokabular-Schnittstellen Algorithmen zur Bewertung der Match-Qualität in Bezug auf die zuvor importierten Entitäten wie Personen. In Projektphase I als PHP-basierte, reine Serverlösung konzipiert, wurde die Toolbox an der ThULB in Projektphase II auf eine Python-Basis umgestellt, um flexibel von GND-Agenturen genutzt und weiterentwickelt werden zu können – gerade mit Blick auf komplexe Entitäten wie Bauwerke.

Ausgehend davon soll mit der Posterpräsentation ein Einblick in die Normdatenbedarfe einer regionalen GLAM-Serviceeinrichtung wie der ThULB gegeben und ein Zwischenfazit dazu gezogen werden, was der Öffnungsprozess der GND und die zum Konferenzzeitpunkt 2-jährige Beteiligung an GND4C für diese Bibliothek bedeutet: Welche Normdatenarbeit findet derzeit an der ThULB statt und zu welchem Zweck? Was leisten die vorhandenen Ressourcen

und Werkzeuge und wo sind ihre Grenzen? Welche Bedarfe können (noch) nicht abgedeckt werden und was müsste dafür geschehen – auf technischer, institutioneller und struktureller Ebene?

Bibliographie

Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg. 2023. „Homepage“. https://www.uni-marburg.de/de/fotomarburg (zugegriffen: 10. November 2023).

Deutsche Digitale Bibliothek. 2023. „Data Preparation Tool“. https://github.com/Deutsche-Digitale-Bibliothek/ddblabs-datapreparationtool/releases (zugegriffen: 10. November 2023).

Deutsche Nationalbibliothek. 2023. „GND für Kulturdaten (GND4C)“. https://www.dnb.de/DE/Professionell/ProjekteKooperationen/Projekte/GND4C/gnd4c_node.html (zugegriffen: 10. November 2023).

digiCULT-Verbund e.G. 2023. „digiCULT.web“. https://www.digicult-verbund.de/de/digicultweb (zugegriffen: 10. November 2023).

Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz). 2023. „lobid-gnd“ http://lobid.org/gnd/search?q= (zugegriffen: 10. November 2023).

Land Software Entwicklung. 2023. „FAUST Standard und FAUST Professional“ http://www.land-software.de/webinfo.FAU?sid=17F5CA797&nr=00000170&art=1 (zugegriffen: 10. November 2023).

OpenRefine. 2023. „Homepage“ https://openrefine.org/ (zugegriffen: 10. November 2023).

Thüringer Universitäts- und Landesbibliothek. 2023. „data4kulthura – Servicestelle für Datenqualität und Normdaten“. https://dksm.thulb.uni-jena.de/data4kulthura/ (zugegriffen: 10. November 2023).

Thüringer Universitäts- und Landesbibliothek. 2023. „GND4C-Toolbox“. https://gnd4c.thulb.uni-jena.de/ (zugegriffen: 10. November 2023).

Verbundzentrale des GBV (VZG). 2023. „WinIBW-Handbuch“. https://wiki.k10plus.de/display/K10PLUS/WinIBW-Handbuch (zugegriffen: 10. November 2023).

InterAnnotator: Interfaces für die Annotation intertextueller Relationen

Horstmann, Jan
jan.horstmann@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0001-8047-2232

Lück, Christian
christian.lueck@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0009-0008-0649-5127

Normann, Immanuel
immanuel.normann@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0003-4702-1282

Stange, Jan-Erik
jan-erik.stange@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0009-0004-7264-2843

Intertextor: Interfaces für die Annotation intertextueller Relationen

Theoriegetriebene Überlegungen zur Annotation von Intertextualität fortführend, wird ein Entwurf zum *Intertextor*, einem Tool zur Annotation intertextueller Relationen, vorgestellt. Während bereits eine breite Palette von Annotationswerkzeugen zur Verfügung steht, ist ein Tool zur Formalisierung von Relationen zwischen Texten und Textstellen bislang ein Desiderat geblieben. Das Poster soll Intertextualitätsforschende einladen, über theoriegetriebene Toolgestaltungsprinzipien zu diskutieren sowie den Intertextor mit uns gemeinsam weiterzuentwickeln und durch Nutzung sukzessive zu verbessern.

Formalisierung als Vorarbeiten

Ausgangspunkt war eine Erhebung über verschiedene Theorien der Intertextualität seit Einführung des Begriffs 1967 durch Julia Kristeva (Kristeva 1972) und das Herausarbeiten ihres gemeinsamen konzeptionellen Kerns (Horstmann, Lück, Normann 2023 und dies. angenommen). Dieser konzeptionelle Kern und einflussreiche Ausprägungen der Theorie (etwa Hypertextualität, Genette 1993), wurden in RDF/OWL formalisiert, was zu einer modularen Ontologie geführt hat. Die Kern-Ontologie beschreibt ein Datenmodell, dessen zentrale Relation die intertextuelle Relation ist: Sie ist eine gerichtete Relation von einem späteren Text bzw. einer Textstelle auf einen Avant-Text bzw. eine Stelle darin; sie verfügt über weitere Properties, insbesondere ist sie entsprechend einer theoretischen Ausprägung klassifizierbar (z.B. als Travestie im Sinne Genettes) und die Anknüpfungsmittel können beschrieben werden (z.B. Übernahme eines Motivs). Zur Anwendung kommt zudem die Web-Annotations-Ontologie zur Referenzierung von Text-

stellen (Sanderson et al. 2017). Auf dieser Grundlage steht der *Intertextor* als Tool für die Annotation von Intertextualität auf verschiedenen Granularitätsstufen.

Technisch gesehen ist der *Intertextor* ein Mensch-Maschine-Interface einer Graph-Datenbank und dient zur Erzeugung und Darstellung von RDF-Relationen aufgrund von User-Interaktionen. Seine Komponenten müssen u.a. folgende Funktionen implementieren: Anlegen, Anzeigen und Durchsuchen von Texten bzw. Korpora; Selektieren von Textstellen und Darstellung solcher Selektionen; Anlegen und Darstellen von intertextuellen Relationen; Klassifizieren und Diskutieren solcher Verbindungen; Erstellung und Darstellung von Anknüpfungsmitteln; Suche und Navigieren in diesen strukturierten Daten. Der *Intertextor* ermöglicht kollaboratives Erfassen und Erforschen von Intertextualität. Die Benutzer*innen erstellen einen globalen Intertextualitäts-Graphen. Offenheit ist daher ein grundlegendes Gestaltungsprinzip des Tools.

Skalierbare Navigation im Netzwerk der Texte als Gestaltungsprinzip

Während in unseren Vorarbeiten formale Methoden und das Datenmodell im Vordergrund standen, folgt die weitere Entwicklung des Tools einem offenen und iterativen Co-Kreationsprozess zur Visualisierung intertextueller Relationen nebst Konzentration auf Prinzipien der User-Experience. In einem Intertextualitätsnetzwerk sind Knoten ganze Texte und Textstellen. Es ist wünschenswert, in das Netzwerk *hinein oder hinaus zoomen* zu können, um Intertextualität sowohl als Struktur als auch in ihren Details explorierbar zu machen. Eine *skalierbare Navigation* ist deswegen die Grundidee des User-Interfaces: Dargestellt wird Text je nach Zoomstufe 1) als Punkt, 2) als Balken, 3) als Fläche eines ins Unleserliche verkleinerten Textes oder 4) als lesbarer Text.

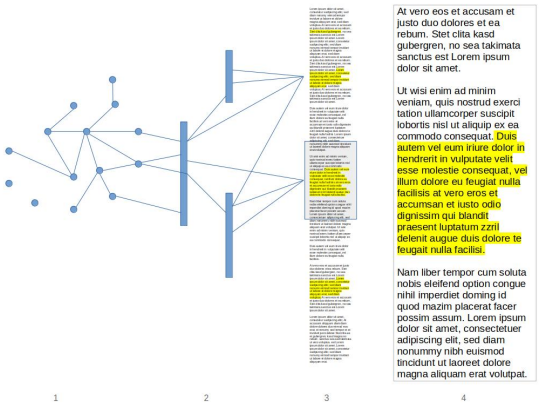


Fig. 1: Text in vier differenzierten Zoomstufen

Ausgehend von dieser Grundidee soll der *Intertextor* Forschenden verschiedene Konfigurationen der Zoomstufen als Zugänge zur Analyse und Annotation von Intertexualität anbieten:

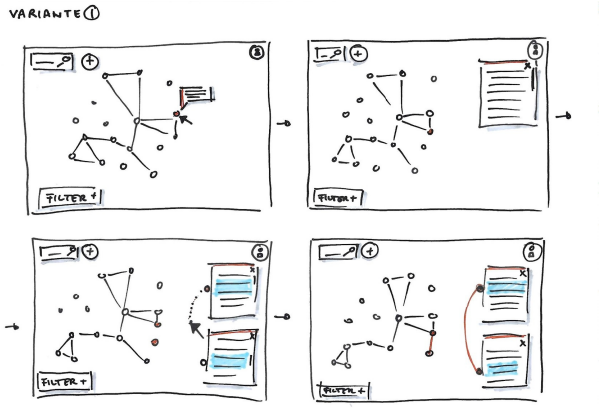


Fig. 2: Globales Netzwerk

In Fig. 2 bildet das gesamte Netzwerk den Hintergrund, auf dem sich die Interaktion mit den Texten abspielt. Einzelne Nodes können selektiert werden, woraufhin sich ein Textfenster öffnet, in dem Textstellen markiert und Verbindungen über mehrere Texte hinweg vorgenommen werden können.

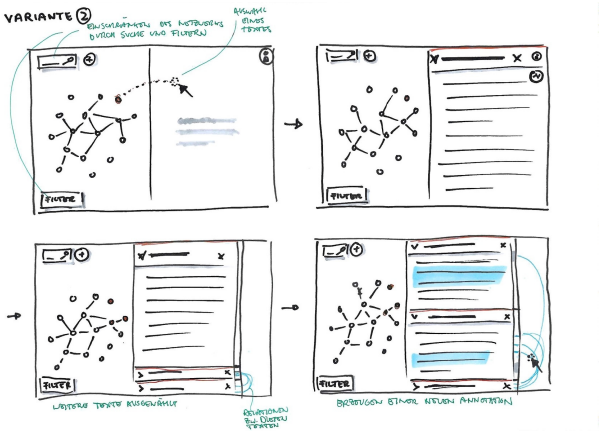


Fig. 3: Text und Netzwerk

In Fig. 3 ist der Screen in zwei Bereiche unterteilt, links das Netzwerk, rechts die Textansicht. Aus dem Netzwerk lassen sich Texte in die Textansicht herüberziehen, um mit ihnen zu arbeiten. Durch das Markieren von Textstellen in expandierten Texten können Annotationen erzeugt werden, zwischen denen wiederum Verbindungen gezogen werden können. Rechts neben den lesbaren Texten werden diese als Balken sowie ihre Verbindungen als Bogendiagramm repräsentiert, was die Navigation intertextueller Relationen zweier Texte ermöglicht. Diese Konfiguration ist besonders nützlich bei Annotation und Analyse intertextueller Rela-

tionen einzelner Textstellen. Für hermeneutische Zugänge und Close-Reading-Ansätze ist sie besonders wertvoll.

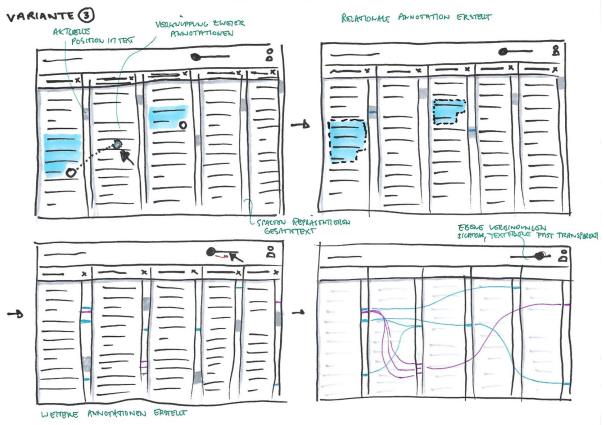


Fig. 4: Synopse

Netzwerk und der als Synopse organisierte Textbereich sind in Fig. 4 visuell stärker voneinander getrennt. In der Skizze sieht man fünf ausgewählte Texte, die jeweils eine eigene scrollbare Spalte zum Lesen haben und eine schmale, detailarme Balken-Repräsentation zur Navigation rechts daneben. Ähnlich wie in den anderen Varianten lassen sich parallel in mehreren Texten Stellen markieren, die dann miteinander verbunden werden können, um intertextuelle Relationen zu erfassen. Auch diese synoptische Konfiguration ist insbesondere für Close-Reading-Methoden geeignet.

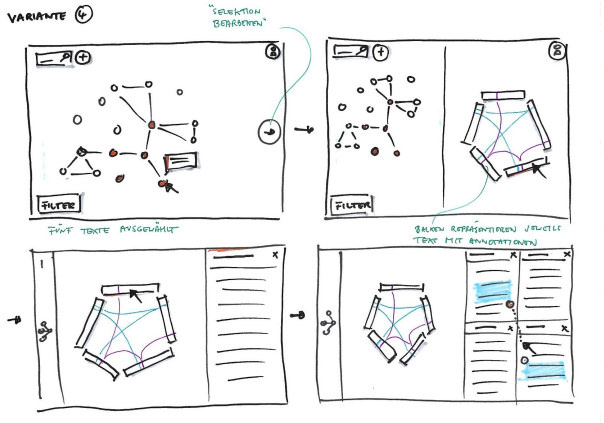


Fig 5: Korpus-Netzwerk

Ähnlich wie in der ersten Konfiguration startet Fig. 5 mit einer bildschirmfüllenden Ansicht des globalen Netzwerks. In diesem lassen sich Nodes auswählen (oder neue Nodes hinzufügen), um ein Korpus zusammenzustellen. Die ausgewählten Texte werden in einem weiteren Bereich als Balken radial angeordnet und bereits existierende Verbindungen zwischen ihnen angezeigt. Die Selektion eines Balkens öffnet eine lesbare Textansicht rechts daneben, mit der wiederum Querverbindungen zwischen Textstellen erzeugt werden können, welche im Radialdiagramm links daneben sichtbar werden.

Bibliographie

Genette, Gerard. 1993. *Palimpseste. Die Literatur auf zweiter Stufe*. Frankfurt am Main: Suhrkamp.

Horstmann, Jan, Christian Lück und Immanuel Normann. 2023. „Textliche Relationen maschinenlesbar formalisieren: Systeme der Intertextualität“. In *DHd 2023 Open Humanities Open Culture*, hg. von Peer Trilcke und Anna Busch. 9. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum", Trier, Luxemburg. DOI: 10.5281/ zenodo .7715368.

Horstmann, Jan, Christian Lück und Immanuel Normann. Angenommen. „Systems of Intertextuality. Towards a formalization of text relations for manual annotation and automated reasoning" In *Working on and with Categories for Text Analysis: Challenges and Findings from and for Digital Humanities Practices*, hg. von Dominik Gerstorfer, Evelyn Gius und Janina Jacke. Digital Humanities Quarterly.

Kristeva, Julia. 1972. „Wort, Dialog und Roman bei Bachtin“. In *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven. Band 3: Zur linguistischen Basis der Literaturwissenschaft II*, hg. von Jens Ihwe, 345–375. Frankfurt am Main: Athenäum.

Sanderson , Robert, Paolo Ciccarese undBenjamin Young. 2017. *Web Annotation Data Model*. URL: https://www.w3.org/TR/annotation-model/ (zugegriffen: 18.07.2023).

KI und Musik in der Lehre - Ein musikwissenschaftliches Projekt der Universität des Saarlandes (UdS)

Schmolenzky, Pascal
s8plschm@stud.uni-saarland.de
Universität des Saarlandes, Deutschland

Klauk, Stephanie
s.klauk@mx.uni-saarland.de
Universität des Saarlandes, Deutschland

Einleitung

Mit der Veröffentlichung von *ChatGPT*[1] durch das US-amerikanische Unternehmen *OpenAi* im vergangenen Jahr erhielt der durch die Coronakrise ausgelöste Digitalisierungsschub in der Hochschullehre einen neuen Impuls. Als Reaktion darauf wurde an der Universität des Saarlandes die Initiative „KI in der Lehre“ ins Leben gerufen, die Teil des seit 2021 von der Stiftung Innovation in der Hochschullehre geförderten Projekts *Digital Teaching Plug-in* (*DaTa-Pin*) bildet. Herausforderungen und Möglichkeiten der Nutzung von KI sollen in innovativen Lehr- und Lernangeboten erprobt und ausgelotet, im Idealfall in Best-Practice-Ansätze überführt werden.

Die Besonderheit des von der Fachrichtung Musikwissenschaft betreuten Teilprojekts „KI und Musik“ bildet sein multimodaler Ansatz. Im Gegensatz zu vielen anderen geisteswissenschaftlichen Fächern sind dort KI-Anwendungen nicht auf Schrift oder Texte beschränkt, sondern beziehen Ton und Musik mit ein. In diesem Beitrag werden das auf zwei Semester angelegte Teilprojekt und erste Ergebnisse vorgestellt.

Projektvorstellung

Die Umsetzung des Teilprojekts „KI und Musik“ im Lehrangebot der Studiengänge Musikwissenschaft und Musikmanagement der UdS bezieht sich auf die Komponenten Textproduktion und Musikproduktion. Beide subsumieren jeweils spezifische Anwendungsszenarien und digitale Werkzeuge. Für den Bereich Textproduktion spielt die Verwendung des Sprachmodells *ChatGPT*, das Studierende unter Einhaltung verbindlicher Regeln einsetzen sollen (Spannagel, 2023), eine wesentliche Rolle. In unterschiedlichen Lehrveranstaltungen, in denen die Generierung verschiedener Textarten im Mittelpunkt steht - wissenschaftliche Hausarbeiten in Seminaren zur Musikgeschichte, Rezensionen und Musikkritiken in Übungen zum musikjournalistischen Schreiben sowie Liedtexte in Praxisveranstaltungen wie Songwriting -, sollen dafür How-to-Konzepte entwickelt werden, die teilweise auch auf die Lehre anderer geisteswissenschaftlicher Fächer übertragbar sind.

Im fachspezifischen Anwendungsbereich werden computergestützte Verfahren und KI-Tools seit Jahrzehnten zur Komposition, Produktion, Aufführung und Analyse von Musik genutzt und in Forschungsprojekten stetig weiterentwickelt. Ein systematischer Einsatz solcher generativen Modelle auf Basis symbolischer Daten oder Audiodaten in der musikwissenschaftlichen Hochschullehre lässt sich allerdings nicht beobachten (vgl. zur Verwendung im Instrumentalspiel Yu et al. 2023). Hier setzt das Projekt mit der Einbindung musikbezogener KI-Werkzeuge in Lehrveranstaltungen zur Analyse und Komposition (*EarMaster*) und Bearbeitung von Musik (*Timbre Transfer*; vgl. Gabrielli et al. 2018) an.

Erste Projektphase: Sommersemester 2023

Aus dem Bereich Musikproduktion wurde zunächst die KI-basierte App *EarMaster* in die musiktheoretische Übung Gehörbildung implementiert. Das Trainingsprogramm zum Üben und Erkennen von Intervallen und Akkorden wird individuell auf die Fähigkeiten und Defizite der Lernenden angepasst und erlaubt gleichzeitig eine stetige Dokumentation des Lernfortschritts (Martínez Villar, 2015). Die Präsenzveranstaltung wurde damit um eine asynchrone Selbstlernphase (Flipped-Classroom, vgl. Bergmann et al. 2012) erweitert, die deutlich stärker genutzt wurde als zusätzliche synchrone Lernangebote wie Tutorien. Eine Lernoptimierung durch die App lässt sich an den Prüfungsergebnissen des Sommersemesters ablesen, die im Vergleich zu Zwischenprüfungen noch ohne die digitale Übemöglichkeit signifikant besser ausgefallen sind.

Für die ersten Schritte im Bereich Textproduktion wurde eine musikjournalistische Übung ausgewählt, in der eine Filmmusik-Rezension verfasst werden sollte. Die Verwendung von *ChatGPT* hat gezeigt, wie wichtig ein qualifizierter Umgang damit ist, für den für die zweite Projektphase erste How-to-Konzepte erarbeitet wurden. Hierzu zählt das Formulieren effektiver Anfragen (vgl. Gimpel et al. 2023), die kritische Beschäftigung mit den Ergebnissen der KI sowie die Kenntnis ihrer Funktionsweise. Außerdem wurden mögliche Grenzen der Automatisierung geistiger Arbeit aufgezeigt, da das kreative Schreiben über das kreative Medium Musik offenbar besondere Herausforderungen für das KI-Tool darstellt.

Zweite Projektphase: Wintersemester 2023/24

In der Kernphase des Projekts werden belastbare Evaluationen durchgeführt und die Erfahrungen aus der ersten Projektphase konzeptionell eingebracht.

Neben flankierenden Kursen wie „Songwriting“ und „Methoden der Analyse“ werden im Seminar „Komponieren im digitalen Zeitalter: AI und Human-Computer Co-Creativity“ KI-basierte Text- und Musikproduktion zusammengeführt. In dem als Flipped-Classroom ausgerichteten Seminar sollen die Studierenden eine wissenschaftliche Hausarbeit zum Thema Musikproduktion mithilfe von *ChatGPT* und verschiedenen Recherche-Tools (*Connected Papers*[2], *Elicit*[3], *Perplexity*[4] u. a.) prozessbegleitend schreiben. Nach der Themenvergabe zu Beginn des Semesters kommen erarbeitete How-to-Konzepte und Prompting-Strategien zur Anwendung. Die inhaltliche Ausrichtung auf KI-gestützte Kompositionen schließt insbesondere auch die Möglichkeit zu eigenen Kompositionsprojekten der Studierenden ein. Dabei kommen Tools wie z. B. *Wekinator*[5], *Jukebox*[6] oder *ChatGPT* zur Generierung von MIDI-Daten zum Einsatz.

Asynchrone Lernphasen werden genutzt, um Texte zu wissenschaftlichen Fragestellungen zu generieren. In synchronen Classroom-Phasen dienen die generierten Texte als Ausgangspunkt einer vorwiegend methodischen Diskussion zwischen Studierenden und Dozierenden (vgl. Weinmann-Sandig, 2023), bei denen ein transparenter Umgang mit Ergebnissen und „Prompts“ gepflegt wird. Die Dokumentation und Evaluation der zweiten Projektphase soll im Poster präsentiert werden.

Fußnoten

1. https://openai.com/blog/chatgpt
2. https://www.connectedpapers.com/
3. https://elicit.com/
4. https://www.perplexity.ai/
5. http://www.wekinator.org/
6. https://openai.com/research/jukebox

Bibliographie

Bergmann, Jonathan und Aaron Sams. 2012. *Flip your classroom. Reach every student in every class every day*. Eugene, OR: International Society for Technology in Education.

Gabrielli, Leonardo, Carmine-Emanuele Cella, Fabio Vesperini, Diego Droghini, Emanuele Principi und Stefano Squartini. 2018. “Deep Learning for Timbre Modification and Transfer: an Evaluation Study.” In *Proceeding of the 144th Audio Engineering Society Convention*. https://www.researchgate.net/publication/330842404_Deep_Learning_for_Timbre_Modification_and_Transfer_an_Evaluation_Study (zugegriffen: 19. Juli 2023).

Gimpel, Henner, Kristina Hall, Stefan Decker, Torsten Eymann, Luis Lämmermann, Alexander Mädche, Maximilian Röglinger, et al. 2023. *Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education*. https://digital.uni-hohenheim.de/fileadmin/einrichtungen/digital/Generative_AI_and_ChatGPT_in_Higher_Education.pdf (zugegriffen: 17. Juli 2023).

Martínez Villar, Sofía. 2015. “Online Ear Training Programmes and Systems: Advantages and Disadvantages.” In *3. CEIMUS Congress*. https://sonograma.org/2015/06/ear-training-programmes-systems/ (zugegriffen 19. Juli 2023).

Spannagel, Christian. 2023. *Rules for Tools*. https://csp.uber.space/phhd/rulesfortools.pdf (zugegriffen: 17. Juli 2023).

Weinmann-Sandig, Nina. 2023. *ChatGPT – Eine Chance zur Wiederbelebung des kritischen Denkens in der Hochschullehre*. https: //hochschulforumdigitalisierung.de/de/blog/praxistest-chatgpt-weimann-sandig (zugegriffen: 17. Juli 2023).

Yu, Xiaofei, Ning Ma, Lei Zheng, Licheng Wang und Kai Wang. 2023. “Developments and Applications of Artificial Intelligence in Music Education.” In *Technologies* 11. https://www.mdpi.com/2227-7080/11/2/42 (zugegriffen: 18. Juli 2023).

Kolophone digital Digital Humanities Tools in der Anwendung

Berns, Nils
berns@ub.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland

Christ, Andreas
christ@ub.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland
ORCID: 0000-0002-3591-2355

Dahm, Margit
dahm@germsem.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland

Diebel, Richard
diebel@ub.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland
ORCID: 0000-0001-7525-7175

Klemenz, Arne
klemenz@ub.uni-kiel.de
Christian-Albrechts-Universität zu Kiel, Deutschland
ORCID: 0000-0002-3450-8748

Projektziele

Das vorgestellte DFG-geförderte Forschungsprojekt (wissenschaftliche Arbeit am Germanistischen Seminar der CAU Kiel, Datenmanagement und technische Realisierung in Kooperation mit der Universitätsbibliothek Kiel, Webseite: www.kolophone.de) hat die systematische Erfassung und Erforschung von Kolophonen in deutschsprachigen Handschriften des Mittelalters als Ziel. Kolophone sind SchreiberInnenzusätze, die zumeist am Ende von Handschriften bzw. der enthaltenen Texte stehen. Sie enthalten Informationen wie Schreibernamen, Entstehungszeit und -ort der Handschrift, sie werden aber auch zur Kommentierung der Schreibtätigkeit und der Texte genutzt und sind eine der wenigen Quellen, die Rückschlüsse auf das Text-

und Selbstverständnis der Schreibenden erlauben (Dahm, 2020). Mit dem Projekt erfolgt erstmalig eine systematische Sammlung von Kolophonen in deutschsprachigen Handschriften, die der scientific community in digitaler Form verfügbar gemacht wird. Zugleich bietet die Datenbank, die auch inhaltliche Parameter erfasst, umfassende Möglichkeiten zur quantitativen wie auch qualitativen Untersuchung von Kolophonen.

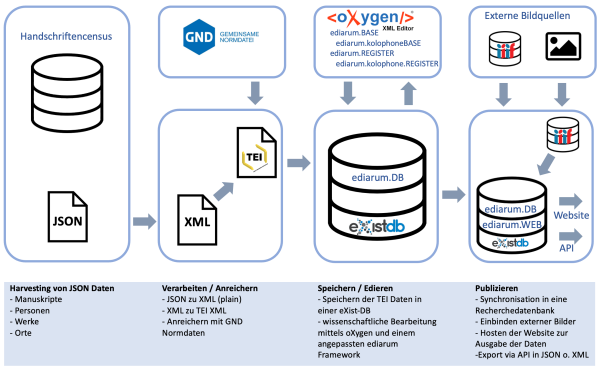
An die digitale Umsetzung des Vorhabens wurden folgende Anforderungen gestellt:

1. Zusammenführung der Transkriptionen der Kolophontexte sowie der zentralen Metadaten der jeweiligen Handschriften in einer digitalen Datensammlung nach einem semantischen und strukturierten Erfassungsschema
2. Anreicherung der Datensätze zu Kolophonen bzw. Handschriften mit weiteren Informationen wie Normdaten und Typisierungen
3. Niederschwellige, dezentrale Datenerfassung für Forschende und Implementierung von Mechanismen zur Qualitätssicherung
4. Nachnutzung existierender Datenbestände
5. Datenformat mit Anschlussfähigkeit für weitere Forschung (insb. Kodikologie und Mediävistik)
6. Open-Access-Veröffentlichung in einer Webanwendung mit Recherche- und Analysemöglichkeiten sowie eine Langzeitarchivierungslösung für die Daten

Das Projekt mit einer Laufzeit von drei Jahren ist 08/2021 gestartet, die Veröffentlichung der Datenbank wird Mitte 2024 erfolgen.

Werkzeugkasten und Daten

Zur Dokumentation der Kolophone und der zugehörigen Handschriften wurde das Dokumentenformat TEI-XML gewählt, da die P5-Guidelines mit dem Modul Manuscript Description mittlerweile den internationalen De-Facto-Standard bei der Beschreibung mittelalterlicher Handschriften darstellen (Schaßan, 2019, 201). Der mensch- und maschinenlesbare TEI-Standard ermöglicht die strukturelle und semantische Auszeichnung der erfassten Texte und stellt eine sinnvolle Grundlage für die quantitative/statistische Analyse und Durchsuchbarkeit sowie die Nachnutzung der Daten in anderen Forschungsprojekten dar. Das Datenmodell des Kolophon-Projekts orientiert sich am etablierten Vorgehen bei der Handschriftenkatalogisierung in TEI-XML (Schaßan 2019, 189–193). Über eine ODD-Datei wird XML-Schema erzeugt, in dem kontrollierte Vokabulare hinterlegt werden, so dass für viele Felder Vorschlagslisten zur Verfügung stehen (Beispiel bei Schaßan, 2017, 170–171).



Für das Projekt werden Metadaten im JSON-Format zu den händisch über Kataloge recherchierten Handschriften und zugehörige Register (Orte, Personen, Werke) aus dem Handschriftencensus (Wolf, 2009) (www.handschriftencensus.de) abgerufen. Es erfolgt eine schrittweise Konvertierung nach TEI-XML mithilfe der Python-Bibliotheken json (Python Software Foundation, 2020) und xml.etree.ElementTree (Lundh, 2008) und des Kommandozeilenwerkzeuges XMLStarlet (Grushinskiy, 2002), gefolgt von einer Anreicherung der Handschriften-Metadaten im TEI-Header durch Abruf von Normdaten (GND) über OpenRefine (Delpeuch und Huynh, 2010-2021). Anschließend werden die Daten in die Editionsdatenbank importiert. Diese Datenbank ist eine native XML-Datenbank namens eXist-db (Retter, 2022). Für die editorische Arbeit kommt der XML-Editor oXygen (Syncro Soft SRL, 2021) zum Einsatz. Die grundlegenden Funktionalitäten im Zusammenspiel der beiden zuletzt genannten Technologien und der nutzungsfreundlichen Arbeitsoberfläche liefert das Framework ediarum (Fechner et al., 2020; Fechner et al., 2021; Fechner und Dumont, 2022), welches an die speziellen Bedürfnisse des Projekts angepasst wurde. Ediarum ist eine etablierte und an komplexe editorische Anforderungen anpassbare Lösung.

Aus der Editionsdatenbank werden die XML-Dateien regelmäßig automatisch in die öffentliche Recherchedatenbank synchronisiert. Die Recherchedatenbank ist ebenfalls eine eXist-db, auf der die ebenfalls auf ediarum (Fechner, 2022) basierende, öffentliche Rechercheplattform aufsetzt. Zur Einbindung der Digitalisate wird auf den Standard IIIF gesetzt, d. h. externes Material über entsprechende Schnittstellen abgerufen und im Mirador-Viewer (Project Mirador, 2021) im Projektportal angezeigt. Die Abbildungen werden entweder von den IIIF-Image-Servern externer Repositorien oder von einem Projekt-Server basierend auf der open-source Software Cantaloupe (Cantaloupe-Project, 2021) abgerufen. Die Rechercheplattform bietet eine über das Web frei zugängliche Schnittstelle, die die Handschriftendaten im JSON-Format und als TEI-XML abrufbar macht. Ediarum ermöglicht diese Funktionalität. Die Schnittstelle steht für andere Vorhaben zur Nutzung bereit und kann bspw. dazu dienen, die Daten zum Handschriftencensus zurück zu transferieren. Eine Langzeitarchivierung

der XML-Dateien und ggf. Bilder wird mittels Datenübertragung in das TextGrid Repository (TextGrid Konsortium, 2006-2014) umgesetzt.

Fazit und Ausblick

Durch das Zusammenwirken von Kompetenzen aus germanistischer Mediävistik, Digital Humanities, Informatik und bibliothekarischem Metadatenmanagement sowie die Adaption etablierter DH-Technologien werden die Projektanforderungen effizient erfüllt. Das Vorhaben leistet einen signifikanten Beitrag zur Grundlagenforschung in der germanistischen Mediävistik und soll Ausgangspunkt weiterer literaturwissenschaftlicher und kodikologischer Forschungen zu Kolophonen als einem bislang kaum erforschten Bereich sein. Dabei werden perspektivisch verbesserte Recherche- und Visualisierungsmöglichkeiten (Karte, Zeitstrahl) sowie die Vernetzung mit weiteren Projekten zur Handschriftenerforschung wie dem Handschriftenportal angestrebt.

Bibliographie

Cantaloupe-Project. 2021. „Cantaloupe“. V. 5.0.5. University Library, University of Illinois. https://github.com/cantaloupe-project/cantaloupe (zugegriffen: 18. Juli 2023).

Dahm, Margit. 2020. „Auf den Spuren des Schreibers - Kolophone in deutschsprachigen Handschriften des Mittelalters.“ *editio - Internationales Jahrbuch für Editionswissenschaft* 34: 23-44.

Delpeuch, Antonin und David Huynh. 2010-2021. *OpenRefine*. V. 3.4.1. Ubuntu 20.04. https://github.com/OpenRefine/OpenRefine (zugegriffen: 18. Juli 2023).

Fechner, Martin, Nadine Arndt, Stefan Dumont und Sascha Grabsch. 2020. *ediarum.REGISTER.edit*. V. 1.0.3. TELOTA – BBAW. https://github.com/ediarum/ediarum.REGISTER.edit (zugegriffen: 18. Juli 2023).

Fechner, Martin, Nadine Arndt, Stefan Dumont, Sascha Grabsch und Lou Klappenbach. 2021. *ediarum.BASE.edit*. V. 2.0.0. TELOTA – BBAW. https://github.com/ediarum/ediarum.BASE.edit (zugegriffen: 18. Juli 2023).

Fechner, Martin. 2022. *ediarum.WEB*. V. 2.1.1. TELOTA – BBAW. https://github.com/ediarum/ediarum.WEB (zugegriffen: 18. Juli 2023).

Fechner, Martin und Stefan Dumont. 2022. *ediarum.DB*. V. 4.0.2. TELOTA – BBAW. https://github.com/ediarum/ediarum.DB (zugegriffen: 18. Juli 2023).

Grushinskiy, Mikhail. 2002. *XMLStarlet*. V.1.6.1. Ubuntu 20.04. https://sourceforge.net/projects/xmlstar/files (zugegriffen: 18. Juli 2023).

Lundh, Fredrik. 2008. *ElementTree*. V. 1.3.0. Python Software Foundation. Ubuntu 20.04. https://github.com/python/cpython/blob/3.9/Lib/xml/etree/ElementTree.py (zugegriffen: 18. Juli 2023).

Project Mirador. 2021. *Mirador*. V. 3.3.0. Project Mirador. https://github.com/projectmirador/mirador (zugegriffen: 18. Juli 2023).

Python Software Foundation. 2020. *JSON encoder and decoder*. V. 2.0.9. Python Software Foundation. https://github.com/python/cpython/tree/3.9/Lib/json (zugegriffen: 18.07.2023).

Retter, Adam. 2022. *eXist-db – Open Source Native XML Database*. V.6.0.1. eXist Solutions. Debian 11. https://github.com/eXist-db/exist (zugegriffen: 18. Juli 2023).

Schaßan, Torsten. 2017. „Some Roads to Script Classification: Via Taxonomy and Other Ways.“ In *Kodikologie und Paläographie im digitalen Zeitalter* 4, hg. von Hannah Busch, Franz Fischer, Patrick Sahle, 165-177. Norderstedt: BoD.

Schaßan, Torsten und Timo Steyer. 2019. „Vom lokalen Bestand zur weltweiten Vernetzung. Mittelalterliche Handschriften im Netz.“ *Das Mittelalter* 24: 188–204.

Syncro Soft SRL. 2021. *oXygen XML Editor*. V. 23.1. https://www.oxygenxml.com/ (zugegriffen: 18. Juli 2023).

TextGrid Konsortium. 2006–2014. *TextGrid: Virtuelle Forschungsumgebung für die Geisteswissenschaften*. Göttingen: TextGrid Konsortium. https://www.textgrid.de (zugegriffen: 18.07.2023).

Wolf, Jürgen. 2009 „Handschriftencensus – Eine Bestandsaufnahme.“ ZfdA 138: 279f http://www.zfda.de/beitrag.php?id=806&mode=maphilinet (zugegriffen: 18. Juli 2023).

Kompetenzprofile und Qualifikationsziele des Weiterbildenden Studiengangs Digitales Datenmanagement (DDM) an der Schnittstelle zwischen Digital Humanities, Informationswissenschaft und Data Science

Wuttke, Ulrike
ulrike.wuttke@fh-potsdam.de
Fachhochschule Potsdam, Deutschland

ORCID: 0000-0002-8217-4025

Alrez, Wassim
wassim.alrez@dainst.de
DAI, Zentrale Wissenschaftliche Dienste, Deutschland
ORCID: 0009-0001-2105-082X

Neuroth, Heike
heike.neuroth@fh-potsdam.de
Fachhochschule Potsdam, Deutschland
ORCID: 0000-0002-3637-3154

Petras, Vivien
vivien.petras@ibi.hu-berlin.de
Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin, Deutschland
ORCID: 0000-0002-8113-1509

Einleitung

Bereits seit vier Jahren wird der Weiterbildende Masterstudiengang „Digitales Datenmanagement“ (kurz DDM[1]) von der Humboldt-Universität zu Berlin und der Fachhochschule Potsdam gemeinsam angeboten. Der DDM-Master ist an der Schnittstelle zwischen Digital Humanities, Informationswissenschaft und Data Science konzipiert (Kindling und Rothfritz, 2019). Mit seiner spezialisierten und innovativen Ausrichtung auf digitales Datenmanagement erweitert er als weiterbildender Studiengang, ganz im Sinne des lebenslangen Lernens, die beruflichen Fähigkeiten der Teilnehmenden.

Der DDM-Master spricht sehr heterogene Studierendengruppen an. Die Teilnehmenden kommen aus Einrichtungen der Wissenschaft, Verwaltung, Wirtschaft und Kultur aus dem gesamten Bundesgebiet sowie dem angrenzenden Ausland und bringen berufliche Erfahrungen aus verschiedenen Arbeitsbereichen mit, z. B. Forschungsreferent*innen, Mitarbeiter*innen im Forschungsservice oder der Informationsinfrastruktur, Daten-Produzent*innen oder -Verarbeiter *innen.

Forschungsfrage

Rollen und Berufsbilder im Bereich des Datenmanagements sowie angrenzenden Bereichen mit erhöhten Bedarfen an Personal mit hohen Datenkompetenzen sind aufgrund der digitalen Transformation im Sinne von Open Science[2] und den FAIR-Prinzipien (Wilkinson et al., 2016) permanenten Wandel unterworfen und befinden sich hierdurch in der Weiterentwicklung, was die Definition von Kompetenzbereichen betrifft (Rothfritz et al., 2021). Teilnehmende aus weiterbildenden Studiengängen sind oft besonders aufgeschlossen gegenüber neuen Entwicklungen. Ihre Erfahrungen und Motivationen können wichtige Impulse für die konzeptionelle Weiterentwicklung des Studiengangs sowie seine fachliche Verortung zwischen fächerübergreifenden und fachspezifischen Anforderungen und sich erst über verschiedene Matrikel abzeichnende Trends liefern, die auch über DDM hinaus von Interesse sind. Themen, wie z. B. Forschungsdatenmanagement und Data Stewardship sind, nicht zuletzt durch die Bewilligung von verschiedenen fachspezifisch den Geisteswissenschaften zuzuordnenden NFDI-Konsortien[3], in den Digital Humanities angekommen.

Der jüngste Bericht zum sich neu entfaltenden und professionalisierenden Berufsfeld Data Stewardship versucht dieses neue Gebiet genauer zu erfassen und Handlungsempfehlungen für seine Umsetzung zu geben (Seidlmayer et al., 2023). Als Ergebnis wird eine Art Baukastensystem für Kapazitäten und Bedarfe als Grundlage verschiedener Modelle von Data Stewardship vorgeschlagen (ebd. S. 8). Ein weiteres wichtiges Feld sind Ausbildungen zur Qualifizierung von Data Stewards. Bis auf DDM stellen Qualifizierungen bezüglich Data Stewardship bislang nur Teilaspekte in den untersuchten BA/MA-Studiengängen dar. Viele der DDM-Teilnehmenden scheinen sich gezielt diesen Studiengang mit Fokus “Data Stewardship” auszuwählen, um sich ein theoretisches, informationswissenschaftliches Grundgerüst für den Umgang mit (Forschungs-)Daten (Neuroth, 2023) zu verschaffen, verbunden mit praktischen Kenntnissen der Datenaufbereitung und -verarbeitung (Data Science), situiert in verschiedenen Domänen, insbesondere konkret in den Digital Humanities (Cremer et al. 2018), wie z. B. repräsentiert in Modulen zu Metadatenstandards, aber auch in Reallaboren mit z. B. OpenRefine[4] zu erwerben, um damit Querschnittsaufgaben in fachspezifischen bzw. fachübergreifenden Kontexten zu übernehmen.

Dieses “Rezept” scheint gut aufzugehen, wie der Erfolg des Studiengangs und seiner Absolvent*innen beweist, ist jedoch bisher noch nicht systematisch untersucht. Dafür werden nun mittels der Auswertungen der vier Erstsemesterbefragungen erste Erkenntnisse geliefert, die zukünftig sukzessive um z. B. Absolvent*innenbefragungen[5] ergänzt werden sollen.

Datenbasis, Methode, Ergebnisse

Anhand der systematischen Auswertung von Daten aus den Erstsemesterbefragungen der DDM-Jahrgänge 2020 bis 2023 (quantitativ und qualitativ), die insbesondere Fragen zu Hintergrund und Motivation der Teilnehmenden enthalten (Kompetenzprofile und Qualifikationsziele), werden u. a. folgende Fragestellungen adressiert:

- Aus welchen (fachspezifischen) Bereichen bzw. Domänen, z. B. Wissenschaft, Gedächtnisinstitutionen (wie Bibliothek, Museum, Archiv), Wirtschaft oder Verwaltung kommen die Teilnehmenden?
- Welche beruflichen Vorerfahrungen mit digitalem Datenmanagement sind vorhanden?

- Welche Erwartungen und Motivation bringen sie mit?
- In welchen Themengebieten identifizieren sie ihre größten Qualifizierungslücken?
- Wie werden die Teilnehmenden vom Arbeitgeber unterstützt, z. B. monetär oder in Form von zeitlichen Freistellungen für die Lehrveranstaltungen?

Die Befragung wurde online mit LimeSurvey durchgeführt. Insgesamt nahmen über die 4 Matrikel (2020-2023) 82 von 97 Studierenden an der Befragung teil, mit einer über die Jahrgänge variierenden Rücklaufquote zwischen 63-100%. Die Auswertung der Daten zu den genannten verschiedenen Aspekten wird zusätzlich durch O-Töne angereichert. Eine zusätzliche, die Selbstevaluierung durch Studierende im ersten Semester bzw. in einem weiteren Schritt Absolvent*innen systematisch ergänzende, externe Evaluation des Studiengangs ist aus Kapazitätsgründen nur im mehrjährigen Akkreditierungsrhythmus vorgesehen und kann daher zu diesem Zeitpunkt nicht einbezogen werden.

Analyse und Ausblick

Im Mittelpunkt der Analyse steht die inhaltliche Evaluierung des Konzepts des Studiengangs, also die Frage, ob der inhaltliche Spagat zwischen fächerübergreifender und fachspezifischer Ausrichtung aufgeht bzw. welche Anpassungen notwendig sind. So scheint der weit überwiegende Anteil der Teilnehmenden einen geisteswissenschaftlichen Hintergrund zu haben und offenbar vor der Herausforderung zu stehen, sich “digital zu qualifizieren” (Allianz der deutschen Wissenschaftsorganisationen, 2020) mit besonderem Fokus auf den Umgang mit digitalen Daten. Das Poster zeigt auf, welche Impulse der weiterbildende Masterstudiengang DDM zur strategischen Weiterentwicklung der Data Stewardship gibt.

Fußnoten

1. https://ddm-master.de/
2. https://www.open-science-conference.eu/
3. https://www.nfdi.de/konsortien/
4. https://openrefine.org/
5. Erste Absolvent*innenbefragungen finden 2 Jahre nach dem Abschluss des Studiums des ersten Jahrgangs im Jahr 2024 statt.

Bibliographie

Cremer, Fabian, Lisa Klaffki und Timo Steyer. 2018. “Der Chimäre auf der Spur: Forschungsdaten in den Geisteswissenschaften,” o-bib 5, Nr. 2: 142–162, https://doi.org/10.5282/o-bib/2018H2S142-162 (zugegriffen am 18. Juli 2023).

Neuroth, Heike. 2023. “Forschungsdaten”. In Grundlagen der Informationswissenschaft, hg. von Rainer Kuhlen, Dirk Lewandowski, Wolfgang Semar, Christa Womser-Hacker, 339-349. Berlin: De Gruyter, https://doi.org/10.1515/9783110769043 (zugegriffen am 18. Juli 2023).

Rothfritz, Laura, Vivien Petras, Maxi Kindling und Heike Neuroth. 2021. “Aus- und Weiterbildung für das Forschungsdatenmanagement in Deutschland.” In Praxishandbuch Forschungsdatenmanagement, hg. von Markus Putnings, Heike Neuroth, Janna Neumann, 255-276. Berlin: DeGruyter. https://doi.org/10.1515/9783110657807-015 (zugegriffen am 18. Juli 2023).

Kindling, Maxi und Laura Rothfritz. 2019. “Data Literacy Education in der Bibliotheks- und Informationswissenschaft: Über den neuen Masterstudiengang Digitales Datenmanagement”. In Forschungsdaten – Sammeln, sichern, strukturieren. 8. Konferenz der Zentralbibliothek, Forschungszentrum Jülich, WissKom 2019, 4.-6. 06. 2019, hg. von Bernhard Mittermaier, 229-245. Jülich: Zentralbibliothek. http://hdl.handle.net/2128/22277 (zugegriffen am 18. Juli 2023).

Allianz der deutschen Wissenschaftsorganisationen. 2020. "Wege zur digitalen Qualifikation": Ein Diskussionspapier. https://doi.org/10.2312/allianzoa.038 (zugegriffen am 18. Juli 2023).

Seidlmayer, Eva, Fabian Hoffmann, Jens Dierkes, Birte Lindstädt, Ralf Depping, Konrad U. Förstner. 2023. Forschung unterstützen: Empfehlungen für Data Stewardship an akademischen Forschungsinstitutionen, ZB MED Informationszentrum Lebenswissenschaften: Köln. https://doi.org/10.4126/FRL01-006441397 (zugegriffen am 18. Juli 2023).

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, Nr. 1 (März): 160018. https://doi.org/10.1038/sdata.2016.18 (zugegriffen am 18. Juli 2023).

Kompetenzzentrum OCR – Automatische Texterkennung als Serviceangebot

Will, Larissa
larissa.will@uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID: 0009-0004-6220-8939

Huff, Dorothee
dorothee.huff@uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID: 0000-0003-0866-9967

Weil, Stefan
stefan.weil@uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID: 0000-0002-0524-9898

Kamlah, Jan
jan.kamlah@uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID: 0000-0002-0417-7562

Durchsuchbare Volltexte historischer Drucke und Handschriften bieten einen zeitgemäßen, umfassenden Zugang zum Kulturgut vergangener Epochen. Sie ermöglichen Forschenden, auf eine breite Palette von Informationen zuzugreifen und diese für ihre wissenschaftliche Arbeit zu nutzen. Darüber hinaus dienen sie als Grundlage für Anwendungen im Bereich der Data Science, da umfangreiche Textdatenbanken bereitgestellt werden können, die für Analysen und Erkenntnisse genutzt werden können (Weil und Kamlah, 2019).

Die Möglichkeiten, die die verschiedenen Texterkennungsprogramme in diesem Bereich mittlerweile bieten, sind breit, jedoch ist die Anwendung sowie die Vor- und Nachverarbeitung nicht immer intuitiv. Im Projekt OCR-BW haben die Universitätsbibliotheken Mannheim und Tübingen seit 2019 das „Kompetenzzentrum Volltexterkennung von handschriftlichen und gedruckten Werken“ aufgebaut und beraten seitdem Informationseinrichtungen und wissenschaftliche Projekte in Baden-Württemberg und darüber hinaus zu diesem Thema (Weil und Kamlah, 2020; Projektübersicht OCR-BW, 2023).

Das Kompetenzzentrum kann ein breites Know-how für unterschiedliche Programme wie z. B. Tesseract (2023), Transkribus (READ-COOP, 2023) und eScriptorium (Scripta, 2023) vorweisen. Die UB Mannheim ist zudem bis Ende 2023 mit zwei Teilprojekten an OCR-D (2023) beteiligt, wodurch auch hier Synergien entstehen (Projekte der UB Mannheim, 2023). Nach Auslaufen des Projekts OCR-BW 2022 werden die Services im Sinne der Nachhaltigkeit und aufgrund der weiterhin bestehenden Bedarfe als Teil des bibliothekarischen Portfolios fortgeführt. Durch die Volltexterkennung von Handschriften und historischen Drucken werden sowohl Forschenden neue Möglichkeiten im Umgang mit Quellen in der wissenschaftlichen Arbeit ermöglicht als auch Bibliotheken ein doppeltes Tätigkeitsfeld eröffnet (Gehrlein et. al, 2020). Neben dem Einsatz für die Bereitstellung von Volltexten zum Zweck der weiteren Erschließung von eigenen Beständen ist das Thema auch für den wissenschaftsunterstützenden Dienst einer Bibliothek relevant (Weil, 2018). Bedarf für die Verwendung von Texterkennungsprogrammen besteht nicht nur in den Geisteswissenschaften, sondern – wie sich gezeigt hat – auch für konkrete Forschungsfragen aus anderen Disziplinen. Zum einen können mithilfe von automatischer Texterkennung große Textkorpora bearbeitet werden, zum anderen wird der Zugriff auf Originalquellen auch ohne paläographische Kenntnisse erleichtert. So werden Kurrent-, Sütterlin- oder Frakturschrift in vielen geisteswissenschaftlichen Studiengängen nur rudimentär behandelt, Naturwissenschaftler*innen fehlt die paläographische Grundausbildung oftmals gänzlich.

Die Anwendung der Texterkennungssoftware und das Lesen des Quellenmaterials stellen jedoch nicht die einzigen Hürden dar, sondern auch zahlreiche andere Fragestellungen müssen im Vorfeld geklärt werden: Welchen rechtlichen Beschränkungen unterliegen die Werke? Ist die Bereitstellung von durchsuchbaren Volltexten im Einzelfall kritisch zu bewerten? Nach welchen Richtlinien werden die Texte transkribiert und Trainingsmaterial erzeugt? Wie wird mit Fehlerraten (sog. Character Error Rate oder Word Error Rate) umgegangen? Und ist die Nachnutzung des Trainingsmaterials oder sogar der Modelle möglich und wie können diese gemäß den FAIR-Prinzipien bereitgestellt werden? Wenn ja, unter welchen Einschränkungen?

Das Angebot des Kompetenzzentrums umfasst ein breites Portfolio. Neben individueller Beratung und Unterstützung werden verschiedene Dokumentationen zu Texterkennungsprogrammen sowie auch Infrastruktur für Forschende z. B. in Form einer Instanz der Texterkennungs- und Transkriptionsplattform eScriptorium zur Verfügung gestellt (eScriptorium/Universitätsbibliothek Mannheim, 2023).

Seit November 2022 bietet das Kompetenzzentrum zudem das niedrigschwellige Angebot einer offenen Sprechstunde via Zoom an (Will, 2022). Hier können sich Interessierte aus allen Bereichen mit Fragen rund um das Thema automatische Texterkennung an das Team des Kompetenzzentrums wenden. Diese Sprechstunde wird ergänzt durch eine stetig aktualisierte FAQ-Sektion auf der Projekthomepage (OCR-BW, 2023).

Auf diesem Poster soll das Serviceangebot der Universitätsbibliotheken Mannheim und Tübingen mit Fokus auf der Erzeugung FAIRer Ground-Truth-Daten vorgestellt werden. Dabei werden alle Schritte von der Datenauswahl über die Erzeugung der Ground-Truth-Daten selbst bis hin zur Veröffentlichung und Nachnutzung beleuchtet.

Bibliographie

Gehrlein, Sabine, Jan Kamlah, Matthias Pintsch, Irene Schumm und Stefan Weil. 2020. “Vom Papier zur Datenanalyse. ‘Neue’ historische Forschungsdaten für die Wirtschaftswissenschaften.” In E-Science-Tage 2019: Data to Knowledge, herausgegeben von Vincent Heuveline, 598:140–52. Heidelberg: heiBOOKS. https://doi.org/10.11588/heibooks.598.c8423.

„Home - READ-COOP“. READ-COOP. Abgerufen am 27.04.2023. https://readcoop.eu/.

„OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen“. OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen. Abgerufen am 13.07.2023. https://ocr-bw.bib.uni-mannheim.de/.

„OCR-D“. OCR-D. Abgerufen am 12.07.2023. https://ocr-d.de/.

„Projekte der UB | Universität Mannheim“. Universitätsbibliothek | Universität Mannheim. Abgerufen am 12.07.2023. https://www.bib.uni-mannheim.de/ihre-ub/projekte-der-ub/.

„Projektübersicht | OCR-BW.“ OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen. https://ocr-bw.bib.uni-mannheim.de/projektuebersicht/

„Scripta / escriptorium - GitLab“. GitLab, 27.04.2023. https://gitlab.com/scripta/escriptorium/.

„GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository)“. GitHub. Abgerufen am 12.07.2023. https://github.com/tesseract-ocr/tesseract.

Universitätsbibliothek Mannheim, „eScriptorium - Homepage“, OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen, abgerufen am 12.05.2023, https://ocr-bw.bib.uni-mannheim.de/escriptorium/.

Weil, Stefan. 2018. “126 Jahre Zeitung online - Fundgrube für historisch Interessierte und Motor für die Bibliotheks-IT: 126 years of the newspaper online.”, präsentiert bei 107. Deutscher Bibliothekartag, Berlin, Deutschland.

Weil, Stefan und Jan Kamlah. 2019. “Forschungsdaten aus Digitalisaten.” In E-Science-Tage 2019: Data to Knowledge, herausgegeben von Vincent Heuveline, 598:189. Heidelberg: heiBOOKS.

Weil, Stefan und Jan Kamlah. 2020. “OCR-BW – Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen: Texterkennung von historischen Drucken mit OCR-D und Tesseract.”, präsentiert bei Dokumentenerbe digital - Digitalisierung historischer Bestände baden-württembergischer Bibliotheken, Online.

Will, Larissa (2022, 28. Oktober). Projektende OCR-BW und 1. offene OCR-Sprechstunde | OCR-BW. OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen. https://ocr-bw.bib.uni-mannheim.de/2022/10/28/projektende-ocr-bw-und-1-offene-ocr-sprechstunde/

Ökonomien des Raums: Ein historisches Findmittel digital denken

Hodel, Tobias
tobias.hodel@unibe.ch
Universität Bern, Walter Benjamin Kolleg, Schweiz
ORCID: 0000-0002-2071-6407

Burkart, Lucas
lucas.burkart@unibas.ch
Universität Basel, Departement Geschichte, Schweiz
ORCID: 0000-0002-9011-5113

Hitz, Benjamin
benjamin.hitz@unibas.ch
Universität Basel, Departement Geschichte, Schweiz
ORCID: 0000-0002-3208-4881

Aeby, Jonas
jonas.aeby@unibas.ch
Universität Basel, Departement Geschichte, Schweiz

Prada Ziegler, Ismail
ismail.prada@unibe.ch
Universität Bern, Walter Benjamin Kolleg, Schweiz;
Universität Basel, Departement Geschichte, Schweiz
ORCID: 0000-0003-4229-8688

Vonwiller, Aline
a.vonwiller@unibas.ch
Universität Basel, Departement Geschichte, Schweiz

Die Aufbereitung historischer Daten nicht zuletzt aus historischen Findmitteln ist eine enorm gewinnbringende, jedoch auch komplexe Angelegenheit. Dabei werden Fragen zur Aufbereitung und Auswertung mit quellenkritischen Fragen vermischt. Und wenn zusätzlich das Findmittel selbst ein umfangreicher Bestand ist, müssen Unsicherheitsfaktoren, die durch die Anwendung maschineller Lernverfahren entstehen, minimiert, berücksichtigt und adressiert werden. Findmittel sind Hilfsmittel, um sich in den Beständen von Archiven zu orientieren. Häufig konzentrieren sich Findmittel auf bestimmte Bereiche von Archivbeständen.

Alle diese Herausforderungen finden sich kondensiert im Projekt “Ökonomien des Raums”, das am Beispiel der Stadt Basel das Wirtschaften mit dem städtischen Grundbesitz untersucht, das aus der ganzen Breite der schriftlichen Überlieferung im Staatsarchiv Basel-Stadt rekonstruiert wird. Darunter fallen Verzeichnisse zu Pfändungen,

Verkaufsurkunden und eine Vielzahl weiterer schriftlich überlieferter Dokumente. Greifbar wird die Überlieferung durch den Einsatz besagter *machine learning* Verfahren, sowie computergestützter Auswertungsverfahren, mit denen das Historische Grundbuch der Stadt Basel (HGB) für die digitale Nutzung aufbereitet wird.

Das Projekt und damit die Repräsentation als Poster reiht sich ein in die Ansätze zur Massenaufbereitung und Informationsextraktion, wobei der Fokus auf die Datenmodellierung (insbesondere mit Blick auf Geodaten) und auf die Extraktion und Kategorisierung von Ereignissen gelegt wird. Besonders der zweite Schritt nutzt dabei Ansätze des *deep learning* und *token*-Vektorisierung.

Das Historische Grundbuch der Stadt Basel (HGB)

Um 1900 wurden alle damals greifbaren Archivalien für das HGB in knappen Quellenauszügen exzerpiert und handschriftlich auf Zettelkarten notiert. Eine Karte steht dabei typischerweise für eine Transaktion oder eine Eintragung in einem Quellenstück. Der Inhalt des Quellenstücks wurde exzerpiert und meist in der Quellensprache auf der Zettelkarte abgeschrieben. Die Form des Exzerpts soll und kann in der digitalen Form nicht rückgängig gemacht werden. Dieser Umstand und das HGB als Findmittel ist für die Auswertungsstrategien und -methoden zentral. Damit können wir indirekt auf die zumeist integral erhaltenen Quellenbestände zurückgreifen, die wiederum über stichprobenartige Überprüfungen ausgewählter Bestände analysiert werden können.

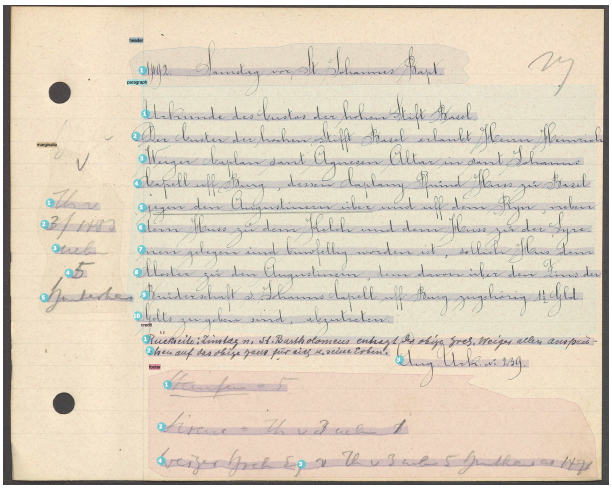


Abb. 1: Beispiel einer semantisch analysierten Seite. Screenshot aus Transkribus.

Die Strukturierung der Information auf diesen Karteikarten ist nicht «neutral», sie folgt den Interessen der Urheber des HGB ebenso wie Konventionen der Informationserfassung; somit besteht für die computergestützte Auswertung strukturell eine Differenz zwischen den digital verfügbaren Informationen und einer qualitativen Lektüre der Quellen.

Auf sprachlicher Ebene hingegen stellt sich das Problem eines spezifischen Einflusses durch die Urheber weniger, weil die Angaben meist als Kurzzitat in den Begriffen und Formulierungen der Quellen selbst erfolgt sind. Wenn auch aus dem Kontext der Dokumente gerissen, ermöglicht dies dennoch eine Analyse der Quellensprache, ihrer Begrifflichkeit und Semantiken.

Datenmodell und Workflow

Das Poster diskutiert das Projekt, wobei Datenmodell und Workflows zur Aufbereitung von rund 120’000 für uns relevanten digitalisierten Bilddateien zentral präsentiert werden. Dabei wird großer Wert auf Reproduzierbarkeit und die Möglichkeit zum wiederholten Prozessieren der genutzten Algorithmen gelegt.

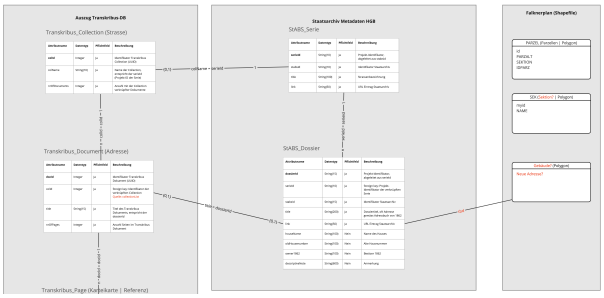


Abb. 2: Auszug aus der Modellbeschreibung der relationalen Projektdatenbank, umgesetzt in PostgreSQL. Screenshot: Projektdokumentation.

Das Datenmodell bildet aus unterschiedlichen Perspektiven und mit unterschiedlicher Granularität die vorhandenen Informationen ab. Einerseits wird auf archivische Metadaten (Signatur und Adressen als Strings, bezogen von einem SPARQL-Endpoint) und bereits vorliegende Geodaten (in Form von “Shapefiles”) zurückgegriffen, andererseits wird aus den gescannten Dokumenten in einem mehrstufigen Prozess Informationen extrahiert: Beginnend auf der Stufe “Sammlung” (Strasse), über “Einzeldokument” (Adresse), zur “Seite” (Karteikarte mit typischerweise einem Ereignis) und schliesslich der Textregion mit den zentralen Informationen (in einem “header” und einem “paragraph”).

Auf technischer Ebene bedeutet dies, dass die Dokumentenmassen in semantisch angereicherte Textregionen aufgeteilt werden (Quirós 2017), bevor eine Texterkennung der Handschriften durchgeführt wurde. Für beides, Segmentierung und Texterkennung, wurden spezifische Modelle erstellt, die eine zuverlässige Identifikation ermöglichen. Insbesondere die Texterkennung mit PyLaia (Puigcerver [2017] 2022), die mit mehr als zehn verschiedenen Händen umgehen musste (Hodel u. a. 2021; Hodel 2023; Pinche 2023), erzielt für ein grosses Modell überzeugende Fehlerraten im Bereich von 4% auf einem Testset. Die Resultate werden mit 3,5% noch besser, wenn nur auf

die zentralen Bereiche (die angesprochenen “header” und “paragraph”) fokussiert wird. Die mit vier Prozent Character Error Rate erzeugten Texte sind für die darauf folgenden Auswertungsschritte, insbesondere die Extraktion von benannten Entitäten (Personen, Orte, Organisationen), deren Relationen und genannte Events mit Hilfe spezifischer Sprachmodelle ausreichend (Torres Aguilar und Stutzmann 2021; Cafiero u. a. 2021; Hodel, Prada Ziegler, und Schneider 2023).

Datenvisualisierungen

Basierend auf den extrahierten Daten und der Kombination mit den identifizierbaren Grundstücken, wird es möglich, nach Begriffen (im Projekt “Eventtypen”) zu suchen und diese auf Karten zu visualisieren, etwa unter Berücksichtigung von Zeitschnitten, wie in Abb. 3 demonstriert.

Die Analyse von Events und Transaktionen bedingt Erkenntnisse zum normativen und semantischen Wandel des Liegenschaftsmarktes. Nur so können Überlieferungslücken, sprachliche Veränderungen und historischer Wandel auseinandergehalten und entsprechend visualisiert werden. Referenz ist dabei der HGB-Bestand: wie dicht ist die Überlieferung in einem Zeitraum und wie gross der Anteil von kategorisierbaren Karteikarten in der Erkennung von Events.

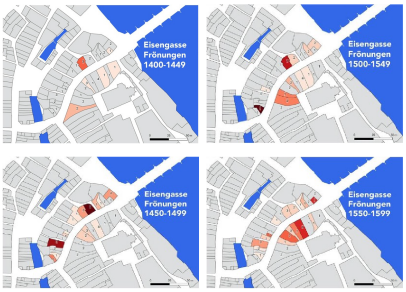


Abb. 3: Beispiel einer visuellen Auswertung mit der Filterung «Frönungen» (Event) gekoppelt mit der Häufigkeit des Auftretens über Zeitschnitte.

Zwischenresultate

Obwohl sich das Projekt noch in einer frühen Phase befindet und hinsichtlich der Auswertung weitere Verfeinerungen in den kommenden Jahren erwartet werden dürfen, wagen wir erste Aussagen.

Mit dem erarbeiteten Datenmaterial wird der Faktormarkt Boden im Kontext der dynamischen urbanen Ökonomien des Spätmittelalters und dessen formelle und informelle Strukturierungen als Voraussetzung seiner Funktionsweise analysiert.

Die Datenbasis eines digitalen HGB bietet die Chance, Untersuchungen des vormodernen städtischen Immobilienmarktes in synchroner und diachroner Perspektive einerseits, in räumlicher Skalierbarkeit von einzelnen Häusern bis zum gesamten Stadtraum andererseits, durchzuführen. Sie bietet die Grundlage für eine Heuristik ausgesprochen dichter Überlieferung, mit der in einer Zeit-Raum-Matrix Akteure, Praktiken und Vokabularien der «Ökonomien des Raums» variabel analysiert werden können.

Gleichzeitig zeigen die eingesetzten Methoden, dass auch für grosse Datensätze die maschinellen Lernverfahren einen reifen Status erreicht haben. Die Auswertung von Massenquellen und komplexer Findmittel kann also in Angriff genommen werden.

Bibliographie

Cafiero, Florian, Thibault Clérice, Paul Fièvre, Simon Gabay, und Jean-Baptiste Camps. 2021. „Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre“. *Journal of Data Mining & Digital Humanities* 2021 (Februar). https://doi.org/10.46298/jdmdh.6485.

Hodel, Tobias. 2023. „Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft. Anwendung, Einordnung und Methodenkritik“. *Historische Zeitschrift* 316 (1): 151–80. https://doi.org/10.1515/hzhz-2023-0006.

Hodel, Tobias, Ismail Prada Ziegler, und Christa Schneider. 2023. „Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern“. Graz, Austria, Juni 30. https://doi.org/10.5281/zenodo.8107616.

Hodel, Tobias, David Schoch, Christa Schneider, und Jake Purcell. 2021. „General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example“. *Journal of Open Humanities Data*, Journal of Open Humanities Data, 7. https://doi.org/10.5334/johd.46.

Pinche, Ariane. 2023. „Generic HTR Models for Medieval Manuscripts The CREMMALab Project“. https://hal.science/hal-03837519.

Puigcerver, Joan. (2017) 2022. „PyLaia“. Python. https://github.com/jpuigcerver/PyLaia.

Quirós, Lorenzo. 2017. „P2PaLA: page to PAGE layout analysis toolkit“. https://github.com/lquirosd/P2PaLA.

Torres Aguilar, Sergio, und Dominique Stutzmann. 2021. „Named Entity Recognition for French medieval charters“. In *Workshop on Natural Language Processing for Digital Humanities*. Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop. Helsinki, Finland. https://hal.archives-ouvertes.fr/hal-03503055.

Mehrsprachige Digital Literacy und Digital Humanities in der Lehre

Beers, Theodore
theo.beers@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0002-5129-5748

Kraneiß, Natalie
n.kraneiss@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0003-3584-285X

Müller-Laackman, Jonas
jonas.mueller-laackman@sub.uni-hamburg.de
Staats- und Universitätsbibliothek Hamburg, Deutschland
ORCID: 0000-0003-2279-6751

Thies, Antonia
antoniathies22@googlemail.com
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland
ORCID: 0000-0001-6153-7109

Wagner, Cosima
cosima.wagner@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0003-4957-3478

In der akademischen Lehre - nicht nur, aber vor allem in den Geisteswissenschaften - werden zunehmend digitale Methoden und Quellen eingebunden. Dabei stellt sich die Frage, wie das bereits vorhandene Wissen über Methoden und Tools sinnvoll und zielführend vermittelt werden kann. Diese Vermittlung wird zu einem Problem, wenn Lehrkräfte nicht über das nötige Fachwissen verfügen, um Digital Humanities als Teil des wissenschaftlichen Methodenkasten lehren zu können und es keine ausgebildeten DH-Lehrkräfte im Fachbereich gibt, die diese Aufgaben übernehmen können. Oft sind besonders die sogenannten “Kleinen Fächern” (https://www.kleinefaecher.de/) davon betroffen.

Hintergrund

Auf dem Deutschen Orientalistentag 2022 an der Freien Universität Berlin (FUB) fand eine durch die DHd AG Multilingual DH in Kooperation mit dem Multilingual DH Lab des Ada Lovelace Center for Digital Humanities (ADA) und dem Seminar für Semitistik und Arabistik der FUB veranstaltete Paneldiskussion zum Thema „Disrupting Digital Monolingualism“ statt. Neben der Diskussion von Herausforderungen der Abbildung nicht-lateinischer Schriften in digitalen Wissenschaftsinfrastrukturen und ihrer Verwendung in DH-Kontexten allgemein rückte schnell das Thema Ausbildung und Lehre in den Mittelpunkt.

Dabei stellte sich heraus, dass die sogenannten „Kleinen Fächer“ im Vergleich zu Fachdisziplinen, die über eine verhältnismäßig gute Versorgung mit DH-Infrastruktur verfügen, nicht nur in der allgemeinen DH-Praxis mit Problemen zu kämpfen haben, sondern vor großen Herausforderungen im Hinblick auf die Vermittlung von DH-Methodenwissen und die Förderung von Digital Literacy stehen. Hier fehlt es oft an Personal, das sowohl das nötige fachspezifische Wissen mitbringt, etwa Kenntnisse der Schrift und Sprache des Forschungsfeldes und -gegenstands, als auch über gute Kenntnisse in DH-Methoden sowie der Vermittlung von Digital und Data Literacy verfügt. Das wird umso mehr zu einem Problem, wenn die Institute und Zentraleinrichtungen nicht über ausreichend geeignetes Personal verfügen und eine universitäre Anstellung für Personen außerhalb der Wissenschaft kaum eine attraktive Alternative darstellt.

Neben den fachspezifischen Herausforderungen bezüglich digitaler Mehrsprachig- und -schriftlichkeit betrifft auch DH-Lehrende in den sogenannten Kleinen Fächern das Problem, dass Studien- und Prüfungsordnungen häufig nicht die notwendigen Voraussetzungen bieten, um von den üblichen Prüfungsformen abzuweichen und alternative Leistungen, wie produzierte Codes oder kollaborative Erschließungs- und Kodierungsarbeiten, anerkennen zu können.

Workshop „Digital Literacy in der multilingualen und -skriptualen Lehre“

Aufbauend auf dieser Paneldiskussion fand Anfang Mai 2023 der Workshop „Digital Literacy in der multilingualen und -skriptualen Lehre“[1] an der Staats- und Universitätsbibliothek Hamburg statt, bei dem Stakeholder der verschiedenen Statusgruppen zusammenkamen, um die Probleme und mögliche Lösungen zu diskutieren. Es nahmen Vertreter:innen der Studierendenschaft, Doktorand:innen, Professor:innen, Lehrpersonal, Postdocs und Vertreter:innen von wissenschaftlichen Bibliotheken teil, die sich in drei Arbeitsgruppen mit den Arbeitspaketen “Lehre und Methode”, “Infrastruktur und Rolle von Bibliotheken” und “Netzwerk” beschäftigten. Obwohl der Fokus auf den Bedarfen der “Kleinen Fächer” lag, sind die diskutierten Lösungen ebenso auf andere Disziplinen anwendbar.

Ergebnisse

Die Notizen und Ergebnisse der Diskussionen wurden auf einem Miro-Board (https://miro.com/app/board/uXjVMMx9Ck0=/?share_link_id=291913974430) festgehalten und bilden die Grundlage für ein Positionspapier sowie ein Living Handbook für Best Practices. Dieses im Entstehen begriffene Handbuch dient zum einen als Handreichung für Lehrende und Studierende, die sich mit den beschriebenen Problemen konfrontiert sehen, andererseits aber als stetig wachsende Informationsquelle über bestehende Bedarfe im Bereich der Lehre.

Arbeitspaket I: Lehre und Methoden

Problemfeld Lehre

Bei ihrem Bestreben, digitale Kompetenzen und Methoden in die Forschung und Lehre zu integrieren, begegnen Forscher:innen und Lehrende einer Vielzahl von Hindernissen und Problemen, die eine effiziente Lehre erheblich einschränken.

Lehrende, die bereits über sichere Methodenkenntnisse verfügen, haben diese meist autodidaktisch erworben. Dies birgt die Gefahr, dass Einzelpersonen ein isoliertes, fachspezifisches Know-How ohne Transferpotenzial aufbauen. In Verbindung mit einem Mangel an finanziellen Ressourcen führt dies häufig dazu, dass jene Forschenden dauerhaft als DH-Expert:innen eingesetzt werden, sodass sie nicht mehr in ihren ursprünglichen Fächern unterrichten können. Dies ist insbesondere für die “Kleinen Fächer” keine wünschenswerte Zukunftsperspektive. Vielmehr werden Lehrkräfte benötigt, die fachliche Expertise und Methoden der DH verknüpfen.

Um diesen Hindernissen begegnen zu können, ist es notwendig, dass DH in der Lehre als elementarer und relevanter Teil des Methodenkasten anerkannt wird. Nötig sind einerseits Einführungsveranstaltungen in den Digitalen Geisteswissenschaften. Dies muss auch die Reflexion über (etablierte) Methoden umfassen. Dafür ist es erforderlich, dass Lehrpersonal auch universitätsübergreifend angefragt werden darf, was den interdisziplinären Austausch und damit auch die zunehmende Vernetzung über Instituts- und Fachgrenzen hinweg fördern kann.

Andererseits sollte die Förderung des forschungsnahen Lernens auch in diesem Bereich stärker genutzt werden. Die Kombination klassischer Methoden mit computationellen Ansätzen kann etwa durch anwendungsbezogene Blockseminare sowie instituts- und universitätsübergreifende Co-Teachings und Teach-ins realisiert werden. Außerdem sollten geeignete Veranstaltungen für Studierende anderer Universitäten geöffnet werden, so dass notwendige Kenntnisse bedarfsgerecht und flexibel erworben werden können.

Darüber hinaus ist es von zentraler Bedeutung, dass Lehrende transparent mit ihren eigenen Lernerfahrungen umgehen, um gemeinsam Best Practices und Lösungen zu entwickeln und die Bereitschaft zur Weiterbildung universitätsübergreifend zu fördern. Auch kurze Workshops für Studierende und Dozierende und an Lehrveranstaltungen anschließende begleitende Tutorien sowie die Schaffung von DH-Helpdesks können den Weg zu besseren digitalen Kompetenzen ebnen.

Problemfeld multilinguale und multiskripturale Herausforderungen

Es gibt gute Software für zahlreiche DH-Anwendungsfälle, die jedoch oft eine unzureichende Funktionalität für Datensets in nicht-lateinischen Schriften aufweisen. So gibt es etwa Probleme mit Dialekten, Schriftsprachen sozialer Medien, (Syntax-)Morphologien oder Worttrennungen bspw. des Arabischen oder Japanischen. Auch Literaturprogramme wie Citavi und Zotero haben hier Defizite. Forschende sind daher oft gezwungen, eigene Lösungen zu entwickeln. Das ist nicht nur zeit- und ressourcenintensiv, sondern bedeutet eine Hürde für Forschungsvorhaben und Fachbereiche, deren Fokus etwa auf nicht-lateinischen Schriften liegt.

Erschwerend kommt hinzu, dass Weiterbildungsmaßnahmen für Forschende, sog. Train-the-Trainer-Angebote, fehlen. Auch hier sind überregionale und internationale Kollaborationen sowie transparente Lerndesigns elementar. Fächer wie Arabistik, Japanologie und Sinologie können als ideale Orientierungshilfen dienen, da Forschende in diesen Bereichen durch eben jene Hindernisse zum Teil schon Best Practice Lösungen entwickelt haben, von denen andere Fachdisziplinen profitieren können.[2]

Arbeitspaket II: Infrastruktur und Rolle von Bibliotheken

Dieses Arbeitspaket beinhaltet die Frage nach der Rolle von Bibliotheken als Akteurinnen und Partnerinnen im Bereich Digitalität und Lehre. Dabei sind fünf Elemente herauszustellen: Administratives, Sichtbarkeit, Ressourcen, Services und Kooperation.

Administratives

Bibliotheken verfügen über eine langjährige Expertise im Umgang mit Daten (insbesondere Metadaten und Onthologien). Mittlerweile beschäftigen viele wissenschaftliche Bibliotheken zusätzlich Personal, das explizit den Bereich Forschungsservices und Digital Humanities abdeckt. Diese Expertise ist aber im universitären Lehrkontext bislang kaum nutzbar. Die rechtlichen Rahmenbedingungen lassen es etwa kaum zu, dass Bibliothekspersonal niedrigschwellig Lehr- und Prüfungsverantwortung an Universi-

täten übernehmen kann. Dazu kommt, dass Lehrangebote seitens der Bibliothek im Einklang mit den meist recht strengen Vorgaben für universitäre Curricula stehen müssen, um überhaupt als solche bewertet und damit ins Vorlesungsverzeichnis integriert werden zu können.[3] Auch hier kann die Anerkennung neuer Formen von Prüfungsleistungen ein Problem sein, etwa wenn im Rahmen eines Hackathons eine gemeinschaftliche Coding-Leistung bewertet werden soll.

Sichtbarkeit

Die Notwendigkeit für eine Verbesserung der Sichtbarkeit von bibliothekarischen Angeboten und Expertise beschränkt sich aber nicht nur auf die Einbindung entsprechender Angebote in die Vorlesungsverzeichnisse. Vielmehr ist eine breite Einbindung des bibliothekarischen Lehrangebots in die Öffentlichkeitsarbeit der Universitäten notwendig, gerade auch für die Bekanntmachung von Services für Lehre und Studium ebenso wie für die Forschung. Auf diese Weise ließe sich auch eine nachhaltige Zusammenarbeit auf den Themenfeldern Digital Literacy und Informationskompetenz erreichen.

Ressourcen

Bibliotheken verfügen über andere Möglichkeiten der Bereitstellung dauerhafter Ressourcen als Fachbereiche oder Institute. In Kooperation mit den lokalen Rechenzentren können hier Räume geschaffen werden, in denen Arbeitsplätze und Technologien zur freien Nutzung für alle Statusgruppen zur Verfügung stehen. Davon profitieren Lehrende für ihre Seminare, ebenso wie Studierende, die für eigene Projekte und Abschlussarbeiten auf diese Räume und Ressourcen zurückgreifen können.

Services

Gerade durch die dauerhaft verfügbare Expertise an Bibliotheken können diese auch als Servicedienstleisterinnen für die Lehre auftreten. Das kann etwa die Betreuung von Lehrveranstaltungen sein, die nicht von DH-Fachpersonal durchgeführt werden, oder auch die Beratung im Hinblick auf die Planung und Durchführung von im geisteswissenschaftlichen Bereich nach wie vor neuartigen Lehrformaten und Aktivitäten, wie etwa Hackathons. Auch die Betreuung von Studierenden oder die Bereitstellung von Selbstlernmaterialien können Gegenstand eines solchen Services sein (Stichwort: Schreibzentren in Bibliotheken, Teaching Librarian als Stellenbezeichnung).

Kooperation

Bibliotheken sind durch ihre Stellung außerhalb der Fachbereichs- und Institutsstrukturen hervorragend als Schnittstellen zwischen Kooperationspartner:innen geeignet. Sie könnten ein Bindeglied schaffen zwischen Forschungseinrichtungen, Instituten, Zentraleinrichtungen und auch zu über- oder außeruniversitären Stakeholdern, wie z.B. dem Chaos Computer Club oder dem DHd-Verband.

Arbeitspaket III: Netzwerk

In diesem Arbeitspaket stellt sich die Frage nach der Zusammenarbeit und Vernetzung zwischen Stakeholdern im Bereich Digitalität und Lehre. Hervorzuheben ist hier vor allem die Notwendigkeit, zwischen DH-affinen Fachpersonen und solchen akademischen Stakeholdern, die in der Lehre tätig sind, aber nur wenig oder keinerlei Kenntnis von digitalen Methoden und Ansätzen haben, zu vermitteln. Dafür bedarf es nicht nur der personellen Zusammenarbeit, sondern auch der Aggregierung und der Vermittlung von Wissen.

Es gibt an Universitäten bereits Beispiele, wie eine solche Zusammenarbeit innerhalb des Universitätsbetriebs funktionieren kann. So gibt es etwa an der Universität Münster das Service Center für Digital Humanities (SCDH), das als Schnittstelle zwischen Bibliothek, DH-Forschung und anderen Stakeholdern fungiert und als Ansprechpartner für digitales Know-how sichtbar ist. Mit der Einführung des Zertifikats „Digital Humanities" wurde hier die Möglichkeit geschaffen, dass sich Studierende Workshops des SCDH anerkennen lassen können. An der Freien Universität Berlin betreibt das Multilingual DH Lab am Ada Lovelace Center für Digital Humanities ein Wiki (https://wikis.fu-berlin.de/display/nlsdh), in dem diverse Materialien zu Digital Humanities und Mehrsprachigkeit gesammelt werden.

Als Desiderat wurde hier eine überuniversitäre Vernetzung festgestellt. So kam als Idee die Etablierung einer „Co-Teaching Coupling Platform" auf, die ermöglichen soll, dass Lehrende bei Bedarf Expert:innen aus bestimmten Fachgebieten anfragen können, die dann im Rahmen eines Teach-In die entsprechenden Seminare besuchen. Damit wäre eine pragmatische Gelegenheit geschaffen, wie fehlende Fachkenntnis kurzfristig ausgeglichen werden könnte. Das stellt allerdings für das Problem der Einschränkungen seitens der Studien- und Prüfungsordnung oder der Planung von Curricula keine Lösung dar.

Sinnvoll wäre auch ein institutionell unabhängiger "Knowledge Hub" für das an den Institutionen gesammelte Wissen und die Erfahrungen und Erkenntnisse aus der Lehre, wenngleich hier die Frage geklärt werden müsste, wer eine solche Infrastruktur aufsetzt und betreibt.

Schließlich stellt sich hier auch die Frage nach der weiteren Vernetzung der an dieser Initiative beteiligten Personen. Wie kann eine größere Öffentlichkeit und mehr Aufmerksamkeit geschaffen werden für die Herausforderungen in der Lehre? Die DHd als Verband könnte hier etwa als Multiplikator und Interessenvertretung der DH-Seite auftreten. Entsprechend wurden bereits erste Anknüpfungspunkte in den AGs gesucht, etwa in der AG Referenzcurriculum.

Postersession

Im Rahmen der DHd 2024 möchten wir über die Erkenntnisse und Erfahrungen aus der bisherigen Arbeit berichten. Dabei sollen besonders die Ergebnisse des Workshops und die Elemente des geplanten Positionspapiers auf dem Poster dargestellt sowie erste Überlegungen über Form und Inhalt eines Living Handbooks angestellt werden. Zudem möchten wir einen Ausblick geben auf weitere Schritte der Vernetzung und Zusammenarbeit, auch mit Blick auf die mögliche Rolle des DHd-Verbandes in diesem Kontext.

Fußnoten

1. Für die Workshop-Einladung siehe https://blog.sub.uni-hamburg.de/?p=35416, zwei Berichte zum Workshop finden sich auf DH3 (https://dh3.hypotheses.org/794) und dem CtG-Blog (https://ctg.hypotheses.org/45)
2. Ein solches Pionierprojekt ist die Berlin-Hamburg-Kooperation “Closing the Gap in Non-Latin Script Data”, eine Plattform, auf der verschiedene DH-Projekte gesammelt werden, wodurch sich Forschenden durch Austausch und Vernetzung weiterbilden können. Siehe: https://m-l-d-h.github.io/Closing-The-Gap-In-Non-Latin-Script-Data/
3. Ausnahmen gibt es etwa in Sonderförderungen wie denen des DDLitLab der Uni Hamburg: https://www.i-sa.uni-hamburg.de/ddlitlab.html

Bibliographie

BMBF. o. J. *„Kleine Fächer – Große Potenziale - BMBF“*. https://www.bmbf.de/bmbf/de/forschung/geistes-und-sozialwissenschaften/kleine-faecher/kleine-faecher_node.html (zugegriffen: 07.07.2023)

Fiormonte, Domenico. 2021. „Taxation Against Overrepresentation? The Consequences of Monolingualism for Digital Humanities“. In: *Alternative Historiographies of the Digital Humanities*. 333–76. Earth: punctum books. https://doi.org/10.53288/0274.1.00.

Cordell, Ryan. 2016. “How Not to Teach Digital Humanities”. In: *Debates in the Digital Humanities 2016*. 459-474. Minneapolis: University of Minnesota Press.

Fyfe, Paul. 2018. “Reading, Making, and Metacognition: Teaching Digital Humanities for Transfer”. In: *Digital Humanities Quarterly 12(2)*: 1-12.

Grallert, Till, Xenia Monika Kudela, Eliese-Sophia Lincke, Colinda Lindermann, Jana-Katharina Mende, Jonas Müller-Laackman, Larissa Schmid. 2023. *Umgang mit Multilingualität im DACH und DHd Verband* (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.7957187.

Rehbein, Malte und Christiane Fritze. 2012. “Hands-on Teaching Digital Humanities. A Didactic Analysis of a Summer School Course on Digital Editing”. In: *Digital Pedagogy: Practices, Principles and Politics*. 47-78. Open Book Publishers.

Rosenblum, Brian, Frances Devlin, Tami Albin, Wade Garrison. 2015. “Collaboration and CoTeaching. Librarians Teaching Digital Humanities in the Classroom”. In: *Digital Humanities in the Library: Challenges and Opportunities for Subject Specialists*. 151-177. Association of College & Research Libraries.

Spence, Paul. 2021. *Disrupting Digital Monolingualism: A report on multilingualism in digital theory and practice*. London: Language Acts & Worldmaking project. https://doi.org/10.5281/zenodo.5743283 (zugegriffen: 07.07.2023).

Mentions technique, future re-invented? Eine quantitative Analyse literarischer Referenzen im niederländischen, deutschen und britischen Parlament.

Blank, Lina Lucy
lina.blank@uni-oldenburg.de
Carl von Ossietzky Universität Oldenburg, Deutschland

Der hier vorliegende Konferenzbeitrag soll Ergebnisse einer automatisierten Analyse von Parlamentsdebatten des Datenpakets *Parlspeech V2* (cf. Rauh und Schwalbach 2020a) präsentieren, in der die These einer (relativ) großen Skepsis niederländischer Eliten gegenüber niederländischer, zeitgenössischer Literatur (cf. Grüttemeier 2016; Grüttemeier 2018) anhand von Autor*innennennungen (*mentions*) quantitativ überprüft wird. Hierbei handelt es sich um eine der ersten digitalen Adaption der *mentions technique*, die durch die Betrachtung eines erweiterten Datensatzes im Vergleich zu vorangegangenen Untersuchungen (cf. van Dijk 2000: 374) eine signifikante Erweiterung für die literatursoziologische Forschung darstellt.

Die Analyse der Nennung von Autor*innennamen stellte Karl-Erik Rosengren in den 1980er Jahren als literatursoziologisches Instrument vor. Mithilfe der sogenannten *mentions technique* wird empirisch der geteilte literarische Referenzrahmen von Sprecher*innen analysiert (cf. Rosengren 1987: 298). Im Rahmen des DFG-Projektes ‚Literaturkonzepte der juristischen und politischen Eliten in den Niederlanden im 20. Jahrhundert‘ wird diese explorative, computationelle Analyse zur Modellierung des Verhältnisses zwischen dem politischen und dem literarischen Feld genutzt. Die theoretische Basis hierfür bietet die Bour-

dieusche Feldtheorie sowie ihre Erweiterungen (vgl. Bourdieu 2001, Bourdieu 2004, Witte 2015, van Rees & Dorleijn 1993, Dorleijn & van Rees 2006). Konkret soll dabei den Fragen nachgegangen werden: Wie sieht der literarische Referenzrahmen von niederländischen Politiker*innen aus, der sich mit Hilfe von Rosengrens *mentions technique* im direkten Vergleich mit deutschen und britischen Politiker*innan rekonstruieren lässt? Und inwiefern lassen sich literaturhistorisch etablierte Bilder des Verhältnisses von literarischem und politischem Feld an ihm überprüfen?

Die prädigitale Technik von Rosengren wird hierfür mit Hilfe digitaler Datenverarbeitungsformen transformiert. Bei dem eigens für dieses Projekt adaptierten Workflow wurde eine pythonbasierte algorithmischen Suche verwendet, um in dem Volltext-Korpus von Parlamentsdebatten *Parlspeech V2* (cf. Rauh und Schwalbach 2020a) nach einer Liste von Autor*innen-Namen zu suchen. Verwendet wurden zum europäischen Vergleich die drei Datenpakete: *Tweede Kamer* , *House of Commons* und *Bundestag* (Scrape date: 22.01.2019 – 06.01.2020). Die verwendeten Daten umfassen alle Redebeiträge, die in den Parlamenten zwischen dem 20.12.1994 und dem 14.12.2018 gehalten wurden. Für die Niederlande bedeutet dies zum Beispiel 1.099.588 Redebeiträge von 1.192 Sprecher*innen mit insgesamt 141.692.179 Wörtern. Aufgrund der hohen Frequenz an genannten Personen, Orten und Organisationen in diesem Datensatz wurde von der Verwendung von NER abgesehen. Für die Suche nach Autor*innen wurde ein Datensatz angelegt, in den 100 Autor*innen der sogenannten nationalen Kanons der Niederlande (cf. Deinsen et al. 2022) aufgenommen wurden. Dadurch sind sowohl zeitgenössische als auch historische Autor*innen im Datensatz enthalten. Als internationalen Vergleichswert wurden alle 250 Autor*innen, die in Bloom's *The Western Canon* (cf. Bloom 1994) genannt werden, in die Suche aufgenommen. Zur Suche nach Autor*innen wurde mit Synonymen, Alias und Regulären Ausdrücken [*RegEx*] gearbeitet.

Die Ergebnisse legen nahe, dass Autor*innen-Nennungen durch Politiker*innen den bestehenden Eindruck einer (relativ) geringen strukturellen Wertschätzung von niederländischen politischen Eliten für niederländische Literatur nicht nur erhärten, sondern dass die Skepsis innerhalb der Eliten in den Niederlanden: 1. Nicht nur bis in die 60er Jahre besteht sondern auch in der Gegenwart vorhanden ist, sich 2. besonders im europäischen Vergleich zeigt und 3. nicht nur eine (relative) Skepsis gegenüber zeitgenössischer niederländischer Literatur, sondern gegenüber Literatur im Allgemeinen aufweist. Hierauf deutet zunächst die schiere Summe der *mentions* im europäischen Vergleich hin. So konnten im niederländischen Parlament nur 473 *mentions* ausfindig gemacht werden, während im Bundestag 1.084 und im Britischen *House of Commons* ganze 3.012 Autor*innenverweise auftraten. Diese Verhältnisse spiegeln sich auch in der Anzahl der Nennungen pro Autor*in wider. Am häufigsten wird in den Niederlanden Gerard Reve genannt (47), in Deutschland Goethe (352) und in Großbritannien Shakespeare (660). Die Annahme eines strukturellen Unterschieds im literarischen Referenzrahmen von niederländischen gegenüber deutschen und britischen Politiker*innen verfestigt sich weiter, denn nennen niederländische Politiker*innen Autor*innen, so handelt es sich wahrscheinlicher um hochgradig kanonisierte (Reve, Multatuli, Vondel etc.) als um zeitgenössische (Grunberg, Lanoye, Spit etc.). 45 % der zehn häufigsten Nennungen sind hierbei international, im Vergleich zu 11% bzw. 7% in Deutschland und Großbritannien. Der vorgestellte Workflow bietet ein hohes Maß an Anschlussfähigkeit und wirft Fragen auf, die in Anschlussprojekten beantwortet werden sollten. Er ist open-source-basiert, skalierbar und kombiniert Errungenschaften aus den Digital Humanities wie das Volltext-Korpus von Parlamentsdebatten und der etablierten Feldtheorie.

Bibliographie

Bloom, Harold. 1994. *The Western Canon: The Books and School of the Ages.* New York, SanDiego, London: Harcourt Brace & Company.

Bourdieu, Pierre. 2001. *Die Regeln der Kunst. Genese und Struktur des literarischen Feldes* . Frankfurt/ Main: suhrkamp taschenbuch wissenschaft

Bourdieu, Pierre. 2006. *Der Staatsadel.* Konstanz: UVK Verlagsgesellschaft mbH.

van Deinsen, Lieve; Anthe Sevenants und Freek van de Velde. 2022. *De Nederlandstalige Literaire Canon(s) Anno 2022: Een Enquête naar de literaire Klassieken: Rapportage (Voorpublicatie).* Gent: Koninklijke Academie voor Nederlandse Taal en Letteren.

Van Dijk, Nel. „Das Zitat als Autorenverweis: Ein prestigebestimmendes Instrument". In: K. Beekman, R. Grüttemeier *Instrument Zitat: Über den literaturhistorischen und institutionellen Nutzen von Zitaten und Zitieren* . Amsterdam, Atlanta: Rodopi. 367-392.

Dorleijn, Gillis & van Rees, Kees [Hg.]. 2006. *De productie van literatuur. Het literaire veld in Nederland 1800– 2000* . Nijmegen: Uitgeverij Vantilt.

Grüttemeier, Ralf. 2016. "Nederland en de Nobelprijs voor Literatuur 1901– 1965" In: *Nederlandse Letterkunde* 21, 131– 156.

Grüttemeier, Ralf. 2018. "Nederlandse reserves tegenover moderne Nederlandstalige literatuur. Het beeld van Multatuli in literatuurgeschiedenissen" In: J. Bel, R.R. Honings und J. Grave [Hg.]: *Multatuli nu: Nieuwe perspectieven op Eduard Douwes Dekker en zijn werk.* Hilversum: Verloren, 69– 90.

Rauh, Christian und Jan Schwalbach. 2020a. "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies", Harvard Dataverse.

Rauh, Christian und Jan Schwalbach. 2020b. "Release Note", Harvard Dataverse.

van Rees, Kees & Dorleijn, Gillis. 1993. *De impact van literatuuropvattingen in het literaire veld. Aandachtsgebied literaire opvattingen van de*

Stichting Literatuurwetenschap. Den Haag: Sichting Literatuurwetenschap.

Rosengren, Karl Erik. 1987. *Literary Criticism: Future Invented.* In: Poetics 16, 295– 325.

Schumacher, Mareike. 2018. „Named Entity Recognition (NER)“. In: *forTEXT. Literatur digital erforschen* . URL: https://fortext.net/routinen/methoden/named-entity-recognition-ner [Zugriff: 01. Dezember 2023].

Witte, Daniel. 2015. „Umstrittene Grenzen: Das Feld der Macht als Ort von Deutungskämpfen um Recht und Religion.“, In: W. Gephart, J.C. Sunturp [Hg.]: *Rechtsanalyse als Kulturforschung II* , Frankfurt/ Main: C.H. Beck. 357– 391.

Narrativität visualisieren - Eine Rezeptionsstudie zur Evaluation der heuristischen Qualität von Narrativitätsgraphen

Hatzel, Hans Ole
hans.ole.hatzel@uni-hamburg.de
Universität Hamburg, Deutschland
ORCID: 0000-0002-4586-7260

Stiemer, Haimo
stiemer@linglit.tu-darmstadt.de
Technische Unversität Darmstadt, Deutschland
ORCID: 0000-0002-4407-2415

Biemann, Chris
chris.biemann@uni-hamburg.de
Universität Hamburg, Deutschland
ORCID: 0000-0002-8449-9624

Gius, Evelyn
evelyn.gius@tu-darmstadt.de
Technische Unversität Darmstadt, Deutschland
ORCID: 0000-0001-8888-8419

Die visuelle Repräsentation von literarischen Phänomenen ist ein etablierter Ansatz in den Computational Literary Studies, um die aus Texten extrahierten Daten bzw. abgeleiteten Strukturen zu explorieren und zu interpretieren (cf. Baillot u. Lassner 2022; Krämer 2014). Vor diesem Hintergrund soll das vorgeschlagene Poster die Ergebnisse einer Rezeptionsstudie präsentieren, mit der die heuristische Qualität von den im Projekt EvENT ("Evaluating Events in Narrative Theory") bislang generierten Narrativitätsgraphen überprüft wurde. Ziel der Studie war es, die Erkennbarkeit der den Graphen zugrunde liegenden Texte zu untersuchen, um hieraus Rückschlüsse für die Weiterentwicklung des EvENT-Ansatzes wie auch die Anwendbarkeit der Graphen für die literaturwissenschaftlich-hermeneutische Praxis zu ziehen. Dabei stellen die Graphen das Ausmaß der Narrativität (auf der y-Achse) über den Textverlauf (auf der x-Achse) dar.

Die Narrativitätsgraphen basieren auf den vier, im EvENT-Projekt auf der Grundlage des narratologischen Forschungsstands konzipierten Ereigniskategorien (Zustandsveränderungen, Prozesseereignisse, statische Ereignisse und Nicht-Ereignisse) und der ihnen zugewiesenen Narrativitätsgrade (cf. Vauth u. Gius 2021). Die über die automatisierte Annotation von Verbalphrasen auf der Textoberfläche detektierten Ereignisse (cf. Hatzel 2022) werden im EvENT-Projekt verwendet, um die Narrativität von Texten über den Textverlauf als Narrativitätsgraphen abzubilden und damit auch die Handlung ihrer Geschichten zu modellieren (cf. Vauth et al. 2021; Gius u. Vauth 2022). Eine Annahme bei dieser Modellierung war, dass die Graphen ebenso die "Erzählwürdigkeit" der Ereignisse in den Texten indizieren und sich damit dem in der Narratologie als Event II diskutierten Phänomen annähern (cf. Hühn 2009, S. 80). Event II stellt einen Ereignistyp mit zusätzlichen Merkmalen im interpretativen Kontext dar, wie z. B. Relevanz oder Unerwartetheit, geht also über die vier oben genannten, grundlegenden Ereigniskategorien hinaus.

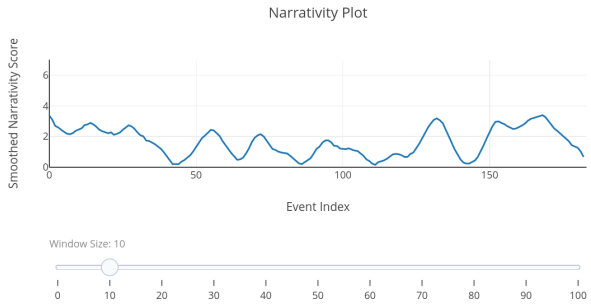


Abbildung: Beispiel EvENT-Narrativitätsgraph (Franz Kafkas *Schakale und Araber*, 1917)

Die Studie wurde als Webanwendung konzipiert. React und Plotly wurden im Front-End verwendet und die Antworten an ein fastAPI-basiertes Back-End übermittelt, welches diese in einer PostgreSQL-Datenbank protokollierte. Auf diese Weise wurde eine schnelle Iteration des Studiendesigns sowie die Teilnahme auf allen wichtigen Plattformen ermöglicht.

Die 19 Teilnehmer:innen der Studie (aktuelle und ehemalige Studierende der Germanistik), wurden gebeten, einem literarischen Text den richtigen Narrativitätsgraphen zuzuordnen, wobei für jeden Text vier Graphen als Antwortmöglichkeiten ausgegeben wurden. Die Textdarbietung er-

folgte auf drei unterschiedlichen Abstraktionsniveaus bzw. in drei Phasen. In der ersten Phase wurden die Teilnehmenden gebeten, zwei kurze deutschsprachige literarische Texte zu lesen und diese jeweils dem richtigen Graphen zuzuordnen. In der zweiten Phase erfolgte die Auswahl der Graphen auf der Grundlage von Zusammenfassungen, in denen die wesentlichen Ereignisse des jeweiligen Textes in chronologischer Reihenfolge präsentiert werden. In der dritten Phase wurden den Teilnehmenden nur die Titel der Texte angeboten, denen ein entsprechender Graph zugewiesen werden sollte. Bei diesen Texten handelt es sich um kanonische Texte (z. B. Hänsel und Gretel von den Brüdern Grimm), deren allgemeine Bekanntheit vorausgesetzt werden konnte.

Zu den auf dem Poster zu präsentierenden Resultaten der Studie gehört, dass der Anteil der korrekten Zuordnungen der Graphen zu den Texten durch die Teilnehmenden bei 25,56 % und damit nicht signifikant über dem Zufallsprinzip liegt. Allerdings gibt es eine positive Korrelation bei Teilnehmenden mit EvENT-Erfahrung (durch Mitarbeit im Projekt, Kenntnis der Annotationsguidelines), welche die korrekten Texte in 47.5% der Fälle ausgewählt haben, ein Wert der mittels Binomialtest als statistisch signifikant identifiziert wurde ($p < 0.01$). Wir konnten zudem mit statistischer Signifikanz zeigen, dass die Zeit, die sich Teilnehmende zur Beantwortung einer Frage nahmen, mit der Quote der richtigen Antworten korreliert. Auch waren lange Texte für unsere Annotator:innen statistisch signifikant schwieriger zu identifizieren als kurze (hier ist anzumerken, dass alle vier Optionen für eine Identifikation stets so gewählt waren, dass sie ähnliche Längen aufwiesen). Offenkundig ist die Identifikation der textzugehörigen Graphen voraussetzungsreich. Im Anschluss an die Zuordnungsaufgaben beantworteten die Teilnehmenden zwei offene Fragen nach ihren Entscheidungsgrundlagen im Verlauf der Studie. Die Antworten legen nahe, dass die Narrativitätsverläufe der Graphen als Repräsentationen von Event II-Vorkommen und die Amplitudenausschläge damit als Verweise auf besonders handlungsrelevant erscheinende Textpassagen interpretiert wurden. Als entscheidungsrelevant wurden von den Teilnehmenden außerdem nicht nur die Peaks der Graphen, sondern auch deren Anzahl sowie Anfang und Ende eines Narrativitätsverlaufs ausgewiesen.

Bibliographie

Baillot, Anne und David Lasser. 2022. “Von Graphen zu Word Embeddings – Zur Entwicklung des mathematischen und visuellen Instrumentariums der Literaturwissenschaft.” Germanica 71/2, Landkarten und Zeitleisten: zur Funktion von Bildern in der Literaturgeschichte / La carte et la frise: les images de l'histoire littéraire, entre visualisation et modélisation: 191-203. https://doi.org/10.4000/germanica.19002

Gius, Evelyn und Michael Vauth. 2022. Inter Annotator Agreement und Intersubjektivität – Ein Vorschlag zur Messbarkeit der Qualität literaturwissenschaftlicher Annotationen. DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands “Digital Humanities im deutschsprachigen Raum”. Potsdam. https://doi.org/10.5281/zenodo.6328209.

Hatzel, Hans Ole. 2022. Event Narrativity Classifier. Zenodo. https://doi.org/10.5281/zenodo.6821142.

Hühn, Peter. 2009. Event and Eventfulness. In: Handbook of Narratology, hg. von Peter Hühn; Christoph Meister, John Pier u. Wolf Schmid, 80-97. Berlin/New York: De Gruyter.

Krämer, Sybille. 2014. Zur Grammatik der Diagrammatik: Eine Annäherung an die Grundlagen des Diagrammgebrauches. Zeitschrift für Literaturwissenschaft und Linguistik, 176/4: 11-30. http://dx.doi.org/10.1007%2FBF03377227

Vauth, Michael und Evelyn Gius. 2021. Richtlinien für die Annotation narratologischer Ereigniskonzepte. Zenodo. https://doi.org/10.5281/zenodo.5078174.

Vauth, Michael, Hans Ole Hatzel, Evelyn Gius und Chris Biemann. 2021. “Automated Event Annotation in Literary Texts“. In: CHR 2021: Computational Humanities Research Conference, 333-345. Amsterdam. http://ceur-ws.org/Vol-2989/short_paper18.pdf.

Nutzergruppenspezifische Zugänge zu mündlichen Korpora aus dem Archiv für Gesprochenes Deutsch: neue Tools, neue Forschungsperspektiven

Frick, Elena
frick@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Deutschland

Helmer, Henrike
helmer@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Deutschland

Das Archiv für Gesprochenes Deutsch (AGD) am Leibniz-Institut für Deutsche Sprache (IDS) ist ein Forschungsdatenzentrum für Korpora des gesprochenen Deutsch (Stift und Schmidt, 2014). Es übernimmt Korpora aus abgeschlossenen Forschungsprojekten, archiviert sie und stellt sie anderen Forschenden für die Nachnutzung bereit. Seine Sammlung, hauptsächlich bestehend aus Gesprächs-, Interview- und Variationskorpora, wurde in verschiedenen

Regionen im deutschsprachigen Raum und in extraterritorialen deutschsprachigen Gebieten erhoben und enthält umfangreiche Metadaten. Ein breites Spektrum von Themen und Kommunikationssituationen aus privaten, institutionellen und öffentlichen Lebensbereichen sowie digitalisierte und mit dem Audio-/Video-Signal alignierte Transkripte mit linguistischen Mehrfachannotationen ermöglichen die Erforschung vielseitiger linguistischer Fragestellungen und bieten darüber hinaus eine wichtige Datengrundlage auch für die Nachnutzung in anderen Disziplinen, wie etwa der empirischen Sozialforschung und Oral History Studies.

Zum Zweck der Nachnutzung werden am IDS Methoden und Technologien entwickelt, die dem Erstellen, Aufbereiten und wissenschaftlichen Arbeiten mit mündlichen Korpora dienen. Aus Nutzerperspektive ist hier vor allem die Datenbank für Gesprochenes Deutsch (DGD[1]; Schmidt, 2017) zu nennen, die aktuell 40 Korpora mit insgesamt fast 5000 Stunden Audio-/Videoaufnahmen für Nutzende nach einer Registrierung und im Rahmen wissenschaftlicher Forschung und Lehre bereitstellt. Die DGD ermöglicht webbasiertes Browsen und systematisches Durchsuchen der Korpora und ist eine international etablierte, breit genutzte Korpusanalyseplattform (aktuell ca. 16.500 registrierte NutzerInnen).

Mit Blick auf neue Nutzergruppen ist in den letzten Jahren im Rahmen des Projekts *ZuMult*[2] und in Kooperation mit dem Hamburger Zentrum für Sprachkorpora (HZSK) sowie dem Herder-Institut der Universität Leipzig eine ganze Reihe von neuen Webanwendungen entstanden (Fig. 1). Vor allem für den Bereich Fremdsprachendidaktik, DaF/DaZ-Forschung und -Lehre wurden Korpusnutzungsszenarien entwickelt und entsprechende Online-Angebote wie *ZuMal*[3] und *ZuViel*[4] implementiert. Diese Webanwendungen bieten eine Filterung der einzelnen Interaktionen nach für DaF-/DaZ-relevanten Parametern (z.B. Niveaustufenzugehörigkeit des enthaltenen Wortschatzes, hoher/niedriger Anteil an Mündlichkeitsphänomenen) und eine Visualisierung der schwierigkeitsbezogenen Phänomene in einzelnen Transkripten (z.B. Sprechgeschwindigkeit), was eine schnelle Beurteilung der Eignung des entsprechenden Korpusabschnittes für die Lehre erlaubt. Auch für korpusbasierte lexikologische/lexikographische Forschung wurde ein eigener nutzergruppenspezifischer Zugang zu mündlichen Korpora geschaffen: Es handelt sich um eine Funktionalität, die es erlaubt, eine benutzerdefinierte Liste von Lemmata in *ZuRecht*[5] hochzuladen, Transkripte nach der Anzahl der Lemmata aus dieser Liste zu filtern und somit schnell passende Gespräche als Belege beim Verfassen z.B. von Wörterbuchartikeln zu finden. Die im Projekt konzipierte und als *Open Source* verfügbare Softwarearchitektur[6] für die neuen Korpuszugänge bietet hohe Flexibilität. Dank des Drei-Ebenen-Softwaremodells und der objektorientierten Modellierung der Korpusbestandteile in Kombination mit aktuellen Standards[7] lassen sich unkompliziert und nachhaltig neue Applikationen für weitere Nutzergruppen unabhängig von Standort und Korpusdatenformaten entwickeln (vgl. dazu Schmidt et al. 2023).

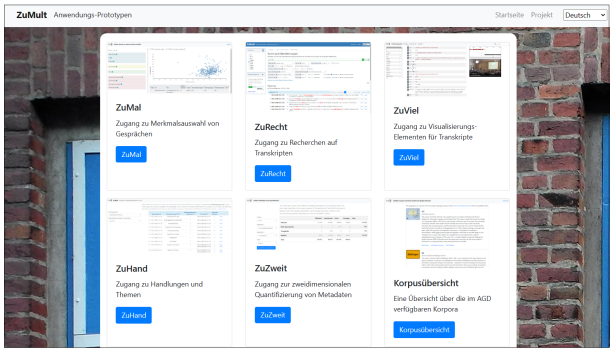


Fig. 1: Webseite mit ZuMult-Anwendungen

Ein weiteres Tool, das derzeit auf Basis der *ZuMult*-Architektur in enger Zusammenarbeit zwischen SoftwareentwicklerInnen und GesprächsforscherInnen entwickelt wird, orientiert sich spezifisch an den Bedarfen der Gesprächsforschung und macht Wiederholungen in der gesprochenen Sprache auffindbar. Das Konzept umfasst eine MTAS[8]/Lucene-basierte Suchmaschine mit einer mächtigen korpuslinguistischen Suchanfragesprache in Kombination mit einem nutzerfreundlichen GUI-Filter (Fig. 2), der speziell für die gesprächsanalytische Forschung entwickelt wurde und auf Spezifika von Gesprächskorpora ausgerichtet ist (z.B. durch die Berücksichtigung von nonverbalen Phänomenen und Sprecherüberlappungen). Dank der linguistischen Annotationen und umfangreichen Metadaten der bereitgestellten Korpora können Wiederholungen von komplexen sprachlichen Phänomenen in transkribierter, normalisierter oder lemmatisierter Form in festlegbaren pragmatischen Kontexten gesucht werden (z.B. eine Mehrwortsequenz mit optionalen Häsitationsphänomenen, die von einer Sprecherin am Ende ihres Beitrags realisiert und von einem männlichen Sprecher eventuell mit einer abweichenden Wortfolge direkt nach dem Sprecherwechsel und außerhalb einer Sprecherüberlappung wiederholt wird). Darüber hinaus wurde GermaNet (Henrich und Hinrichs, 2010) in die Anwendung integriert, was das Finden von Wiederholungskonstruktionen ermöglicht, die Synonyme, Hyperonyme oder Hyponyme enthalten können. So können etwa Fälle gefunden werden, in denen Wiederholungen zum Zweck einer Begriffsklärung verwendet werden (z.B. *käschte/rechteck*, *gynäkologe/frauenarzt*).

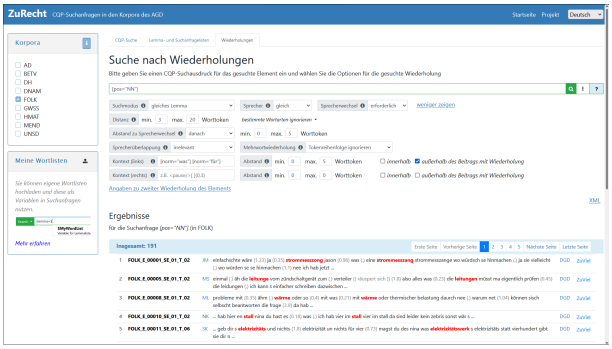


Fig. 2: Tool für die Suche nach Wiederholungen

Das Vorkommen von Wiederholungen und ihre Funktion im Rahmen von sprachlichen Praktiken (z.B. im Spracherwerb) wurden bereits vielfach und in verschiedenen Disziplinen untersucht. Viele Fragestellungen wurden allerdings hauptsächlich für die englische Sprache und noch nicht anhand von Eigenheiten des Deutschen überprüft. Außerdem gibt es Fragestellungen, die bis jetzt noch nicht in den Blick genommen wurden wie z.B. die Relevanz von multimodalen Ressourcen (Gestik, Gesichtsausdrücke usw.) beim Wiederholen von Äußerungen. Auch können phonetische und prosodische Besonderheiten etwa von Wiederholungen ganzer Äußerungen (z.B. nach Unterbrechungen) systematisch untersucht werden. Das neue Tool bietet Forschenden die Möglichkeit, solche und ähnliche Fragestellungen, in denen Wiederholungen eine Rolle spielen, effizienter zu untersuchen, in dem es über eine systematische Suche in Audio/Video-Daten gezielt gewünschte Sequenzen und Kontexte auffindbar macht, ohne dass Forschende mühsam manuell komplette Transkripte gesprochener Sprache sichten müssen. Eine solche Funktion gab es bisher nicht für Korpora des gesprochenen Deutsch.

Das Vorkommen von Wiederholungen und ihre Funktion im Rahmen von sprachlichen Praktiken (z.B. im Spracherwerb) wurde bereits vielfach und in verschiedenen Disziplinen untersucht. Viele Fragestellungen wurden allerdings hauptsächlich für die englische Sprache und noch nicht anhand von Eigenheiten des Deutschen überprüft. Außerdem gibt es Fragestellungen, die bis jetzt noch nicht in den Blick genommen wurden wie z.B. die Relevanz von multimodalen Ressourcen (Gestik, Gesichtsausdrücke usw.) beim Wiederholen von Äußerungen. Auch könnten phonetische und prosodische Besonderheiten etwa von Wiederholungen ganzer Äußerungen (z.B. nach Unterbrechungen) systematisch untersucht werden. Das neue Tool bietet Forschenden die Möglichkeit, solche und ähnliche Fragestellungen, in denen Wiederholungen eine Rolle spielen, effizienter zu untersuchen, in dem es über eine systematische Suche in Audio/Video-Daten gezielt gewünschte Sequenzen und Kontexte auffindbar macht, ohne dass Forschende mühsam manuell komplette Transkripte gesprochener Sprache sichten müssen. Eine solche Funktion gab es bisher nicht für Korpora des gesprochenen Deutsch.

Fußnoten

1. https://dgd.ids-mannheim.de
2. https://zumult.org
3. https://zumult.ids-mannheim.de/ProtoZumult/prototype/dist/zuMal.jsp
4. https://zumult.ids-mannheim.de/ProtoZumult/jsp/zuViel.jsp?transcriptID=FOLK_E_00349_SE_01_T_01
5. https://zumult.ids-mannheim.de/ProtoZumult/jsp/zuRecht.jsp
6. https://github.com/zumult-org/zumultapi
7. Media: PCM-WAV/MP3, MPEG-4; Metadaten: XML, CMDI; Transkriptionen/Annotationen: ISO 24624:2016; Korpussuche: Lucene, CQP
8. https://textexploration.github.io/mtas

Bibliographie

Henrich, Verena und Erhard Hinrichs. 2010. „GernEdiT – The GermaNet Editing Tool.“ In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC)*. Valletta, Malta, Mai 2010, 2228-2235.

Schmidt, Thomas. 2017. „DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim.“ In *Zeitschrift für germanistische* Linguistik 45: 3. Berlin / Boston: de Gruyter, 451-463.

Schmidt, Thomas, Christian Fandrych, Elena Frick, Matthias Schwendemann, Franziska Wallner und Kai Wörner. 2023. „Zugänge zu mündlichen Korpora für DaF und DaZ – Projekt, Datengrundlagen, technische Basis.“ In *KorDaF* 3(1): 1–12.

Stift, Ulf-Michael und Thomas Schmidt. 2014. „Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch.“ In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache, 360-375.

Ob Werkzeugkoffer, Werkstatt oder Baumarkt: offene, community-kuratierte Tool Registries mit Wikidata

Grallert, Till
till.grallert@fu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID: 0000-0002-5739-8094

Eckenstaler, Sophie
sophie.eckenstaler.1@ub.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Tirtohusodo, Samantha
samantha.tirtohusodo.1@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Schlesinger, Claus-Michael
claus-michael.schlesinger@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID: 0000-0001-6718-5773

Tools oder Werkzeuge, im Folgenden verstanden als Software und Verfahren, sind nicht nur Mittel zum Zweck, sondern als solche auch Gegenstand von Verzeichnissen. Dabei richten sich Toolverzeichnisse, die im Kontext der Digital Humanities angelegt werden, weniger auf die Verzeichnung von Objekten, die in absehbarer Zeit auch historisch werden, sondern bedienen zunächst ein instrumentelles Interesse. In diesem Zusammenhang hat sich das Toolverzeichnis in den Digital Humanities inzwischen als eigenes Genre etabliert: von DiRT zu Bamboo und TAPoR (3.0)[1] (Grant u. a. 2020), großen EU-Projekten wie dem *Social Sciences and Humanities Open Marketplace*,[2] den Konsortien der deutschen Nationalen Forschungsdateninfrastruktur (NFDI), den Fachinformationsdiensten (FID) oder individuellen Bibliotheken und Instituten. Allen ist gemeinsam, dass sie einen Bedarf der Forschungscommunities nach einem Überblick über computergestützte Werkzeuge mit kuratorischen Ansätzen der Wissensorganisation bedienen. Für die meisten Ansätze gilt, dass sie über keine dauerhafte Finanzierung verfügen, dass sie primär auf die Kuratierung durch (unbezahlte) Expert_innen-Gremien oder Crowdsourcing setzen, diesen Prozess aber nicht dauerhaft und nachhaltig gewährleisten können, dass Datensilos mit proprietären Infrastrukturen (Datenmodelle, Backends und Frontends) geschaffen werden, und dass nur in geringem Maße APIs angeboten und dokumentiert werden. Die zum Teil umfassenden Ansprüche an möglichst vollständige Katalogisierung verfügbarer Möglichkeiten computergestützter Werkzeuge sind dabei mit einem Gegenstand konfrontiert, der oft nur schwer abgrenzbar ist und sich im Rhytmus von Release-Zyklen ständig verändert. Der Anspruch eines umfassenden oder repräsentativen Abbilds aktuell verfügbarer Werkzeuge für die computergestützte Forschung in den Digital Humanities ist daher kaum einzulösen und kann, wo er formuliert wurde, als gescheitert gelten. (Dombrowski 2021). Die Bedarfe bleiben natürlich bestehen. Wünschenswert wäre daher, dass zumindest die Daten von Verzeichnisprojekten dauerhaft, mit Bezug auf ein Referenzdatenmodell, permanenten URIs und in einem stabilen, maschinenlesbaren Format als Basis zur Verfügung stünden.

Der Beitrag stellt unseren Vorschlag einer offenen Basisinfrastruktur für Toolverzeichnisse vor. Dabei steht Wikidata[3] als eine verteilte, community-kuratierte Normdatei und offene Softwareplattform im Zentrum unseres Vorschlages. Wikidata erlaubt es, minimale Datenmodelle iterativ zu entwickeln, Datensätze zu pflegen und sie in Wikiprojekten zu kuratierten Sammlungen zusammenzustellen. Auf der Datenebene erlaubt Wikidata die unmittelbare Nutzung sämtlicher Informationen als Linked Open Data über SPARQL, APIs sowie das etablierte Webinterface. Wikidata ist außerdem eine der Quellen für das *Virtual International Authority File* (VIAF)[4] und für zusammenfassende Informationen in den Ergebnislisten der dominanten Suchmaschinen, was die Sichtbarkeit der Datensätze enorm erhöht. Darüber hinaus bieten Wikidata und ihre Schwesterprojekte eine etablierte Governancestruktur für nutzergenerierte und -kuratierte Inhalte. Jede_r kann die Einträge beitragen und pflegen, die für ihre je konkrete Forschung relevant sind. Anders als bei vielen Infrastrukturen der Digital Humanities ist die Vielsprachigkeit von Interfaces und Datensätzen ein grundlegendes Feature. Auf dieser Datenbasis lassen sich dann Fachcommunity-spezifische Toolverzeichnisse kuratieren und anreichern. Denkbar ist etwa eine Klassifizierung unter Anwendung der TaDiRAH-Taxonomie[5] (Borek u. a. 2021) oder die Hinterlegung von Anwendungsbeispielen, Publikationen oder Tutorials im angereicherten Datensatz.

Um für verteilt angelegte Datensätze ein gemeinsames Referenzmodell zu schaffen schlagen wir ein reduziertes Basisdatenmodell für DH-Werkzeuge vor, das minimale bibliografische und technische Eigenschaften definiert und anschlussfähig bleibt für bereits existierende Datenmodelle wie dem des Software Preservation Network oder von RIDE (Christopherson u. a. 2022; Sichani und Spadini 2022). Dieses Basisdatenmodell garantiert den Zugriff auf die gemeinsam verbindlich vereinbarten Basisdaten für die Nachnutzung von Einträgen in eigenen kuratierten Sammlungen.

Die Wikidata-Plattform erlaubt die Umsetzung des Ansatzes ohne weitere Software. Denkbar ist aber auch, Wikidata ausschließlich als Normdatei und Datenprovider für

eigene Frontends einzusetzen, so wie es z.B. Scholia[6] für die Profile von Wissenschaftler_innen tut (Nielsen, Mietchen, und Willighagen 2017). Schließlich adressiert unser Vorschlag die Nachhaltigkeit von Projektförderungen durch den kontinuierlichen Beitrag von Daten zu den *Digital Commons* (Wittel 2013) in Gestalt von Wikidata während der Projektlaufzeit und die Weiternutzung dieser Daten nach der Projektlaufzeit. Damit ist unser Vorschlag Teil einer Bewegung, Wikidata in der Wissenschaft und GLAM-Institutionen nicht mehr nur als Anbieter von Inhalten wahrzunehmen (vgl. Zhao 2022; Fischer und Ohlig 2019).

Fußnoten

1. https://tapor.ca/
2. https://marketplace.sshopencloud.eu/
3. https://wikidata.org/
4. https://viaf.org/
5. https://vocabs.dariah.eu/tadirah/
6. https://scholia.toolforge.org/

Bibliographie

Borek, Luise, Canan Hastik, Vera Khramova, Klaus Illmayer, und Jonathan D. Geiger. 2021. „Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR“. In *Information Between Data and Knowledge*, 321–32. Schriften Zur Informationswissenschaft 74. Glückstadt: Werner Hülsbusch. https://doi.org/doi.org/10.5283/epub.44951.

Christopherson, Allan, Elena Colón-Marrero, Dianne Dietrich, Patricia Falcao, Claire Fox, Karen Hanson, Allen Kwan, und Matthew McEniry. 2022. „Software Metadata Recommended Format Guide“. Cornell University Library. https://doi.org/10.7298/XE9S-3B15.

Dombrowski, Quinn. 2021. „The Directory Paradox“. In *People, Practice, Power: Digital Humanities Outside the Center*, herausgegeben von Anne B. McGrail, Angel David Nieves, und Siobhan Senier. Debates in the Digital Humanities. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/people-practice-power/section/ca87ec4c-23a0-452d-8595-7cfd7e8d6f0c.

Fischer, Barbara, und Jens Ohlig. 2019. „‚GND Meets Wikibase‘ - Eine Kooperation. Eine Bundesbehörde Geht Auf Expedition Im Wikiversum: Ein Neues Testfeld Für Wikibase“. *GND* (blog). 8. Mai 2019. https://wiki.dnb.de/pages/viewpage.action?pageId=147754828.

Grant, Kaitlyn, Quinn Dombrowski, Kamal Ranaweera, Omar Rodriguez-Arenas, Stéfan Sinclair, und Geoffrey Rockwell. 2020. „Absorbing DiRT: Tool Directories in the Digital Age“. *Digital Studies / Le Champ Numérique* 10 (1). https://doi.org/10.16995/dscn.325.

Nielsen, Finn Årup, Daniel Mietchen, und Egon Willighagen. 2017. „Scholia, Scientometrics and Wikidata“. In *The Semantic Web: ESWC 2017 Satellite Events*, herausgegeben von Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz,, Fabio Ciravegna, und Olaf Hartig, 237–59. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-70407-4_36.

Sichani, Anna-Maria, und Elena Spadini, Hrsg. 2022. *RIDE*, Nr. 15: Tools and Environments (Dezember). https://ride.i-d-e.de/issues/issue-15/.

Wittel, Andreas. 2013. „Counter-commodification: The economy of contribution in the digital commons“. *Culture and Organization* 19 (4): 314–31. https://doi.org/gmqgqq.

Zhao, Fudie. 2022. „A systematic review of Wikidata in Digital Humanities projects“. *Digital Scholarship in the Humanities*, Dezember, 1–22. https://doi.org/10.1093/llc/fqac083.

Open Public Peer Review auf dem Prüfstand: Community-Einfluss auf den Referenzrahmen für eigenständige digitale Wissenschaftskommunikation durch Forschende

Seltmann, Melanie Elisabeth-H.
melanie.seltmann@tu-darmstadt.de
Universitäts- und Landesbibliothek Darmstadt, Deutschland
ORCID: 0000-0002-7588-4395

Frick, Claudia
claudia.frick@th-koeln.de
Technische Hochschule Köln, Deutschland
ORCID: 0000-0002-5291-4301

Wissenschaftskommunikation ist nicht erst seit der letzten DHd-Jahrestagung ein Thema in den digitalen Geisteswissenschaften. Sie bezeichnet die Form der Kommunikation, „bei der Wissenschaft und Gesellschaft miteinander über wissenschaftliche Fragestellungen in Austausch treten“ (Frick und Seltmann, 2023). Einen speziellen Fall der externen Wissenschaftskommunikation (vgl. u. a. Frick et al., 2021 sowie Seltmann, 2023) nimmt der *Referenzrahmen für eigenständige digitale Wissenschaftskommunikation durch Forschende* in den Blick. Er möchte Kompetenzen und Herausforderung dieser Form transparent machen und damit Anhaltspunkte liefern, die in der Lehre, in Fortbildungen sowie für Forschende selbst relevant sind. Die

Kompetenzen liegen dabei auf methodischer, technischer sowie sozialer Ebene. Die Betrachtung eben jener Kompetenzen fehlt bisher zumeist in der wissenschaftlichen Ausbildung, geschweige denn, dass sie bisher systematisch erfasst wäre.

Dafür setzt der Referenzrahmen in zehn Bereichen verschiedene Niveaustufen von disziplinenübergreifenden Teilkompetenzen an, ähnlich wie es beispielsweise auch der Gemeinsamen Europäischen Referenzrahmen für Sprachen (Trim und Quetz, 2001) handhabt. Die Niveaustufen bedienen dabei *passive (elementare) Kompetenzen* (A1/A2), *aktive (selbständige) Kompetenzen* (B1/B2) sowie *interaktive (nachhaltige) Kompetenzen* (C1/C2). Natürlich sind auch disziplinspezifische Kompetenzen für gute Wissenschaftskommunikation notwendig, diese sind jedoch nicht Teil des Referenzrahmens, sondern werden vorausgesetzt. Die zehn betrachteten Bereiche umfassen folgende:

- Impulse geben
- Plattform nutzen
- Einheit produzieren
- Community managen
- Kanal konzipieren
- Rolle leben
- Zielgruppe erreichen
- Ton treffen
- Wissenschaft öffnen
- Mit emotionalen Herausforderungen umgehen

Dabei stehen die verschiedenen Kompetenzbereiche miteinander in Verbindung und ergänzen sich gegenseitig (vgl. Figure 1). Durch die schematische Darstellung werden zudem Wechselwirkungen und Voraussetzungen über Teilkompetenzen hinweg detailliert verdeutlicht. Der Referenzrahmen stellt einen theoretischen Rahmen dar, den es noch empirisch zu erproben gilt. Die theoretische Fundierung beschreibt jeweils für einen Kompetenzbereich die entsprechenden Teilkompetenz in den sechs Niveaustufen, formuliert eine passende Frage und fügt ein Beispiel an. Daran schließen sich drei Fallbeispiele an, in denen jeweils die verschiedenen Kompetenzniveaus Anwendung finden.

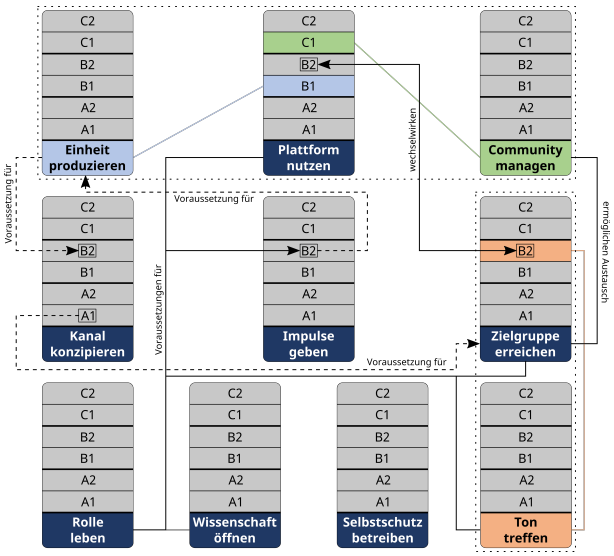


Figure 1: Schematische Darstellung des Referenzrahmens für eigenständige digitale Wissenschaftskommunikation mit seinen zehn Teilkompetenzen und deren Verknüpfungen und Abhängigkeiten. [aus Frick und Seltmann, 2023]

Da ein Referenzrahmen getestet werden muss, wurde sich entschlossen, ihn zur Diskussion zu stellen und im Open-Public-Peer-Review-Verfahren (vgl. ZfdG, 2023) bei der Zeitschrift für digitale Geisteswissenschaften als Working Paper zu veröffentlichen. Denn wer könnte besser beurteilen, wie hilfreich und richtig der Referenzrahmen ist als die Community, die Wissenschaftskommunikation betreibt, unterstützt oder begleitet? Vorteile eines solchen Reviewverfahrens, wie es Pöschl (2010; 2012) beschreibt, sind, dass es die wissenschaftliche Diskussion fördert, die Wirksamkeit und Transparenz wissenschaftlicher Qualitätssicherung maximiert, eine kurzfristige Veröffentlichung neuer wissenschaftlicher Erkenntnisse ermöglicht sowie diese Veröffentlichungen frei zugänglich macht (vgl. Van Edig, 2016, 29), auch wenn der letzte Aspekt wohl allgemein auf Open-Access-Publikationen unabhängig des verwendeten Reviewverfahrens zutrifft. Betrachtet man das Open Public Peer Review genauer, so lässt sich feststellen, dass hier verschiedene Definitionen von Open Peer Review miteinander verbunden werden: Es werden zum einen *Open Reports*[1] verwendet, die ebenso eine *Open Interaction*[2] ermöglicht und als *Open Pre-Review Manuscripts*[3] durchgeführt wird. Es handelt sich zum anderen um eine *Open Participation*[4] und ebenso um *Open Identities*[5] (vgl. Ross-Hellauer, 2017, 7) – mit der Einschränkung, dass die verwendeten Nutzernamen nicht immer Rückschluss auf die dahintersteckende Person zulassen.

Ein Beispiel für funktionierendes Open Peer Review in interner Wissenschaftskommunikation zeigt beispielsweise Frick (2020), indem sie Peer-Review-Verfahren von Studien während der Corona-Pandemie kritisch betrachtet. Sie schließt ihre Ausführungen mit der Erkenntnis, dass sich mit der Veröffentlichung von Preprints und mit der Verwendung von Open-Peer-Review-Verfahren „schnelle Verbreitung neuer wissenschaftlicher Erkenntnisse und wis-

senschaftliche Begutachtung […] gegenseitig bereichern [können]“ (Frick, 2020).

Das Poster schließt sich der Betrachtung von Open-Peer-Review-Verfahren an und betrachtet den Use Case des beschriebenen Referenzrahmens. Es gibt nicht nur einen Überblick über den Referenzrahmen selbst, es zeigt auch den Einfluss dieses Reviewprozesses auf das Working Paper. Es beantwortet die Fragen, welche Art von Kommentaren annotiert wurde, ob sich Diskussionen innerhalb der Community entsponnen und wie die Anmerkungen in die zweite Version des Referenzrahmen einflossen.

Fußnoten

1. Bei *Open Reports* werden die Reviews mit dem Artikel gemeinsam veröffentlicht (vgl. Ross-Hellauer, 2017, 7).
2. In der *Open Interaction* werden direkte Unterhaltungen zwischen Autor:innen und Reviewer:innen bzw. untereinander ermöglicht (vgl. Ross-Hellauer, 2017, 7).
3. Ein *Open Pre-Review Manuscript* macht das Manuskript vor jeglicher Form von formalen Peer Review als Preprint öffentlich verfügbar (vgl. Ross-Hellauer, 2017, 7).
4. Mit der *Open Participation* wird die weitere Community in den Reviewprozess eingeladen (vgl. Ross-Hellauer, 2017, 7).
5. Bei *Open Identities* sind die Identitäten von sowohl Autor:innen als auch Reviewer:innen öffentlich (vgl. Ross-Hellauer, 2017, 7).

Bibliographie

Frick, Claudia. 2020. „Peer-Review im Rampenlicht: Ein prominentes Fallbeispiel“. Informationspraxis 6 (2). 10.11588/ip.2020.2.74406.

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. 2023. „Referenzrahmen für eigenständige digitale Wissenschaftskommunikation durch Forschende“. Zeitschrift für digitale Geisteswissenschaften, Nr. 3. 10.17175/WP_2023B.

Pöschl, Ulrich. 2010. „Interactive Open Access Publishing and Public Peer Review: The Effectiveness of Transparency and Self-Regulation in Scientific Quality Assurance“. IFLA Journal 36 (1): 40–46. 10.1177/0340035209359573.

———. 2012. „Multi-Stage Open Peer Review: Scientific Evaluation Integrating the Strengths of Traditional Peer Review with the Virtues of Transparency and Self-Regulation“. Frontiers in Computational Neuroscience 6. https://www.frontiersin.org/articles/10.3389/fncom.2012.00033.

Trim, John L. M. und Jürgen Quetz. 2001. Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Deutsch als Fremdsprache. Berlin: Langenscheidt.

Ross-Hellauer, Tony. 2017. „What Is Open Peer Review? A Systematic Review“. F1000Research. 10.12688/f1000research.11369.2.

Seltmann, Melanie Elisabeth-H. 2023. „#PublicDH oder doch nur #WissKomm?“ Trier, Luxemburg. 10.5281/zenodo.7715494.

Van Edig, Xenia. 2016. „Interactive Public Peer ReviewTM: an innovative approach to scientific quality assurance“. In Positioning and Power in Academic Publishing: Players, Agents and Agendas, 28–33. IOS Press. 10.3233/978-1-61499-649-1-28.

ZfdG. 2023. „O wie Open Public Peer Review“. Zeitschrift für digitale Geisteswissenschaften. 2023. https://zfdg.de/o-wie-open-public-peer-review.

Partizipation, Co-Kreation und Citizen Science – Zwischen Grundlagenforschung und digitaler Kulturvermittlung

Brinkmann, Hanna
hanna.brinkmann@donau-uni.ac.at
Universität für Weiterbildung Krems, Österreich

Grebe, Anja
anja.grebe@donau-uni.ac.at
Universität für Weiterbildung Krems, Österreich

Lopin, Melanie
melanie.lopin@museumsverein-korneuburg.at
Stadtmuseum Korneuburg, Kulturvermittlung, Österreich

1. Einleitung

Partizipation, Co-Kreation und Citizen Science bzw. Citizen Humanities sind Ansätze, die in den Digital Humanities in den letzten Jahren verstärkt gefordert und gefördert wurden (Heinisch 2020). In Museen werden sie seit vielen Jahrzehnten erprobt (Pierroux et al., 2020) und kommen in der Kulturvermittlung im Bereich Community Building oder als Outreach-Strategie zum Einsatz. Intergenerative

Formate versprechen im Hinblick auf kollaborative Wissensgenerierung und die co-kreative Entwicklung userzentrierter, digitaler Vermittlungsformate von besonders hoher zukünftiger Bedeutung zu sein.

2. Projektvorstellung

Das Poster stellt die Ergebnisse und Learnings aus dem Pilotprojekt "MuseumsMenschen im Schaudepot" vor (Brinkmann et al., 2022), welches das Ziel verfolgte, durch die oben genannten Ansätze eine bereits bestehende, institutionenübergreifende Web-App für 10 Stadtmuseen in Niederösterreich, um weitere Inhalte anzureichern und in den einzelnen Prozessschritten zu evaluieren. Dieses digitale, partizipative, Grundlagenforschung und Kulturvermittlung verbindende Projekt an der Schnittstelle von analog und digital, wurde vom Land Niederösterreich, Abteilung Kunst und Kultur, gefördert und als Kooperation zwischen der Universität für Weiterbildung Krems und dem Stadtmuseum Korneuburg durchgeführt.

Das Engagement von Bürger:innen für das lokale und regionale Kulturerbe und die Dokumentation kultureller Daten spielte bereits bei der Gründung von Stadtmuseen eine große Rolle, wie das Forschungsprojekt „MuseumsMenschen“ (2017-2020)[1] der Universität für Weiterbildung Krems zeigt, welches den Ausgangspunkt für die webbasierte App darstellt (Grebe, 2019).

3. Umsetzung

Die Web-App „MuseumsMenschen“ vernetzt das physische Objekt, welches vor Ort im Museum betrachtet wird, mit spielerisch aufbereiteten Informationen in leicht verständlicher Sprache. So soll der Zugang zum einzelnen Objekt, aber auch zur Sammlung insgesamt erleichtert und eine intensivere Auseinandersetzung mit den Objekten erreicht werden. Auf diese Weise entsteht ein hybrider Raum zwischen analog und digital (Løvlie et al., 2021; Ergin, 2022).

Ausgehend vom bereits bestehenden Dialog-Ansatz der Web-App, welche die Inhalte im Chat-Format aufbereitet, wurde durch das partizipative Kulturvermittlungsprojekt „MuseumsMenschen im Schaudepot“ eine Brücke zwischen den Generationen geschlagen: Teilnehmer:innen waren eine Gruppe von Jugendlichen, die ihr Wissen und Know-how im Umgang mit digitalen Medien einsetzten, und Senior:innen, die im Sinne der „oral history“ Objektgeschichten lebendig werden ließen. Die Jugendlichen waren über einen Medienaufruf in der lokalen Presse und im Radio rekrutiert worden, während sich die Senior:innen über persönliche Kontakte und durch den Museumsverein fanden. Generationenspezifisches, lokales Wissen zu konkreten Objekten aus der Museumssammlung wurde im Rahmen von Workshops geteilt und gemeinsam in das Medium Web-App überführt. Im Zuge dieses Prozesses wirkten Bürger:innen somit auch an einer Kuratierung der Objekte im digitalen Raum mit, was anschließend allen Museumsbesucher:innen und Nutzer:innen der Web-App zur Verfügung gestellt wurde.

4. Learnings und Ausblick

Während die partizipative Erweiterung der Web-App bereits in einem weiteren Stadtmuseum in Niederösterreich in einer ähnlichen Form umgesetzt werden konnte, zeigte die 2022 durchgeführte Nutzer:innenstudie zur Web-App, dass diese bislang nur zögerlich verwendet wird (Brinkmann 2023). Einige aus dieser Studie gewonnenen Erkenntnisse sollen kurz aufgeführt werden. Um das Potenzial der Web-App voll entfalten zu können, sind auch die Museen gefragt: Museumsintern könnte die Kommunikation über die Existenz der Web-App und eine kurze Information für Mitarbeiter:innen an der Kasse oder in der Aufsicht zu einer stärkeren Nutzung durch die Besucher:innen ein erster Schritt sein. Mit der Aushändigung des zur Verfügung gestellten Flyers mit dem aufgedruckten QR-Code ist ein zweiter wichtiger Schritt getan, die Besucher:innen werden so aktiv auf die Web-App hingewiesen und haben den direkten Zugang persönlich in die Hand gedrückt bekommen. Hilfreich ist es auch, die Objekte, welche in der App besprochen werden, im Museum zu kennzeichnen, wie dies bereits in einigen der beteiligten Museen umgesetzt wurde. Auch wenn die Web-App einen BYOD (bring your own device)-Ansatz verfolgt, kann es sinnvoll sein, im Haus Tablets anzubieten (siehe Abb. 1), auf denen die App-inhalte größer zu sehen sind, was sich gerade für ältere Besucher:innen bewährt hat.

Abb. 1 Foto © Museumsverein Korneuburg

Der partizipative, intergenerative Ansatz des Pilotprojekts soll im inter- und transdisziplinären Folgeprojekt „Industriekultur im Dialog“ weiterentwickelt werden, welches sich mit der Alten Werft in Korneuburg befasst. Das Archiv der 1993 geschlossenen Werft, einschließlich zahlreicher Objekte, befindet sich im Stadtmuseum Korneuburg. Es soll nun digitalisiert und mit Hilfe von Citizen Humanists aufgearbeitet werden und dabei in intergenerativen Teams von Schüler:innen der BHAK Korneuburg und ehemaligen „Werftler:innen“ auch Eingang in die Web-App finden.

Fußnoten

1. Das FTI-Projekt „MuseumsMenschen“ des Departments für Kunst- und Kulturwissenschaften der Universität für Weiterbildung Krems wurde in Kooperation mit dem Museumsmanagement Niederösterreich und den zehn ältesten niederösterreichischen Stadtmuseen durchgeführt: Rollettmuseum der Stadt Baden (1806/1876), Museum St. Peter an der Sperr, Wiener Neustadt (1824), Museum Retz (1833), Stadtmuseum Korneuburg (1863), Stadtmuseum St. Pölten (1879), Stadtmuseum Melk (1879/1880), museumkrems (1884/1889), Krahuletz-Museum Eggenburg (1889/1901), Zeitbrücke-Museum Gars am Kamp (1898/1902), Stadtmuseum Zwettl (1900).

Bibliographie

Brinkmann, Hanna. 2023. „Objektgeschichten aus niederösterreichischen Stadtmuseen. Zur Nutzung, Evaluierung und Erweiterung der MuseumsMenschen-Web-App“. In *Österreich Geschichte, Literatur, Geographie 3. Neue Perspektiven auf 100 Jahre Kulturerbe Niederösterreich* hg. von Anja Grebe, 9-20.

Brinkmann, Hanna, Anja Grebe, Melanie Lopin. 2022. „Ein gemeinsames Online-Tool als Chance für Kooperation und Kollaboration“. *neues museum* 3: 34-36.

Ergin, Gamze. 2022. "Museums in the Digital Age: Hybrid Museum Experience." In *Multidisciplinary Perspectives Towards Building a Digitally Competent Society*, hg. von Sanjeev Bansal, Vandana Ahuja, Vijit Chaturvedi, Vinamra Jain, 51-69. Hershey, PA: IGI Global. DOI 10.4018/978-1-6684-5274-5.ch003.

Grebe, Anja. 2019. “MuseumsMenschen - Stadtmuseen im Fokus.“ *neues museum* 3: 66-68.

Heinisch, Barbara. 2020. “Citizen Humanities as a Fusion of Digital and Public Humanities?”. *magazén*, 1(2): 143-180. DOI 10.30687/mag/2724-3923/2020/02/001.

Løvlie, Anders Sundnes, Karin Ryding, Jocelyn Spence, Paulina Rajkowska, Annika Waern, Tim Wray, Steve Benford, William Preston, Emily Clare-Thorn. 2021. “Playing Games with Tito: Designing Hybrid Museum Experiences for Critical Play.” *Journal on Computing and Cultural Heritage* , 14 DOI 6 10.1145/3446620 .

Palmyre Pierroux, Per Hetland, Line Esborg. 2020. “Traversing citizen science and citizen humanities Tacking stitches”. In *A history of participation in museums and archives: traversing citizen science and citizen humanities*, hg. von Per Hetland, Palmyre Pierroux, Line Esborg, 3-23. Abingdon, Oxon: Routledge.

PUDEL: Paving the Way for Pawsome Data Models and Vocabularies in the Academic Community

Goldhahn, Dirk
goldhahn@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0003-1681-567X

Kretschmer, Uwe
kretschmer@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0002-3685-6519

Muehleder, Peter
muehleder@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0001-6593-5673

Naether, Franziska
naether@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0003-4652-6836

Becker, Anja
becker@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0003-4202-1566

Graiff, Cecilia
graiff@saw-leipzig.de
Sächsische Akademie der Wissenschaften zu Leipzig, Deutschland
ORCID: 0000-0002-2226-7174

Making knowledge accessible in a systematic and standardized manner to allow for extensive use and reuse is one of the key challenges in the Digital Humanities (DH). The process of data modeling (Flanders / Jannidis 2015) to adequately capture knowledge has become a vital part of this task. Yet, it is typically not trivial for researchers to document and publish their models or vocabularies in order to make them available for reuse following principles such as FAIR (Wilkinson et al. 2016) and, consequently, findable etc. for other researchers to discover them.

The "PUDEL" project (Publikationsdienst für wissenschaftliche Datenmodelle und Vokabulare, based at the Saxon Academy of Sciences and Humanities Leipzig) focuses on developing a comprehensive publication service that caters to the specific requirements of data models and vocabularies. The project recognizes the critical role of these resources in facilitating data interoperability, enabling data reuse, and promoting collaborative research across various disciplines.

Therefore, the core objective of PUDEL is to establish a comprehensive platform that enables researchers to publish and document their data models and vocabularies in various formats (XML schemas, RDF-based ontologies, SKOS vocabularies etc.) in a standardized and easily accessible manner, taking into account discipline-specific requirement of such models and their respective meta data. Existing services, like the Vocabs Service of ACDH-CH (Austrian Academy of Sciences 2023) only support the publication of vocabularies in SKOS format, or require the data model to already be published online to be able to reference it, as it is the case with VoCoReg (Fraunhofer Society 2023). PUDEL tries to offer a more inclusive access and focus here, avoiding the trap of being a mere ‘data silo’.

The platform offers a range of robust tools and workflows for documentation, validation, and versioning in a Git-based file storage backend. This systematic approach guarantees the long-term sustainability and usability of the published resources, allowing researchers to confidently share and disseminate their data models and vocabularies.

PUDEL is built around an intuitive web service that acts as an entry point, and various ‘middleware’ services in order to validate data models and create representations based on established best practices (Garijo / Poveda-Villalón 2020; Semantic Web Deployment Working Group 2008).

An initial project phase of PUDEL was funded by the Saxon State Ministry of Science and Arts within the framework of SaxFDM, the Saxon initiative for research data management. During this stage, two main aspects were addressed:

1. the provision of a prototype of the service, and
2. the investigation of the state of the art concerning related projects and tools.

The first task was successfully tackled by developing a functional prototype that provides basic functionality and workflows (to be presented at the Dhd). The second point included creating an overview of projects, which offer similar services in order to identify important features PUDEL needs to offer in order to fulfill typical use cases. In addition, tools and services were investigated that could be utilized during the implementation of the service.

At this point, a second project phase is in preparation. It is planned to extend the prototype and to address the following tasks:

(1) One of the key features of PUDEL will be exploration and discoverability of data models. The platform will incorporate comprehensive search and retrieval mechanisms, allowing researchers to efficiently explore and access specific vocabularies and ontologies etc. relevant to their research interests or project requirements. Additionally, PUDEL will support automatic publication on the open repository Zenodo (European Organization for Nuclear Research, OpenAIRE 2013).

(2) The service is developed as a free and open-source software (FOSS), making it possible for institutions to run their own instance of PUDEL and allowing them to publish data models within their own name spaces. As a set of software tools, it should also be simple to maintain and update. Here, PUDEL follows the approach of minimal computing (Sayers 2016) whenever possible, using as few resources as possible and avoiding a complex system architecture. Fur-

thermore, OpenAPI is used to create a standardized documentation of the service APIs.

(3) In addition to its technical infrastructure, PUDEL recognizes the importance of community building and knowledge dissemination. The project will organize a series of training events, conference talks, and coffee lectures to raise awareness about the significance of sharing data models in research – and addressing more advanced user such as developers. By advancing data publication and interoperability, PUDEL contributes to the acceleration of scientific progress, promotes open science principles, and supports the academic community in their research endeavors.

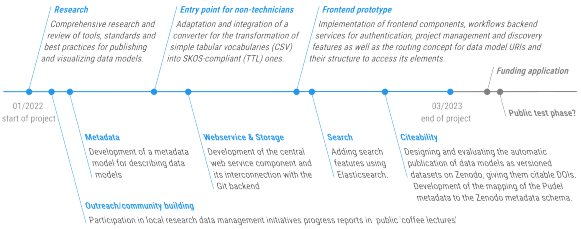


Fig. 1: Timeline of the PUDEL project

Bibliographie

Austrian Academy of Sciences , Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), "Vocabs services", <https://vocabs.dariah.eu/> [17.07.2023].

European Organization for Nuclear Research, OpenAIRE (2013): "Zenodo", in: CERN Publ., doi: 10.25495/7GXK-RD71, < https://www.zenodo.org/ > [17.07.2023].

Flanders, Julia / Jannidis, Fotis (2015): "Knowledge Organization and Data Modeling in the Humanities", < https://nbn-resolving.org/html/urn:nbn:de:bvb:20-opus-111270 > [17.07.2023].

Fraunhofer Society , "VoCol Service on VoCoREG", <https://www.vocoreg.com/> [17.07.2023].

Garijo, Daniel / Poveda-Villalón, María (2020): "Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web", < https://doi.org/10.48550/arXiv.2003.13084 > [17.07.2023].

OpenAPI Initiative (2022): "OpenAPI Specification v3.1.0", in: The Linux Foundation (Publ.), < https://spec.openapis.org/oas/latest.html > [17.07.2023]

Sayers, Jentery (2016): "Minimal Definitions", in: Minimal Computing Working Group of GO::DH (Publ.), < https://go-dh.github.io/mincomp/thoughts/2016/10/02/minimal-definitions/ > [17.07.2023].

Semantic Web Deployment Working Group (2008): "Best Practice Recipes for Publishing RDF Vocabularies" (Work in Progress), < https://www.w3.org/TR/swbp-vocab-pub/ > [17.07.2023].

Wilkinson, Mark D. et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship", in: Sci Data 3: 160018, < https://doi.org/10.1038/sdata.2016.18> [17.07.2023].

Quo tendimus? Visualisierungen in digitalen Editionen am Beispiel der „Hybridedition der deutschsprachigen Werke des Martin Opitz“

Schwaß, Susann
schwass@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Schulz, Daniela
schulz@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0003-3167-5089

Martin Opitz (1597–1639) war ein schlesischer Dichter aber auch Theoretiker und Diplomat, der heute als eine der Schlüsselfiguren der europäischen Literaturgeschichte gilt. Er unternahm vielfach diplomatische Missionen, die sich durchaus förderlich auf die Etablierung eines europaweiten Netzwerks von Intellektuellen auswirkten, zu dem beispielsweise Georg Michael Lingelsheim, Matthias Bernegger und Hugo Grotius gehörten. Dieses Netzwerk, genauso wie die politische Situation des Dreißigjährigen Krieges beeinflussten sein literarisches Schaffen nachhaltig.

Seine deutschsprachigen Werke werden im Rahmen eines DFG-Projektes seit 2018 vom Lehrstuhl für Literaturgeschichte der Frühen Neuzeit der Universität Tübingen und der Herzog August Bibliothek Wolfenbüttel in Absprache mit dem Anton Hiersemann-Verlag (Stuttgart) und der Bibliothek des Literarischen Vereins (Stuttgart) digital aufbereitet und zum Teil neu ediert. Zwischen 1968 und 1990 entstand durch George Schulz-Behrend eine kritische Ausgabe, in der Opitz’ deutschsprachige und lateinische poetisch-literarische Werke bis 1630 Berücksichtigung fanden. Die letzten Lebensjahre von Opitz sowie posthum erschienene Ausgaben fehlten, da die Edition nach dem insgesamt siebten Teilband (Band IV, 2) abbricht. Diese Bände wurden im Projektkontext retrodigitalisiert. Das Projekt strebt außerdem die Vervollständigung der Schulz-Behrend-Ausgabe um die noch fehlenden Texte an. 2021 ist bereits Band 5 (Opitz, 2021) erschienen, Band 6 (Opitz, 2023) wird noch 2023 in den Druck gehen. Das Projekt ist als Hybridedition konzipiert, sodass nach Ablauf einer Moving Wall von

zwei Jahren nach Veröffentlichung der Printausgabe die neuen Bände in einem übergreifenden Portal online publiziert werden.

Gerade durch seine zahlreichen Reisen und das vorhandene Netzwerk, aber auch durch die vielen Personen und Orte, die in Opitz' Texten Erwähnung finden, bietet sich im Projektkontext großes Potenzial für die visuelle Aufbereitung der annotierten Daten. Die Rolle von Visualisierungen in digitalen Editionen und die unterschiedlichen Zwecke (z.B. illustrativ oder als Erkenntnismittel) zu denen sie eingesetzt werden können, sind in den letzten Jahren vielfach diskutiert worden, doch scheint es übergreifend immer noch an einer etablierten Praxis zu mangeln.[1] Die *Spatial Humanities* z.B. beschäftigen sich mit dem Einbezug raumbezogener Daten in die geisteswissenschaftliche Forschung. Aus diesen Kontexten sind zahlreiche Werkzeuge und Hilfsmittel hervorgegangen, die für die Aufbereitung von Kartenansichten in der Opitz-Edition verwendet werden.

So wurde beispielsweise eine *leaflet*-Karte basierend auf der digitalisierten Portolankarte *Nova Et Exquisita Descriptio Navigationum Ad Praecipuas Mundi Partes* von Nicolas de Nicolay entworfen, welche die Handlungsorte der *Argenis* Bände visualisiert. Eine weitere Kartenansicht widmet sich den Lebens- und Schaffensstationen Opitz'. Dabei wurden unter anderem mit *QGIS* historische Kartendigitalisate auf aktuelle Georeferenzsysteme transformiert und mit *geojson.io* durch Geotagging historische Gebiete auf modernen Kartenansichten integriert. Mit dem *DARIAH-DE Geo-Browser* soll ebenfalls eine Überblickskarte aller Orte im Ortsregister bereitgestellt werden.

Neben den Plätzen, die eindeutig mit realweltlichen Orten korrelieren, finden sich in Opitz Dichtung aber auch vergessene, erfundene und mythische Handlungsorte, die einer spezifischen Auseinandersetzung bedürfen. Visualisierungen dieser textlichen Raumkonfigurationen können mitunter helfen, frühneuzeitliches Erzählen unter neuen Aspekten zu erkunden.

Im Rahmen des Opitz-Projektes wurden und werden verschiedene Formen der Visualisierung erprobt und damit auch auf ihren sinnhaften Einsatz hin überprüft. Was kann anhand der Daten visualisiert werden? Welche Darstellungsformen sind angemessen und hilfreich? Inwieweit können Aspekte des werte-sensitiven Designs[2] Berücksichtigung finden? Wie ist mit Unsicherheiten umzugehen?[3]

Datengrundlage sind zum einen die retrodigitalisierten Editionstexte von Schulz-Behrend, zum anderen die ‚Neubände'. Aufgrund unterschiedlicher Workflows im Verlauf des Projekts, wurden die Editionstexte mit unterschiedlichen Verfahren (teils MD2TEI, teils Word2TEI) transkribiert und transformiert. Named Entities sind hauptsächlich händisch ausgezeichnet worden und fließen im Orts- und Personenregister, auf das sich die Visualisierungen stützen, zusammen. Diese Daten werden mit Daten aus dem Leben und Schaffen von Opitz flankiert.

Die Hybridedition möchte nicht nur einen neuen und umfassenden Blick auf die Texte bieten, sondern durch die Visualisierung von Daten, Forschenden einen erweiterten Zugang zu Opitz und seinem Werk ermöglichen. Das Poster gewährt einen Einblick in das Projekt, seine Inhalte und Workflows. Der Schwerpunkt liegt auf dem Bereich der Visualisierungen. Dabei sollen sowohl technische Möglichkeiten vorgestellt, als auch die angewendeten Ansätze des werte-sensitiven Designs von Visualisierungen am Beispiel historischer Geodaten zur Diskussion gestellt werden.

Fußnoten

1. Vgl. zur Begriffsbestimmung den aktuell unter Open Public Peer Review stehenden Beitrag von Linda Freyberg (Freyberg, 2023) im diskursiven Glossar. Nach Schaal und Lancaster (Schaal und Lancaster, 2016) fehlt es im DH-Bereich an einer Best Practice, an der sich Forschende orientieren können, um erkenntnisfördernde Visualisierungen zu erzeugen.
2. Vgl. zu Value Sensitive Design im Kontext der DH Leyrer, 2023.
3. Vgl. hierzu Kräutli und Davies, 2013.

Bibliographie

Freyberg, Linda. 2023. „Visualisierung." In Begriffe der Digital Humanities. Ein diskursives Glossar, hg. von der AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum e. V. (= Zeitschrift für digitale Geisteswissenschaften / Working Papers, 2). 10.17175/wp_2023_014 .

Kräutli, Florian und Stephen Boyd Davis. 2013. „Known Unknowns: Representing Uncertainty in Historical Time". In Electronic Visualisation and the Arts. Swindon u. a., 61–68. 10.14236/ewic/EVA2013.16 .

Leyrer, Katharina. 2023. „Bye, Bye, Bias! Digital-Humanities-Projekte wertebasiert gestalten mit Value Sensitive Design." In Fabrikation von Erkenntnis – Experimente in den Digital Humanities, hg. von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, Ulrike Wuttke. Wolfenbüttel (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5). Erstveröffentlichung 08.09.2021. Version 2.0 vom 21.03.2023. 10.17175/sb005_003_v2 .

Martin Opitz. Gesammelte Werke. Band 5. Die Werke von 1630 bis 1633, hg. von Gudrun Bamberger und Jörg Robert (Bibliothek des literarischen Vereins in Stuttgart 355), Stuttgart: Hiersemann.

Martin Opitz. Gesammelte Werke. Band 6. Der Argenis anderer Theyl, hg. von Gudrun Bamberger und Jörg Robert (Bibliothek des literarischen Vereins in Stuttgart 358), Stuttgart: Hiersemann.

Schaal, Gary S., Kelly Lancaster. 2016. „Ein Bild sagt mehr als 1000 Worte? Visualisierungen in den Digital Humanities." Digital Classics Online 2,3: 5–22. https://doi.org/10.11588/dco.2016.0.30875 .

Quo vadis digitised newspapers and radio? Next steps for the integration of western European collections via impresso II.

Bunout, Estelle
estelle.bunout@uni.lu
Universität Luxemburg, Luxemburg, UL

Düring, Marten
marten.during@uni.lu
Universität Luxemburg, Luxemburg, UL

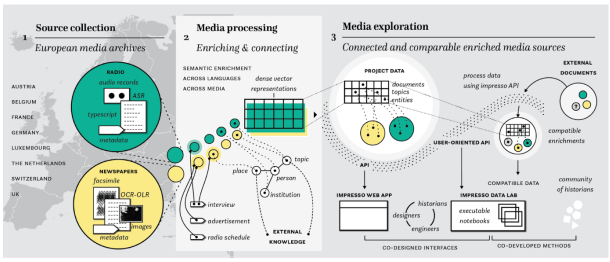
Clematide, Simon
simon.clematide@uzh.ch
Universität Zürich, UZH, Schweiz

Ehrmann, Maud
maud.ehrmann@epfl.ch
Ecole Polytechnique Fédérale de Lausanne, EPFL, Schweiz

Guido, Daniele
daniele.guido@uni.lu
Universität Luxemburg, Luxemburg, UL

Ruppen Coutaz, Rapahëlle
raphaelle.ruppencoutaz@unil.ch
Universität Lausanne, UNIL, Schweiz

Beelen, Kaspar
kaspar.beelen@turing.ac.uk
The Alan Turing Institute, Großbritannien



Over the past few years, interdisciplinary research projects such as Oceanic Exchanges, Living with Machines, Numapresse or NewsEye (see links below) all worked towards the semantic enrichment, integration and computational analysis of historical newspaper collections across institutional, national and linguistic boundaries. Outputs include shared tasks to advance the quality of semantic enrichment of historical text (Ehrmann, 2020), symposia to identify future research directions (Ehrmann, 2023), as well as a wide range of original historical research using computational methods together with reflections on their application (Keck et al., 2022; Oberbichler et al., 2021; Bunout et al., 2022). Most recently, the HAICu project announced the application of machine learning technologies to Dutch cultural heritage collections across different modalities on a national scale.

Against this background, we propose to present a new project: “impresso - Media Monitoring of the Past II. Beyond Borders: Connecting Historical Newspapers and Radio”, which builds on the first *impresso* project that compiled and semantically enriched a corpus of Swiss and Luxembourgish newspapers based on the collections of project partners such as the national libraries of Switzerland and Luxembourg, Neue Zürcher Zeitung, Le Temps, the Valais State Archives and the Swiss Economic Archive.

The first *impresso* application integrates new opportunities offered by semantic enrichments such as word embeddings, topic modelling, and the automated detection of text reuse and languages for content search and discovery as well as comparative and critical perspectives on the available data. Its design was driven by the principles of co-design, generosity (the provision of multiple entry points to the collection) and transparency (reflections on tools, document processing and data quality) (preprint Düring et al., 2023). The application is freely accessible.

This first project has demonstrated the added value of integrating sources from different languages into the same system in order to better facilitate their joint exploration and comparison. The corpus of the second *impresso* project will build on the first in several ways: First, it will broaden the corpus to include radio alongside newspaper sources. Second, it will expand to a Western European scale in partnership with national and state-level libraries as well as archives dedicated to the preservation of audiovisual materials together with new partners in Austria, Belgium, France, Germany, Luxembourg, the Netherlands and Switzerland. Many of our source material will be in German and be connected to contents in French, Dutch, English etc.

Third, these collections will be transformed from noisy and heterogeneous text in multiple languages into rich data, integrated and represented in a shared vector space. Fourth, it is our goal to develop an open and generic technological framework for the seamless exploration of semantically enriched and connected media archives. Fifth, the project will benefit from five (media) historical case studies which will exploit the newly available data under the shared research theme “influences” but also active engagement with research communities in digital humanities and history. Sixth,

and finally, *impresso* will seek to actively support the research community by offering relevant data and interfaces for the exploration of its collections.

With the proposed poster we also seek to share a call for collaboration: Ingrained in the spirit of the *impresso* project is the belief in interdisciplinarity and collaborative design on the intersection between computational linguistics, computer science, history, digital humanities and design. We hope to reach out to researchers interested in contributing to the paradigm shift in the processing, representation and analysis of historical document collections.

Project links:

impresso project : https://impresso-project.ch/
Oceanic Exchanges : https://oceanicexchanges.org/
Living with Machines : https://livingwithmachines.ac.uk/
Numapresse : https://numapresse.hypotheses.org/
NewsEye : https://www.newseye.eu/

Bibliographie

Bunout, Estelle, Maud Ehrmann, Frédéric Clavert, eds. Digitised Newspapers – A New Eldorado for Historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization. Studies in Digital History and Hermeneutics. Berlin: De Gruyter, 2022.

Ehrmann, Maud, Marten Düring, Clemens Neudecker, Antoine Doucet. "Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292)." Dagstuhl Reports 12, no. 7 (2023): 112–79. https://doi.org/10.4230/DagRep.12.7.112.

Ehrmann, Maud, Matteo Romanello, Stefan Bircher, Simon Clematide. "Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers." In Advances in Information Retrieval, 524–32. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-45442-5_68.

Keck, Jana, Mila Oiva, Paul Fyfe. "Lajos Kossuth and the Transnational News: A Computational and Multilingual Approach to Digitized Newspaper Collections." Media History (2022): 1–18. https://doi.org/10.1080/13688804.2022.2146905.

Oberbichler, Sarah, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, Mikko Tolonen. "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, 2021, https://doi.org/10.1002/asi.24565.

Quo Vadis Fachbereiche und Schulen der DHd: Netzwerkanalyse der DHd Abstracts 2014-2023

Haider, Thomas Nikolaus
thomas.haider@uni-passau.de
Universität Passau, Deutschland
ORCID: 0000-0003-1522-4026

Gassner, Sebastian
sebastian.gassner@uni-passau.de
Universität Passau, Deutschland

Rehbein, Malte
malte.rehbein@uni-passau.de
Universität Passau, Deutschland
ORCID: 0000-0002-3252-0604

Einleitung

Die Digital Humanities (DH) sind ein wachsendes Forschungsfeld (Tang et al., 2017), und nach 10 Jahren DHd fragt die Konferenz in Passau 2024 nach ‚Quo Vadis DH?'. Wir möchten mit diesem Beitrag einen Überblick der Fächerlandschaft und des Zitierverhaltens der deutschsprachigen DH Community gewinnen. Dies verfolgen wir mit Methoden der Netzwerkanalyse, wobei wir uns insbesondere Ko-Keywords, Ko-Autorschaft, und Ko-Zitation ansehen. Zum Zitierverhalten werfen wir außerdem einen Blick auf die Länge der Bibliographien und das Alter der Zitationen.

Während bestehende bibliometrische Studien sich mit verwandten Bereichen wie der Computerlinguistik (Bollmann & Elliot, 2020), dem Management (Wieczorek et al., 2021) und den Digital Humanities international (Tang et al., 2017) sowie mit spezifischen Analysen einzelner DHd-Konferenzen (Henny-Kramer & Sahle, 2018; Hoenen, 2019) beschäftigt haben, fehlt ein umfassender Überblick der Zitationslandschaft der DHd-Konferenz. Diese Lücke motiviert unsere Studie, mit dem Ziel eine Bewertung und einen Einblick in die Zitationsdynamik innerhalb dieses wissenschaftlichen Forums zu geben.

Wir verfolgen folgende Analysen:

Ko-Keywords: Durch das Extrahieren und Analysieren von Keywords aus wissenschaftlichen Arbeiten können wir wiederkehrende Themen, gemeinsame Begriffe und Gruppen von verwandten Problemen identifizieren.

Die Beziehungen zwischen Keywords sollten Einblicke in die Teilgebiete der DH bieten.

Ko-Autorschaft: Die Zusammenarbeit ist das Herzstück des wissenschaftlichen Fortschritts, und Koautorenschaftsnetzwerke sind ein Instrument zur Untersuchung der Interaktionen zwischen Forscher:innen. Die Analyse von Koautorenschaftsmustern kann helfen, einflussreiche Forschungsgruppen, intellektuelle Zentren und Communities mit gemeinsamen Forschungsinteressen zu identifizieren. Die Aufdeckung dieser Netzwerke ermöglicht es uns, unterschiedliche Denkrichtungen, ihre geografische Verteilung und möglicherweise ihren Einfluss auf den wissenschaftlichen Fortschritt zu ermitteln.

Zitierverhalten und Ko-Zitation: In wissenschaftlichen Arbeiten werden typischerweise frühere Arbeiten zitiert, um die intellektuelle Abstammung von Ideen zu belegen und grundlegende Beiträge zu würdigen. Durch die Erstellung von Kozitationsnetzwerken können wir etwa einflussreiche Arbeiten aufspüren und nachverfolgen, und Einsicht in die Verbindungen zwischen verschiedenen Sub-Disziplinen bekommen.

Korpus

Als Hauptkorpus verwenden wir die DHd Abstracts von 2016 bis 2023, welche komplett in TEI XML vorliegen (https://github.com/DHd-Verband). Die Volltexte der Jahrgänge 2014 und 2015 liegen nur in .pdf vor. Die Information zu Keywords und Bibliographie ist in diesen frühen Jahrgängen inkonsistent und wird daher nicht verwendet. Allerdings verwenden wir für Ko-Autorschaft die Überblicks-XML (die alle Beiträge der jeweiligen Jahrgänge 2014 und 2015 vereint).

Methodik

Die Hauptaspekte dieses Papiers (Ko-Keywords, Ko-Autorschaft, Ko-Zitation) modellieren wir mit Netzwerkanalyse. Die Analysen bewegen sich auf Dokument-Ebene: Pro DHd-Abstract verwenden wir eine Liste mit Keywords, eine Liste mit Autor:innen, und eine Liste mit Literaturreferenzen.

Die Größe der Knoten ist durchgängig durch Betweenness Centrality bestimmt (wie wichtig ist ein Knoten für das gesamte Netzwerk), und die Kantenstärke ist durch die absolute Frequenz des gemeinsamen Vorkommens von zwei Knoten bestimmt (zum Beispiel zeigt eine Kante im Ko-Autorschaftsnetzwerk an wie oft zwei Autor:innen auf der DHd zusammen publiziert haben).

Für das Layout der Netzwerke verwenden wir Force Atlas 2, und die (Cluster)farben werden durch Modularity bestimmt.

Netzwerke

Keywords-Netzwerk

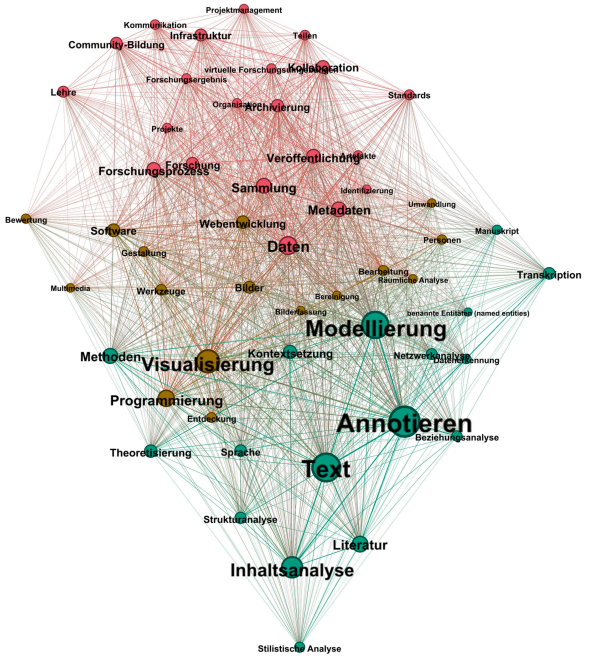


Abbildung 1: Ko-Keywords Netzwerk

Abildung 1 zeigt das Ko-Keywords Netzwerk. Die DHd-Abstracts verfügen über zwei separate Keyword Auszeichnungen, sog. `Keywords‘ und `Topics‘. Es zeigt sich, dass die `Keywords‘ eher sparse sind (da diese durch ein Freifeld eingegeben werden), und die `Topics‘ sehr dicht sind. Daher zeigen die Netzwerke der einzelnen Felder wenig aufschlussreiches. Wenn allerdings beide Annotationen auf Dokumentebene zusammengeführt werden, so zeigt sich, dass die Topics durch die Keywords sinnvoll miteinander verbunden werden. Der Übersicht halber wurden seltene Keywords gefiltert.

Wir finden einige Cluster: Rechts unten die Annotation von Text, insbesondere für die Inhalts- und Stilanalyse in der Literaturwissenschaft, mittig Modellierung und Visualisierung, dazu Methoden und Metadaten. Zwischen den ersten beiden Clustern die Bildwissenschaft. Mittig und Oben finden sich Daten, deren Sammlung, Veröffentlichung, und Archivierung. Oben die Forschung selbst, mit der dazugehörigen Community, Infrastruktur und Lehre.

Ko-Autorschaft Netzwerk

In Abbildung 2, dem Ko-Autorschafts-Netzwerk zeigen sich einige regionale Zentren. In grün mittig links die Verbindung Potsdam-Berlin (Trilcke, Fischer). Rechts in türkis Graz. In orange und grün links und zentral Stuttgart-Darmstadt-Hamburg. Links oben Würzburg-Trier

(Jannidis, Schöch). Oben in blau Leipzig-Regensburg (Burghardt). Zentral hellgrün findet sich ein dichtes Feld mit gemischten Affiliationen, darunter einige Kölner Namen.

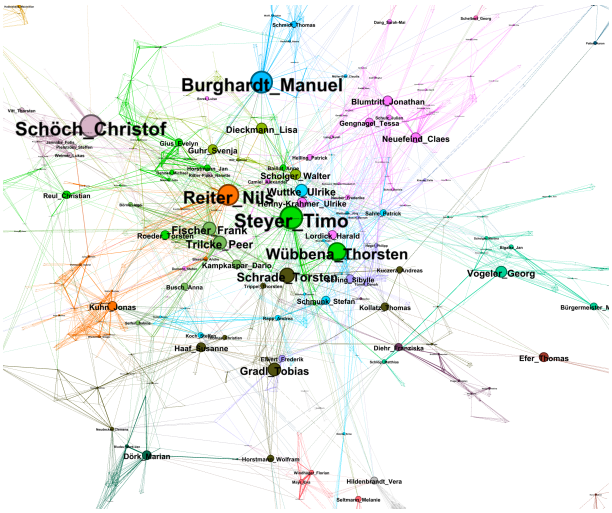


Abbildung 2: Ko-Autorschaft Netzwerk

Zitierverhalten & Ko-Zitation

Das Parsing von Referenzen/Zitationen fokusiert auf first-level Referenzen (also auf Artikel-Ebene, und nicht z.B. in welchem Sammelband etwas erschienen ist). Im Folgenden (Tabelle 1) sind die 20 am häufigsten zitierten Arbeiten aufgeführt. Darunter finden sich unter den Hauptautor:innen drei Frauen, Johanna Drucker, Birgit Hamp, und Verena Henrich. Die Themenwahl reicht von den FAIR Principles über Topic Modelling/Embeddings und Distant Reading über Editionswissenschaft und Korpora bis in die Bildwissenschaft. Es zeigt sich ein ähnliches Bild wie bei der Keywordanalyse.

Title	Year	Authors	Citations
The Fair Guiding Principles for Scientific Data Management and Stewardship	2016	Wilkinson et al.	23
Latent Dirichlet Allocation	2003	Blei et al.	18
Digitale Editionsformen	2013	Sahle	18
Distant Reading	2013	Moretti	14
TEI P5 Guidelines for Electronic Text Encoding and Interchange	(2007)	TEI Consortium	12
Probabilistic Topic Models	2012	Blei	12
Germanet: A Lexicalsemantic Net for German	1997	Hamp & Feldweg	11
Stylometry with R: A Package for Computational Text Analysis	2016	Eder et al.	11
GernEdit: The Germanet Editing Tool	2010	Henrich & Hinrichs	10
Generous Interfaces for Digital Cultural Collections	2015	Whitelaw	10
Digitales Publizieren	2017	Kohle	10
BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding	2018	Devlin et al.	10
Humanities Approaches to Graphical Display	2011	Drucker	9
Efficient Estimation of Word Representations in Vector Space	2013	Mikolov et al.	9
MALLET: A Machine Learning for Language Toolkit	2002	McCallum	9
Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges	2018	Windhager et al.	9
Literaturwissenschaft als Hackathon: Zur Praxeologie der Digital Literary Studies und ihren Epistemischen Dingen	2018	Trilcke & Fischer	9
Digital Humanities: Eine Einführung	2017	Jannidis & Kohle & Rehbein	8
Leistung aus Vielfalt	2016	Rat für Informationsinfrastrukturen (RfII)	8
Das Drama	2001	Pfister	8

Tabelle 1: Top 20 am häufigsten zitierte Arbeiten

Um zu untersuchen wie die Bibliographien beschaffen sind, untersuchen wir deren Länge und das Alter der Referenzen, wie von Bollmann und Elliot (2020) für die ACL gezeigt, um zu untersuchen ob ältere Publikationen `vergessen‘ werden. Es folgen zwei ‚letter value‘ plots, die die Länge der Bibliographie der Abstracts und das Alter der Referenzen zeigen.

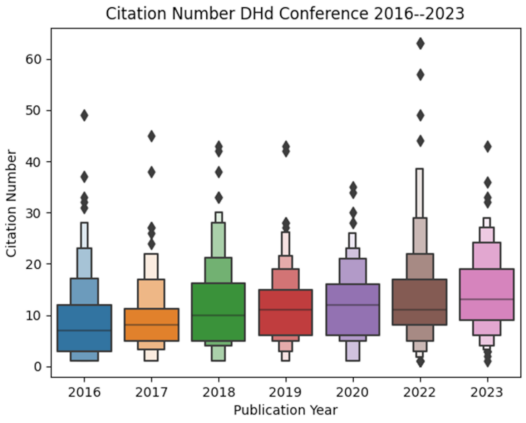


Abbildung 3: Letter Value Plot für Länge der Bibliographien

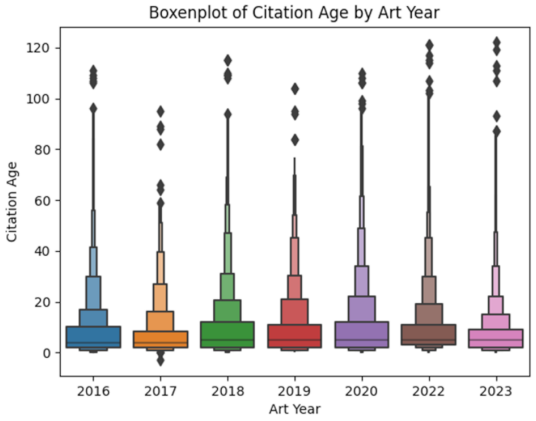


Abbildung 4: Letter Value Plot für Alter der Referenzen

Ingesamt ist zu sagen, dass die Bibliographien über die Jahre immer länger geworden sind (Median 7 bis Median 13), was für die Wissenschaftlichkeit und die Anerkennung des Fachs sicherlich fördernd ist. Zu Abbildung 4: Im Median sind Referenzen vier oder fünf Jahre alt, wenn sie zitiert werden. Es werden auch sehr alte Arbeiten zitiert, aber hier wurde erst ab 1900 berücksichtigt um etwaige Parsingerror zu minimieren. Das arithmetische Mittel scheint die letzten Jahre etwas zu fallen, was darauf hinweisen könnte, dass weniger alte Arbeiten zitiert werden.

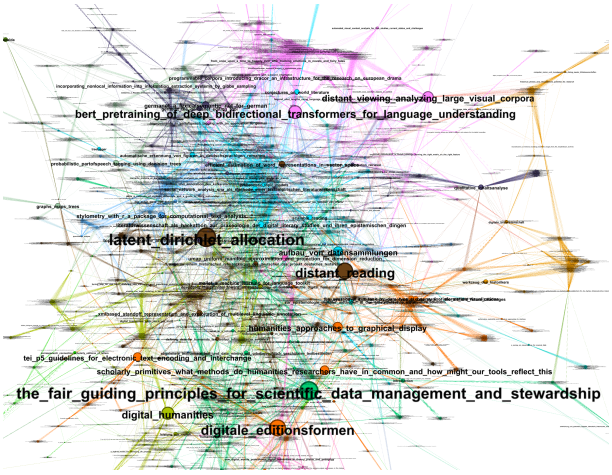


Abbildung 5: Ko-Zitationsnetzwerk

Abbildung 5 zeigt das Ko-Zitationsnetzwerk. Eine Kante wird gebildet, wenn zwei Referenzen zusammen in einer Bibliographie auftauchen, und das Gewicht der Kante bestimmt in wievielen Abstracts dies der Fall ist. Wir sehen wieder ein gewohntes Bild: Oben beginnt das Netzwerk mit Sentiment Analysis und Transformer Modellen los und bewegt sich dann über Computational Literary Studies (mit z.B. Netzwerkanalyse für Dramen) über Topic Analyse (Latent Dirichlet Allocation) und Distant Reading hin zu Forschungsdatenmanagement (FAIR) bis zur Editionswissenschaft.

Zusammenfassung

Wir zeigen auf diesem Poster einen Querschnitt der Themen (Ko-Keywords), der Ko-Autorschaft und das Zitierverhalten der DHd Community. In Zukunft wollen wir einzelne Variablen miteinander verbinden (etwa wer für welche Keywords zitiert wird).

Bibliographie

Bollmann, Marcel, and Desmond Elliott. 2020. "On forgetting to cite older papers: An analysis of the acl anthology." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pp. 7819-7827. 2020.

Henny-Krahmer, Ulrike und Patrick Sahle. 2018. “Einreichungen zur DHd 2018.” Zugriff 19. Juli 2023. https://dhd-blog.org/?p=9001.

Hoenen, Armin. 2019. “Einreichungen zur DHd 2019 - II.” Zugriff 19. Juli 2023. https://dhd-blog.org/?p=11418 .

Tang, Muh-Chyun, Yun Jen Cheng, and Kuang Hua Chen. 2017. "A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses." Scientometrics 113 (2017): 985-1008.

Wieczorek, Oliver, Markus Eckl, Madeleine Bausch, Erik Radisch, Christoph Barmeyer, and Malte Rehbein. 2021. "Better, Faster, Stronger: The Evolution of Co-authorship in International Management Research Between 1990 and 2016." *Sage open* 11, no. 4 (2021): 21582440211061561.

ReflectAI: Reflexionsbasierte künstliche Intelligenz in der Kunstgeschichte

Stalter, Julian
julian.stalter@kunstgeschichte.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-1149-1688

Springstein, Matthias
Matthias.Springstein@tib.eu
L3S Research Center, Leibniz Universität Hannover, Deutschland; TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften, Deutschland
ORCID: 0000-0002-6509-8534

Kristen, Maximilian
max.kristen@campus.lmu.de
Ludwig-Maximilians-Universität München, Deutschland

Schneider, Stefanie
stefanie.schneider@itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-4915-6949

Müller-Budack, Eric
Eric.Mueller@tib.eu
L3S Research Center, Leibniz Universität Hannover, Deutschland; TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften, Deutschland
ORCID: 0000-0002-6802-1241

Ewerth, Ralph
Ralph.Ewerth@tib.eu
L3S Research Center, Leibniz Universität Hannover, Deutschland; TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften, Deutschland
ORCID: 0000-0003-0918-6297

Kohle, Hubertus
hubertus.kohle@lmu.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID: 0000-0003-3162-1304

In der Kunstgeschichte sind Ähnlichkeitsbewertungen von Bildern von großer Bedeutung: Wölfflin analysierte Kunstwerke nach begrifflichen Gegensatzpaaren mit Doppelprojektionen, Warburg beim *Vergleichenden Sehen* nach sogenannten *Pathosformeln* (Wölfflin, 1915; Hensel, 2010). Mit Verfahren aus dem Bereich des maschinellen Sehens (*Computer Vision*) lassen sich derartige Bewertungen heute automatisieren. *State-of-the-Art* -Modelle wie *GPT* (*Generative Pre-trained Transformer*; OpenAI, 2023) oder *CLIP* (*Contrastive Language-image Pre-training*; Radford et al., 2021) können zudem aufgrund erweiterter Rechenkapazitäten effizienter generalisieren und damit auf nicht realweltliche Bilddaten angewandt werden. Eingesetzt wurden solche Ansätze bereits in kunsthistorischen Projekten, insbesondere in der Ähnlichkeits-basierten Bildsuche und -clusterung (Schneider et al., 2022; Offert und Bell, 2023).
Der Einsatz von *Künstlicher Intelligenz* (*KI*) in der kunsthistorischen Forschung eröffnet unbestreitbar neue explorative Potenziale für die Analyse von Bildähnlichkeiten. Aus methodenkritischer Perspektive ist er jedoch zu hinterfragen; insbesondere der „Black Box“-Charakter künstlicher neuronaler Netze wird diskutiert (Crawford und Paglen, 2021). Diesen Problemen soll sich das hier vorgestellte Projekt *ReflectAI* speziell für den Bereich der Kunstgeschichte annehmen. An dem Vorhaben, das im Rahmen des DFG-Schwerpunktprogramms „Das digitale Bild“ von 2022 bis 2025 gefördert wird, sind Forschende aus der Kunstgeschichte und Informatik der Universitäten München und Hannover beteiligt. Eine schematische Darstellung aller Teilmodule innerhalb des Projekts ist in Abb. 1 dargestellt.

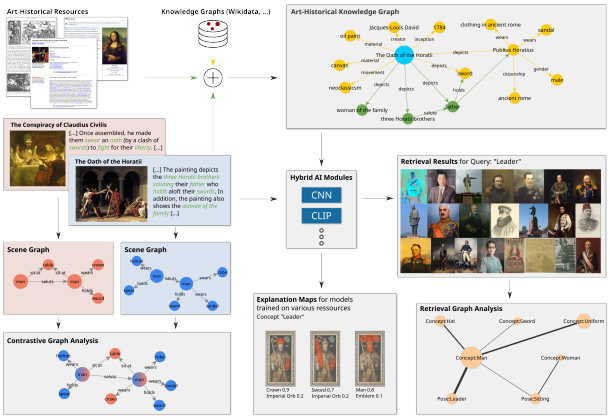


Abb. 1: Schematische Darstellung der Erstellung kunsthistorischer Wissensgraphen auf der Basis domänenspezifischer Textkorpora, des Trainings multimodaler Modelle sowie der Visualisierungentypen, die im Rahmen von *ReflectAI* entwickelt werden.

Herausforderung „Black Box“

Obwohl die Ein- und Ausgabedaten neuronaler Netze bekannt sind, bleiben die internen Verarbeitungsprozesse weitgehend undurchsichtig. Ziel unseres Projekts ist es, ein tieferes Verständnis der Prozesse zu erlangen, die zu den Suchergebnissen führen. Dabei konzentrieren wir uns auf Prozesse, die zur Identifikation von Bildähnlichkeiten, zum Clustering von Bildern und zur Klassifikation anhand spezifischer Bilddetails führen. Im Sinne einer „Explainable AI“ werden Methoden untersucht, die einen Blick in diese „Black Box“ erlauben (Guidotti et al. 2019; Offert und Bell, 2021). Unsere Absicht ist, Verfahren der automatisierten Bildanalyse sowohl aus bildwissenschaftlicher als auch aus wissenschaftshistorischer Perspektive zu analysieren und in eine für kunsthistorische Forschungsprojekte optimierte Anwendung zu überführen. Mit Motivation entwickeln wir im Projekt reflexive Werkzeuge und stellen sie bereit.

Reflexive Komponenten

Expertenwissen

Die Zusammensetzung der Trainingsdaten, auf deren Basis neuronale Netze und Modelle trainiert werden, ist den Anwendern oft nicht bekannt und im Sinne einer Methodenkritik schwer zugänglich. Unser Projekt beinhaltet die Optimierung dieser Modelle mit kunsthistorischen Texten und multimodalen Informationen (wie Text-Bild-Paaren), um domänenspezifisches Wissen zu integrieren.

Dabei wollen wir z.B. auf Visual Language Models wie BLIP-2 (Li, 2023) oder LLAVA (Liu, 2023) zurückgreifen, wobei das Expertenwissen als Trainingsmaterial für Modelle dienen kann oder auch als zusätzlicher Input des Language Models genutzt werden kann. Außerdem trainieren wir Modelle mit Textkorpora: So können beispielsweise die Ergebnisse von Modellen, in die Kunstkritiken des 19. Jahrhunderts eingespeist wurden, mit denen verglichen werden, die auf Texten des 21. Jahrhunderts basieren. Im Rahmen des Projekts wird dazu ein multimodales Datenkorpus aus Bild- und Textquellen aufgebaut. Es umfasst derzeit 60.000 Objektbeschreibungen aus Museen, Sammlungen und Auktionshäusern, 10.000 Lexikonartikel und 50.000 wissenschaftliche Publikationen.

Wissensgraphen

Das so gesammelte Expertenwissen soll nicht nur zum Training der Modelle, sondern auch zur Erstellung von *Wissensgraphen* (*Knowledge Graphs*) verwendet werden. Obwohl es Versuche gibt, kunsthistorisches Wissen aus Wissensgraphen für das Training neuronaler Netze zu nutzen, beschränken sich diese häufig auf einzelne kunsthistorische Attribute (Garcia und Vogiatzis, 2018; Schrade und Söhn, 2022). Neben den Wissensgraphen werden Szenegraphen generiert, die mit zusätzlichen Visualisierungen ein besseres Verständnis der Bildauswahl anhand der in den Suchergebnissen dargestellten Objekte und Beziehungen ermöglichen und auf Verzerrungen („Biases“) oder Ähnlichkeiten hinweisen können (Suhail, 2021). Diese Szenegraphen können Beziehungen zwischen Objekten innerhalb eines Kunstwerks oder im Kontext darstellen. In Kombination mit Wissensgraphen ist eine kontrastive Analyse möglich.

Visualisierung von Bias

Unter Bias versteht man die Verzerrung der Ergebnisse aufgrund falscher Annahmen über die Trainingsdaten oder die Modelle. Dieser Bias kann historisch bedingt sein oder durch fehlerhafte oder diskriminierende Annotationen in den Trainingsdaten entstehen (Pasquinelli und Joler, 2021). Um solche Verzerrungen sichtbar zu machen, werden den Nutzenden im Rahmen des Projekts verschiedene Werkzeuge zur Verfügung gestellt, mit denen Biases visualisiert werden können; ein Beispiel ist in Abb. 2 dargestellt.

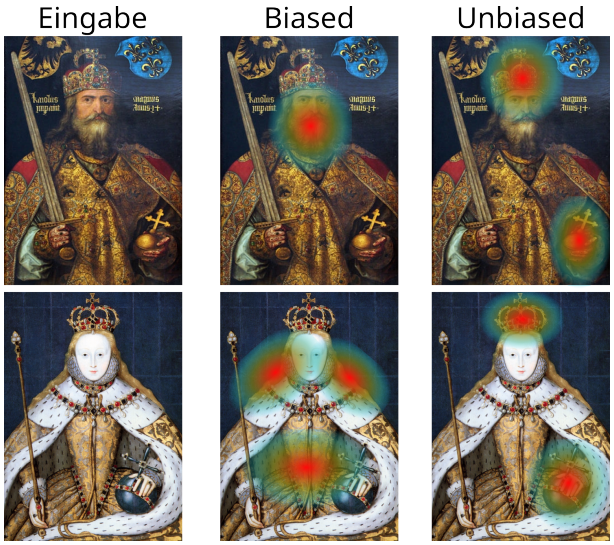


Abb. 2: Illustration der Vorhersage von Modellen zur Erkennung von herrschenden Personen. In verzerrten Modellen werden geschlechtsspezifische Merkmale wie der Bart oder das lange Haar zur Vorhersage verwendet, in unverzerrten Modellen eher Krone und Reichsapfel.

Danksagung

Das Projekt wird von der Deutschen Forschungsgemeinschaft (DFG) unter der Projektnummer 510048106 gefördert.

Bibliographie

Crawford, Kate und Trevor Paglen. 2021. „Correction to: Excavating AI. The Politics of Images in Machine Learning Training Sets.“ *AI & Society* 36.4: 1399 10.1007/s00146-021-01301-1.

Garcia, Noa und George Vogiatzis. 2018. „How to Read Paintings. Semantic Art Understanding with Multimodal Retrieval.“ In *Computer Vision – ECCV 2018 Workshops. Lecture Notes in Computer Science* 11130: 676–691 10.1007/978-3-030-11012-3_52.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti und Dino Pedreschi. 2019. „A Survey of Methods for Explaining Black Box Models.“ In *ACM Computing Surveys* 51.5: 93:1–93:42 10.1145/3236009.

Hensel, Thomas. 2010. „Aby Warburg und die‚Verschmelzende Vergleichsform‘.“ In *Vergleichendes Sehen*, hg. von Lena Bader, Martin Gaier und Falk Wolf, 468–489. München: Fink.

Li, Junnan, Dongxu Li, Silvio Savarese und Steven C. H. Hoi2023. „BLIP-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models.“ In *International Conference on Machine Learning. ICML 2023*, 19730–19742. https://

proceedings.mlr.press/v202/li23q.html (zugegriffen: 3. Dezember 2023).

Liu, Haotian, Chunyuan Li, Qingyang Wu und Yong Jae Lee. 2023. *Visual Instruction Tuning. arXiv:2304.08485*.

Offert, Fabian und Peter Bell. 2023. „imgs.ai. A Deep Visual Search Engine for Digital Art History.“ In *DH 2023. Digital Humanities 2023. Conference Abstracts*, 141–142 10.5281/zenodo.7961822.

Offert, Fabian und Peter Bell. „Perceptual Bias and Technical Metapictures. Critical Machine Vision as a Humanities Challenge.“ *AI & Society* 36.4: 1133–1144 10.1007/s00146-020-01058-z.

OpenAI. 2023. *GPT-4 Technical Report*. arXiv:2303.08774.

Pasquinelli, Matteo und Vladan Joler. 2021. „The Nooscope Manifested. AI as Instrument of Knowledge Extractivism.“ *AI & Society* 36.4: 1263–1280 10.1007/s00146-020-01097-6.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger und Ilya Sutskever. 2021. „Learning Transferable Visual Models From Natural Language Supervision.“ In *Proceedings of the 38th International Conference on Machine Learning. ICML 2021*, hg. von Marina Meila und Tong Zhang, 8748–8763, http://proceedings.mlr.press/v139/radford21a.html (zugegriffen: 19. Juli 2023).

Schneider, Stefanie, Matthias Springstein, Javad Rahnama, Hubertus Hohle, Ralph Ewerth und Eyke Hüllermeier. 2022. „iART. Eine Suchmaschine zur Unterstützung von bildorientierten Forschungsprozessen.“ In *8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum. DHd 2022*, hg. von Michaela Geierhos, 142–147 10.5281/zenodo.6304590.

Schrade, Torsten, und Linnaea Söhn. 2022. *Culture Portal 1.3. Knowledge Graph v.1.0 & Repositorien Überblick*. https://nfdi4culture.de/de/nachrichten/culture-portal-13-knowledge-graph-v10-repository-overview.html (zugegriffen: 19. Juli 2023).

Suhai, Mohammedl, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gérard G. Medioni und Leonid Sigal. 2021. „Energy-Based Learning for Scene Graph Generation.“ In *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2021*, 13936–13945 10.1109/CVPR46437.2021.01372.

Wölfflin, Heinrich. 1915. *Kunstgeschichtliche Grundbegriffe. Das Problem der Stilentwickelung in der neueren Kunst.* München: F. Bruckmann.

„Roads? Where we’re going, we don't need roads.“ Die Zukunft des Publizierens

Dinger, Patrick
patrick.dinger@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0002-2649-4737

Horstmann, Jan
jan.horstmann@uni-muenster.de
Universität Münster, Deutschland
ORCID: 0000-0001-8047-2232

Jansky, Caroline
jansky@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0002-7071-1022

Jurczyk, Thomas
thomas.jurczyk-q88@rub.de
Universität Tübingen, Deutschland
ORCID: 0000-0002-5943-2305

Steyer, Timo
t.steyer@tu-braunschweig.de
Technische Universität Braunschweig, Deutschland
ORCID: 0000-0003-0218-2269

Die DHd-AG Digitales Publizieren

Seit ihrer Gründung 2014 widmet sich die DHd Arbeitsgruppe Digitales Publizieren Diskursen rund um das Thema digitales Publizieren in den Digital Humanities. Die Arbeit der AG ist maßgeblich geprägt von der immer selbstverständlicher werdenden Nutzung digitaler Ressourcen im dynamischen Publikationswesen. Die Spannungsverhältnisse zwischen analogen und digitalen sowie geschlossenen und offenen Publikationswegen und -formen bilden Konstanten in der AG-Tätigkeit. Dabei agiert die AG möglichst eng an den Fragen und Bedarfen der Community und vermittelt und fördert die Vorteile von offenen Publikationen. Der Blick richtet sich dabei sowohl auf aktuelle Entwicklungen als auch auf die Zukunft des Publizierens. Ergebnisse und offene Fragen wurden in zahlreichen Vorträgen, Podiumsdiskussionen, Postern und digitalen Publikationen

(vgl. AG Digitales Publizieren, 2021) an die Community zurückgegeben.

Anlässlich ihres zehnjährigen Bestehens macht die AG mit einem Poster auf der DHd2024 in Form einer Zeitleiste auf ihre Aktivitäten aufmerksam und möchte über Erfolge, aktuelle Vorhaben, aber auch nicht Erreichtes informieren und ins Gespräch kommen. Einen zweiten Schwerpunkt des Posters bildet ein Blick nach vorn: Fünf Thesen zur Zukunft des (digitalen) Publizierens können interaktiv diskutiert und hinsichtlich der Wahrscheinlichkeit ihrer Realisierung eingeschätzt werden.

Fünf Thesen zur Zukunft des (digitalen) Publizierens

1. These: Die AG Digitales Publizieren streicht „Digital“ aus ihrer Benennung.

„Publikationskompetenz“ als wissenschaftliche Schlüsselkompetenz wird unabhängig von der Frage nach dem Publikationsmedium: Digitales Publizieren wird zum Normalfall, das gedruckte Buch oder Heft zur bibliophilen Kostbarkeit. Publizieren ist in Zukunft immer digital.

2. These: Es entsteht eine große Bandbreite an Publikationsformaten, die nicht nur die Lesenden, sondern auch die maschinelle Verarbeitung im Blick haben.

Obwohl schon häufiger tot gesagt, ist PDF derzeit noch das meistverbreitete Publikationsformat, das die analoge Leseerfahrung in den digitalen Raum zu verlagern versucht. Dabei wird auf digitale Features wie Maschinenlesbarkeit oder eine anpassbare Darstellung verzichtet. In Zukunft setzen sich jedoch sogenannte One-Source-Publishing-Ansätze stärker durch, bei denen eine autoritative Kerndatei automatisiert in beliebig viele andere Formate wie PDF, HTML und XML (JATS) transformiert werden kann. So werden die Veröffentlichungen durch die verbesserte Maschinenlesbarkeit selbst zu einer potenziellen Quelle für computergestützte Analysen.

3. These: Anwendungen des Machine-Learnings werden integraler Bestandteil von Prozessen des wissenschaftlichen Schreibens, Reviewings, der Redaktion und des Layouts.

Nach der niedrigschwelligen Verfügbarmachung von Large Language Models durch Anwendungen wie ChatGPT nahmen Diskussionen um die Integration „Künstlicher Intelligenz“ auch in wissenschaftliche Arbeits- und Publikationsprozesse Fahrt auf (vgl. etwa Hoffmann, 2023). In Zukunft nehmen Text- und Bildgeneratoren sowie Prompt-Engineering einen festen Platz in der wissenschaftlichen Arbeit ein. Offene KI-Anwendungen werden Community-basiert so verbessert, dass die generierten Informationen durch Verknüpfungen mit gewährleisteten Wissensressourcen nachvollziehbar sind. Außerdem setzt die wissenschaftliche Gemeinschaft Standards des Umgangs mit KI-Tools etablieren und automatisierte Generierung in Teilprozessen so ein, dass sie die wissenschaftliche Integrität von Texten und Visualisierungen nicht durch ungesichertes, nicht belegbares Wissen kontaminiert.

4. These: Der Tod der individuellen Autorschaft steht kurz bevor.

Durch die Öffnung von Publikationsprozessen bestehen die Möglichkeit und der Anspruch, Art und Umfang der Beteiligung an einem Forschungsergebnis transparent öffentlich darzustellen (vgl. D’Ignazio und Klein, 2020, 18; Heller et al., 2014). Das Prinzip „Autorschaft“ hat damit schon bald ausgedient: Stattdessen rückt die präzise Benennung des jeweils individuellen Beitrags zu einer kollaborativen Publikation in den Mittelpunkt (vgl. National Information Standards Organization, 2022).

5. These: Es werden vermehrt Datensätze statt Texte publiziert.

Mit den neuen Möglichkeiten, Forschungsdaten zu veröffentlichen und zu beschreiben, verändert sich auch die Art, wie in den Digital Humanities publiziert wird (vgl. Cremer, 2018). Die neue Publikationskultur setzt verstärkt auf offene Micro-Formate, Zeitschriftenartikel und Repositorien. Über Plattformen können Forschungsdaten und Ergebnisse zusammengefasst, Darstellungen von Forschungsprozessen als vollwertige digitale Publikation und Vertrauen stiftende Datenquelle akzeptiert werden – ein Trend, der sich bereits jetzt andeutet (vgl. Helling et al., 2022; Busch et al., 2022; NFDI 2023).

Poster interaktiv

Die Postergestaltung stellt die Interaktion zwischen den Vorstellenden und dem Publikum ins Zentrum. Die Zeitleiste, die die AG-Aktivitäten der letzten zehn Jahre darstellt, fächert sich für den Blick in die Zukunft des Publizierens in fünf bewusst pointiert dargestellte Thesen auf: Sie sollen die Besucher*innen zu einer Positionierung provozieren und zum Mitdiskutieren anregen. Über QR-Codes sind Erläuterungen der Thesen abrufbar. Außerdem können die Thesen hinsichtlich ihrer Relevanz und Wahrscheinlichkeit mithilfe eines digitalen Umfragetools bewertet werden. Diese Positionsbestimmung beschränkt sich nicht auf den Zeitraum der Posterpräsentation, sie kann weitergeführt und in anderen Kontexten aufgegriffen werden. Die Ergebnisse liefern der AG wichtige Anhaltspunkte

dafür, welche die drängendsten Fragen an die Entwicklungen des Publikationswesens sind und wie die Arbeit der AG ausgerichtet werden muss, um die Bedarfe der Community auch in Zukunft abdecken zu können.

Bibliographie

AG Digitales Publizieren. 2021. Digitales Publizieren in den Geisteswissenschaften: Begriffe, Standards, Empfehlungen. Wolfenbüttel: Zeitschrift für digitale Geisteswissenschaften / Working Papers. https://doi.org/10.17175/wp_2021_001_v2 (zugegriffen: 1. Dezember 2023).

Busch, Anna, Fabian Cremer, Harald Lordick, Dennis Mischke und Timo Steyer. 2022. „Strukturen und Impulse zur Weiterentwicklung der DHd-Abstracts.“ In DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum". https://doi.org/10.5281/zenodo.6328089 (zugegriffen: 1. Dezember 2023).

Cremer, Fabian. 2018. „Nun sag, wie hältst Du es mit dem Digitalen Publizieren, Digital Humanities?“ In Digitale Redaktion. Editorial zum digitalen Publizieren. https://editorial.hypotheses.org/113 (zugegriffen: 1. Dezember 2023).

D’Ignazio, Catherine und Lauren F. Klein. 2020. Data Feminism. Cambridge, MA: The MIT Press. DOI: https://doi.org/10.7551/mitpress/11805.001.0001 (zugegriffen: 1. Dezember 2023).

Heller, Lambert, Ronald The und Sönke Bartling. 2014. „Dynamic Publication Formats and Collaborative Authoring.“ In: Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing, hg. von Sönke Bartling und Sascha Friesike, 191–211. Cham: Springer. https://doi.org/10.1007/978-3-319-00026-8 (zugegriffen: 1. Dezember 2023).

Helling, Patrick, Anke Debbeler und Rebekka Borges. 2022. „Konferenzbeiträge strategisch publizieren: Automatisierte Workflows zur individuellen Veröffentlichung von Konferenzbeiträgen am Beispiel des Verbands Digital Humanities im deutschsprachigen Raum e.V.“ O-Bib. Das Offene Bibliotheksjournal 3: 1–17. https://doi.org/10.5282/o-bib/5835 (zugegriffen: 1. Dezember 2023).

Hoffmann, Lisa. 2023. „ChatGPT im Hochschulkontext – eine kommentierte Linksammlung.“ In: Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/de/blog/chatgpt-im-hochschulkontext-%E2%80%93-eine-kommentierte-linksammlung (zugegriffen: 1. Dezember 2023).

Nationale Forschungsdateninfrastruktur (NFDI). 2023. https://www.nfdi.de/ (zugegriffen: 1. Dezember 2023).

National Information Standards Organization. 2022. ANSI/NISO Z39.104-2022. CRediT, Contributor Roles Taxonomy. Baltimore, MD: National Information Standards Organization. https://doi.org/10.3789/ansi.niso.z39.104-2022 (zugegriffen: 1. Dezember 2023).

Search, Link, Integrate: The User-Centered Approach in Developing NFDI4Culture’s Antelope (Annotation & Terminology) Service

Rossenova, Lozana
lozana.rossenova@tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland
ORCID: 0000-0002-5190-1867

Bailly, Kolja
kolja.bailly@tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland

Blümel, Ina
ina.bluemel@tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland
ORCID: 0000-0002-3075-7640

Context

In the context of NFDI (National Research Data Infrastructure) in Germany, terminologies encompass domain-specific vocabularies, thesauri and formal ontologies (Anders et al. 2022). In order to comply with the FAIR principles for research data management (GO FAIR, n.d.), NFDI consortia need to identify and align relevant terminologies within their designated communities and achieve broad adoption and application (Kailus 2023). There is a dedicated Terminology Service Working Group within the cross-NFDI Section “Metadata” and a new service is in development, meant to meet the needs of all consortia via the Base4NFDI initiative (Anders et al. 2022).

However, commonly used terminology tools within NFDI consortia, such as the Ontology Lookup Service (Jupp et al. 2015), struggle to accommodate vocabularies and ontologies used in the arts and humanities disciplines due to their typically large size, divergence in serialization

formats, and variety of hierarchical relations within complex category trees. To address the specific needs of the culture and digital humanities (DH) research communities, the NFDI4Culture (Altenhöner et al. 2020) team based at TIB are prototyping a new service which aims to address the gaps in current service provisions. Crucial in the approach of developing the service is the close engagement with the communities represented in NFDI4Culture (Rossenova 2022), including art history, architecture, music and performing arts, as well as the agile development of the service in several stages of light-weight interactive prototypes, speculative design wireframes, and iterative releases of a public beta instance.

Requirements gathering

A dedicated requirements gathering workshop (2022) and two successive rounds of user testing (2023), provided evidence to support the claim that, when it comes to (meta)data enrichment, there is a growing interest in extending data with semantics and standard terminologies (Rossenova et al. 2022). But adoption remains slow due to lack of accessible tools for making ontological and vocabulary choices, and the large volume of manual input required, a problem already widely discussed in the field of ontology search and semantics (Butt et al. 2014; Vandenbussche et al. 2017). Researchers require powerful automation workflows that can traverse multiple terminology sources at once and can be connected directly (via APIs) to their own collection and research data management (RDM) systems. An example scenario discussed during the workshop includes a researcher cataloging a collection of stained glass images and adding structured data for the iconography (Fig.1). The researcher needs to find the correct IDs for iconography concepts from several external vocabularies and add these IDs in the collection management system (Rossenova et al. 2022). Using terminology search functions within separate web services, such as the GND or Iconclass websites (GND Explorer 2023; Iconclass Illustrated Edition, n.d.), or aggregators such as BARTOC or DANTE (BARTOC, n.d.; DANTE, n.d.), is a time-consuming and demanding process when it comes to identifying correct terms within domain-specific ontological hierarchies. Next comes the manual integration of the term IDs into the RDM system.

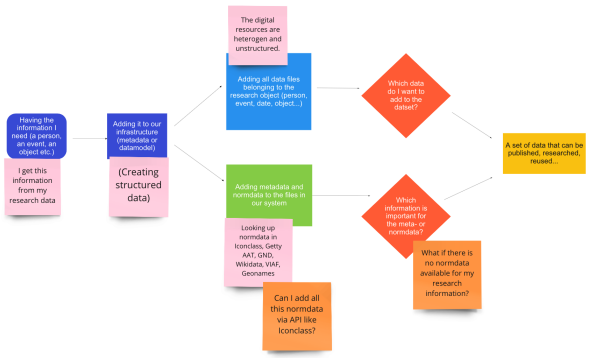


Figure 1: User journey diagramme (Rossenova et al. 2022) by Anja Gerber (Berlin-Brandenburgische Akademie der Wissenschaften Corpus Vitrearum Medii Aevi). CC-BY.

Beyond creating structured descriptive metadata, enrichment with linked entities and annotation of diverse media – from 2D images and 3D-models to AV – is a growing part of cultural and/or humanities data curation workflows, as evidenced by the user stories collected during the formation of NFDI4Culture (NFDI4Culture, n.d.). AI- and ML-assisted workflows – especially in relation to image recognition and entity linking – are commonly referenced in DH literature (Schneider et al. 2022; Vignoli et al. 2023), but are not easy to integrate in the day-to-day tasks of researchers who do not have experience in training their own models or orchestrating the required infrastructure. It is even more challenging when researchers develop niche, domain-specific vocabularies that are unlikely to be included in more general terminology search tools and pre-trained models (Elwert and Rossenova, 2023).

Service development

Antelope extends the functionalities of existing tools (Falcon 2.0 2022; iArt 2023; TIB Terminology Service, n.d.) into a common framework with high impact in meeting existing user needs and closing the gaps between: 1) using semantic concepts in describing digitized cultural objects, 2) annotating different types of media, and 3) introducing automation in assisted data curation and annotation workflows. Aiming for long-term sustainability and light-weight maintenance, the Antelope framework privileges interconnectivity over aggregation. This means the framework connects to existing, individual terminology service APIs and SPARQL endpoints in order to provide users with a uniform interface and search experience. At the same time, the extensible ‘plug-and-play’ approach keeps the framework agnostic to changes in vocabulary versions, serving the latest version available via the respective provider. The Antelope beta prototype is accessible both via a frontend web portal (Antelope, 2023) and a data service API, which can be directly integrated into third-party RDM systems (Fig.2).

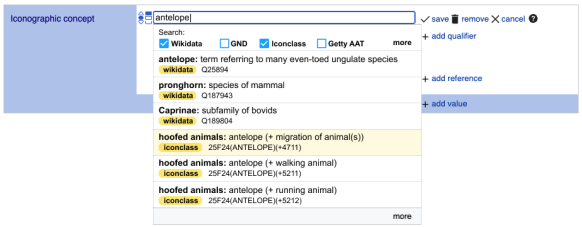


Figure 2: Suggested integration of Antelope service directly into the Wikibase user interface. CC-BY.

This poster presentation will highlight the impact of our research methodology on the choice of tool libraries, functionalities and integration strategies.

Bibliographie

Altenhöner, Reinhard, Ina Blümel, Franziska Boehm, Jens Bove, Katrin Bicher, Christian Bracht, Brand Ortrun, Lisa Dieckmann, Maria Effinger, Malte Hagener, Andrea Hammes, Lambert Heller, Angela Kailus, Hubertus Kohle, Jens Ludwig, Andreas Münzmay, Sarah Pittroff, Matthias Razum, Daniel Röwenstrunk, Harald Sack, Holger Simon, Dörte Schmidt, Torsten Schrade, Annika-Valeska Walzel, and Barbara Wiermann. "NFDI4Culture - Consortium for research data on material and immaterial cultural heritage." Research Ideas and Outcomes 6: e57036 (July 2020). DOI: 10.3897/rio.6.e57036.

Anders, Ivonne, Bailly, Kolja, Baum, Roman, Engel, Felix, Ernst, Felix, Ghiringhelli, Luca M., Kailus, Angela, et al. 2022. "Terminology Services - Working Group Charter (NFDI Section-metadata)". Zenodo. https://doi.org/10.5281/zenodo.6759325.

Antelope. 2023. Accessed December 6, 2023. https://service.tib.eu/annotation/

BARTOC. n.d. Accessed December 6, 2023. https://bartoc.org/

Butt, Anila Sahar, Haller, Armin, and Xie, Lexing. 2014. "Ontology Search: An Empirical Evaluation". In The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science, edited by Mika, Peter et al., vol 8797. Springer, Cham. https://doi.org/10.1007/978-3-319-11915-1_9

DANTE. n.d. Accessed December 6, 2023. https://bartoc.org/en/node/19999

Elwert, Frederik, and Rossenova, Lozana. 2023. "User Testing for Antelope with the DiGA project (Digitization of Gandharan Artefacts)". 18 July, 2023. See: https://github.com/DiGArtefacts/thesaurus

Falcon 2.0. 2022. Last modified January 31, 2022. https://github.com/SDM-TIB/falcon2.0

GND Explorer. 2023. Last modified November 13, 2023. https://explore.gnd.network/en/

GO FAIR. n.d. "I2: (Meta)data use vocabularies that follow the FAIR principles". Accessed July 19, 2023. https://www.go-fair.org/fair-principles/i2-metadata-use-vocabularies-follow-fair-principles/

iArt. 2023. Last modified July 4, 2023. https://github.com/TIBHannover/iart

Iconclass Illustrated Edition. n.d. Accessed December 6, 2023. https://iconclass.org/

Jupp, Simon, Burdett, Tony, Leroy, Catherine, and Parkinson, Helen. 2015. "A new Ontology Lookup Service at EMBL-EBI". Presented at the Workshop on Semantic Web Applications and Tools for Life Sciences. Online. https://www.semanticscholar.org/paper/A-new-Ontology-Lookup-Service-at-EMBL-EBI-Jupp-Burdett/b83bfbfc1f2f08e5b88af5ef65ef2a8687ac4112

Kailus, Angela. 2023. "Handreichung für ein faires management kulturwissenschaftlicher forschungsdaten". Zenodo. https://doi.org/10.5281/zenodo.7716941.

NFDI4Culture. n.d. "User stories". Accessed July 19, 2023. https://nfdi4culture.de/resources/user-stories.html

Rossenova, Lozana. 2022. "Sustainable community building with the Wikibase Stakeholder Group – presentation at FOSDEM 2022". TIB Blog, February 14, 2022. https://blogs.tib.eu/wp/tib/2022/02/14/sustainable-community-building-with-the-wikibase-stakeholder-group-presentation-at-fosdem-2022/

Rossenova, Lozana, Sohmen, Lucia, and Bailly, Kolja. 2022. "First User Research Workshop on Terminology Services in Nfdi4culture". Zenodo. https://doi.org/10.5281/zenodo.7100818.

Schneider, Stefanie, Springstein, Matthias, Rahnama, Javad, Kohle, Hubertus, Ewerth, Ralph, and Hüllermeier, Eyke. 2022. "Iart - Eine Suchmaschine Zur Unterstützung Von Bildorientierten Forschungsprozessen". Zenodo. https://doi.org/10.5281/zenodo.6328175.

TIB Terminology Service. n.d. Accessed July 19, 2023. https://terminology.tib.eu/ts

Vandenbussche, Pierre-Yves, Atemezing, Ghislain A., Poveda-Villalón, María, and Vatant, Bernard. 2017. "Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web". Semantic Web, vol. 8, 437–452, Jan. 2017. DOI: 10.3233/SW-160213.

Vignoli, Michela, Gruber, Doris, and Simon, Rainer. 2023." Revolution or Evolution? AI-Driven Image Classification of Historical Prints". In Digital Humanities 2023: Book of Abstracts, edited by Anne Baillot, Toma Tasovac, Walter Scholger, and Georg Vogeler, 184–185. Zenodo. https://doi.org/10.5281/zenodo.7961822.

Software und Zeitungen - Evaluierung einer Software zur Segmentierung von Überschriften in der NS-Zeitung "Freiheitskampf”

Henning, Tim
Henning-tim@gmx.de
Hannah-Arendt-Institut für Totalitarismusforschung, Dresden, Deutschland

Einleitung

Die Analyse großer Mengen visueller Elemente wie Überschriften, Werbeanzeigen etc. in historischen Zeitungen erweitert im Sinne des “distant reading” die Erschließung historischer Quellen für Kultur- und Geisteswissenschaftler:innen um eine wichtige Dimension. Der seit Jahren bestehende “Digital Gap” (Stumpf 2021) in der NS-Forschung erfordert zudem insbesondere Arbeiten, die sich mit der digitalen quantitativen Auswertung von historischen NS-Dokumenten beschäftigen. Eine solche Analyse erfordert Tools und Workflows, die von der Digital Humanities (DH) Community seit einigen Jahren zur Verfügung gestellt werden. Insbesondere für die Extraktion visueller Elemente in historischen Zeitungen wurden neuronale Netze, vor allem die Faster R-CNN Modelle, bereits erfolgreich getestet und eingesetzt (Funkquist 2022). Viele Tools aus der DH-Community werden jedoch bislang nicht von der eigentlichen Zielgruppe verwendet, oft aus Gründen der mangelnden Benutzerfreundlichkeit (Gibbs und Owens 2012). Die vorliegende Arbeit soll für den genannten Anwendungsfall - der einer NS-Zeitung - das geeignete Tool für Kultur- und Geisteswissenschaftler:innen identifizieren.

Das HAIT führt seit 2009 eine inhaltliche Tiefenerschließung (Hanzig u. a. 2021) der Dresdner Ausgabe der Tageszeitung “Freiheitskampf” mit dem Ziel durch, sie als Quelle zur Untersuchung der Entwicklung des Nationalsozialismus (NS) in Sachsen zu nutzen und für die politische Bildungsarbeit aufzubereiten (Hanzig, Munke, und Thoß 2022). Die Erfassung und quantitative Auswertung der Überschriften, die in gedruckten Zeitungen als “Eyecatcher”, die Lesbarkeit und Wahrnehmung beeinflussen (Holsanova, Rahm, und Holmqvist 2006; Leckner 2012; Ozretić Došen und Brkljačić 2018) soll nun die Untersuchung der Zeitung ergänzen. Exemplarisch für geisteswissenschaftlich ausgerichtete Institute kleiner Größe soll für das Vorhaben Open Source Software verwendet werden, die modifizierbar ist und auf der vorhandenen Hardware laufen kann. Für die Evaluierung geeigneter Tools sind daher neben der Vorhersagegenauigkeit von Machine Learning (ML)-Programmen auch der Umfang der Dokumentation, die Schwierigkeit der Installation, die Anpassbarkeit bzw. Modifizierbarkeit und die Laufzeit zu berücksichtigen.

Methoden

Die Auswahl der Tools fiel auf die Open Source Tools Eynollah v0.2.0, LayoutParser v.0.3.4 und OCR-D v2.49.0 (Neudecker u. a. 2019; Rezanezhad u. a. 2023; Shen u. a. 2021). Proprietäre Anwendungen wie Transkribus und ABBYY Finereader sind aufgrund eingeschränkter Möglichkeiten zur Modifikation für den Anwendungsfall nicht geeignet (Liebl und Burghardt 2020). Der LayoutParser wurde sowohl out-of-the-box als auch in trainierter Version getestet. Für das Training des Layoutparsers wurde ein Fine-Tuning der Faster R-CNN-Implementierung aus dem Detectron2 Model Zoo durchgeführt. Die Trainingsdaten umfassten den Datensatz "Beyond Words"[1] , der annotierte Informationen aus historischen Zeitungen wie Bilder, Karten und Überschriften enthält, sowie 16 eigens annotierte Seiten des Freiheitskampfes aus dem Jahr 1936. Bei Eynollah wurde eine schnellere Light-Version gegen die Standardversion getestet.

Für die Evaluierung wurden 18 Seiten des Freiheitskampfes als Ground Truth (GT) manuell annotiert. Der IoU-Schwellenwert (Intersection over Union), der die Überlappung der vorhergesagten Bounding Boxes der Tools und der Ground Truth angibt, wurde auf >0,5 gesetzt. Eigene experimentelle Tests haben gezeigt, dass eine binäre Klassifikation im Vergleich zu einer Multi-Label-Klassifikation (Titel, Untertitel etc.) von Titeln bessere Mean Average Precisions erzeugte.

Ergebnisse

Fig. 1 zeigt, dass der LayoutParser in der mit eigenen Daten trainierten Version mit einem F1-Score von 57 % die besten Ergebnisse erzielt und mit Abstand am schnellsten ist (Fig. 1 und Fig. 2).[2] Eynollah und OCR-D haben vor allem das Problem langer Laufzeiten. Außerdem gibt es für diese beiden Tools keine ausreichende Dokumentation für die Implementierung selbst trainierter Modelle. Zudem ist die Software bei OCR-D teilweise veraltet, was die Anwendung und Installation einzelner Prozessoren erschwert.

	Precision	Recall	F1-Score	Laufzeit in sek*	Dokumentation	Installation
OCR-D***	**1**	.07	.13	240	-zur Benutzung des Tools: okay (Workflow mit einzelnen Prozessoren etwas kompliziert) -für das Training: unvollständig	-nicht einfach (da einige veraltete Modelle und Prozessoren)
LayoutParser out-of-the-box	**1**	.18	.31	**5**	**-gut** -verständliche Beispiele zur Nutzung in Jupyter Notebooks	**-einfach**
LayoutParser trainiert	.85	**.42**	**.57**	**5**	**-gut** -Beispiele zum trainieren in Jupyter Notebooks	**-einfach**
Eynollah-light	.45	.33	.38	348**	**-zur Benutzung des Tools: gut** -für das Training: unvollständig	**-einfach**
Eynollah-full	.63	.19	.30	381**	-siehe Eynollah-light	-siehe Eynollah-light

* durchschnittliche Laufzeit pro Seite
** gerechnet im Google Colab
*** Die pipeline "good results for slower processors" wurde benutzt:: https://ocr-d.de/en/workflows#good-results-for-slower-processors

Fig. 1: Evaluierung

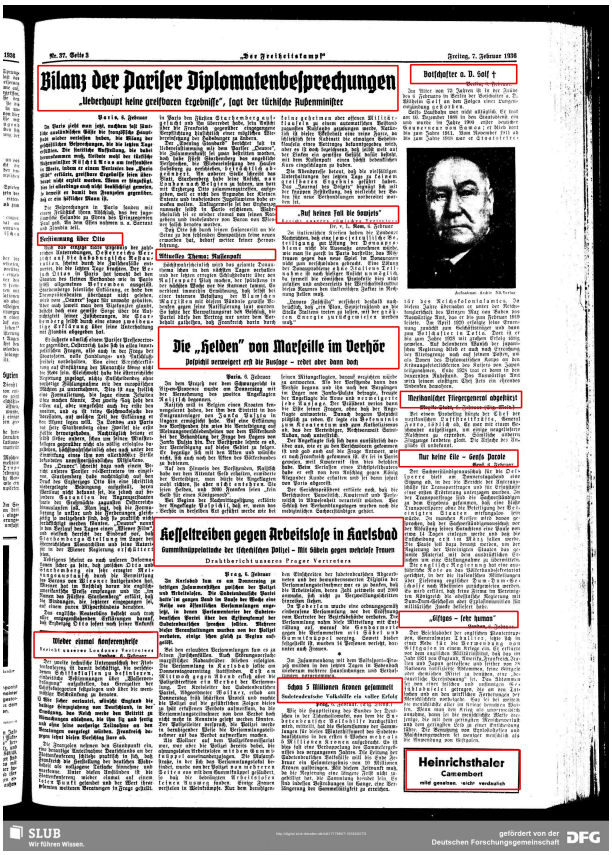


Fig. 2: Erkennung von Überschriften im "Freiheitskampf" mit dem trainierten LayoutParser

Diskussion

Eine wesentliche Einschränkung des Ansatzes ist zu beachten: Die anderen Tools wurden nicht in trainierter Form getestet, wodurch die generell niedrigen F1-Scores der out-of-the-box Anwendungen zu erklären sind. Da dies aber zum jetzigen Zeitpunkt nur mit erheblichen Änderungen im Core-Code der übrigen Tools möglich ist, ist das Kriterium der einfachen Anwendbarkeit und Benutzerfreundlichkeit nicht vollständig gegeben.[3]

Der LayoutParser bietet durch seine gute Dokumentation und Beispielanwendungen in verschiedenen Jupyter Notebooks eine einfache Benutzbarkeit.[4] Geringe Laufzeiten ermöglichen ein schnelles exploratives Testen des Tools auf die eigene Fragestellung. Da neben Überschriften unkompliziert weitere Kategorien trainiert werden können, stellt der *LayoutParser* ein ressourceneffizientes, benutzerfreundliches und anpassbares Tool zur Analyse visueller Elemente in historischen Zeitungen dar.

Fußnoten

1. Erklärung zum Datensatz: https://labs.loc.gov/work/experiments/beyond-words/
2. Die Tests wurden auf einem Laptop ohne Grafikkarte mit i5-7200U-Prozessor und 8 GB RAM durchgeführt.
3. Die Entwickler von Eynollah stellten nach Korrespondenz freundlicherweise Hilfe zur Verfügung, um die Implementierung der trainierten Modelle in den Code zu überprüfen. Die Implementierung erwies sich jedoch als zu zeitaufwendig.
4. Beispielanwendung am Freiheitskampf: https://github.com/TimOps102/Headline_extraction

Bibliographie

Funkquist, Mikaela. 2022. Layout Analysis on Modern Newspapers Using the Object Detection Model Faster R-CNN. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-320791.

Gibbs, Fred, und Trevor Owens. 2012. „Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs". Digital Humanities Quarterly 006 (2).

Hanzig, Christoph, Martin Käseberg, Thomas Lindenberger, und Michael Thoß. 2021. „Tiefenerschließung des ‚Mustergaus' Sachsen: Die Datenbank zur Dresdner Tageszeitung Der Freiheitskampf (1930–1945)". In Nationalsozialismus digital, herausgegeben von Markus Stumpf, Hans Petschar, und Oliver Rathkolb, 1. Aufl., 329–42. Göttingen: V&R unipress. https://doi.org/10.14220/9783737012768.329.

Hanzig, Christoph, Martin Munke, und Michael Thoß. 2022. „Digitising and Presenting a Nazi Newspaper". In Digitised Newspapers – A New Eldorado for Historians?, herausgegeben von Estelle Bunout, Maud Ehrmann, und Frédéric Clavert, 153–72. De Gruyter. https://doi.org/10.1515/9783110729214-008.

Holsanova, Jana, Henrik Rahm, und Kenneth Holmqvist. 2006. „Entry Points and Reading Paths on Newspaper Spreads: Comparing a Semiotic Analysis with

Eye-Tracking Measurements“. Visual Communication 5 (1): 65–93. https://doi.org/10.1177/1470357206061005.

Leckner, Sara. 2012. „Presentation Factors Affecting Reading Behaviour in Readers of Newspaper Media: An Eye-Tracking Perspective“. Visual Communication 11 (2): 163–84. https://doi.org/10.1177/1470357211434029.

Liebl, Bernhard, und Manuel Burghardt. 2020. „From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline“. In . https://www.semanticscholar.org/paper/From-Historical-Newspapers-to-Machine-Readable-The-Liebl-Burghardt/e003f89279efdb9d536cf2a012b33cdb99693324.

Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Matthias Boenig, Kay-Michael Würzner, Volker Hartmann, und Elisa Herrmann. 2019. „OCR-D: An End-to-End Open Source OCR Framework for Historical Printed Documents“. In Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, 53–58. Brussels Belgium: ACM. https://doi.org/10.1145/3322905.3322917.

Ozretić Došen, Đurđana, und Lidija Brkljačić. 2018. „Key design elements of daily newspapers: Impact on the reader’s perception and visual impression“. KOME, August. https://doi.org/10.17646/KOME.75692.93.

Rezanezhad, Vahid, Konstantin Baierer, Mike Gerber, Kai Labusch, und Clemens Neudecker. 2023. „Document Layout Analysis with Deep Learning and Heuristics“. In Proceedings of the 7th International Workshop on Historical Document Imaging and Processing, 73–78. San Jose CA USA: ACM. https://doi.org/10.1145/3604951.3605513.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, und Weining Li. 2021. „LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis“. In Document Analysis and Recognition – ICDAR 2021, herausgegeben von Josep Lladós, Daniel Lopresti, und Seiichi Uchida, 12821:131–46. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-86549-8_9.

Stumpf, Markus. 2021. „Sinnvoll, angemessen und gerecht? Digitale Wiederveröffentlichung von NS-Schrifttum durch Bibliotheken“. In Nationalsozialismus digital, herausgegeben von Markus Stumpf, Hans Petschar, und Oliver Rathkolb, 1. Aufl., 225–66. Göttingen: V&R unipress. https://doi.org/10.14220/9783737012768.225.

TEI2CEI2TEI oder Sapere Aude: habe Mut, dich deines eigenen Schemas zu bedienen!

Atzenhofer-Baumgartner, Florian
florian.atzenhofer-baumgartner@uni-graz.at
University of Graz, Österreich
ORCID: 0000-0001-8157-8629

Lamminger, Florian
florian.lamminger@uni-graz.at
University of Graz, Österreich
ORCID: 0000-0002-1310-6352

Tscherne, Niklas
niklas.tscherne@uni-graz.at
University of Graz, Österreich
ORCID: 0000-0002-7211-7393

Vogeler, Georg
georg.vogeler@uni-graz.at
University of Graz, Österreich
ORCID: 0000-0002-1726-1712

Die Charters Encoding Initiative (CEI) (Vogeler, 2005) hat wichtige Arbeit bei der Standardisierung der XML-Kodierung (historischer) Urkunden geleistet und ist in einer für Monasterium.net angepassten Version auch in über 650.000 Urkunden im praktischen Einsatz. Die Einschränkungen des CEI-Dialekts, welcher von TEI-P4 beeinflusst war, haben jedoch die Notwendigkeit einer Transformation in ein weiter verbreitetes und flexibleres Format hervorgerufen. So konzentrierte sich etwa das CORD-Projekt auf die Konvertierung von CEI in TEI-P5 (Vogeler et al., 2018). Ein angepasstes TEI-Schema wurde erstellt, Transformations-Workflows entwickelt und die Flagship-Sammlung “Illuminierte Urkunden” (Roland, 2014; Bartz & Gneiß, 2018, Roland et al., 2017) von Monasterium.net im Geisteswissenschaftlichen Asset Management System (GAMS) langzeitarchiviert.

Um CEI insgesamt interoperabler zu machen und die gesamte XML-Datenbank hinter Monasterium.net für den Einsatz in einem KI-gestützten ‘Virtual Research Environment’ (Vogeler, 2023) im Rahmen des ERC-Projektes “From Digital to Distant Diplomatics” (DiDip, 2023) vorzubereiten, generalisieren wir existierende Workflows und verbessern ihre Robustheit. Dieser Prozess erfordert eine sorgfältige Analyse der Daten, die Anpassung des bereits erstellten TEI-Schemas, die Erweiterung der Transformation inkl. ausführlicher Validierung sowie entsprechende

Dokumentation. Er beinhaltet eine konservative Einführung neuer Elemente (z. B. Authentifizierungsmittel; Winslow, 2020), Anpassung von Datentypen sowie die Anreicherung, Bündelung und Normalisierung von Daten (z. B. Sprachcodierung, Datumsangaben). Zudem streben wir ein Schema an, das zwar auf die Bedürfnisse der Domäne eingeht, aber gleichzeitig strengere Vorgaben für die Kodierung und Eingabe von Daten bietet. Es gilt weniger das Ziel, den bestehenden Bestand einmalig aufzuräumen, sondern es sollen Lehren aus der Vielfalt der vorhandenen Daten gezogen werden und mit einem Schema langfristig auf bessere und konsistentere Daten hingearbeitet werden.

Insgesamt liegt dem Transformationsprozess ein iterativer Ansatz zugrunde, der sich zunächst auf abgeschlossene und intensiv annotierte Urkundensammlungen in Monasterium.net, wie der des Zisterzienserklosters Fontenay (Stutzmann, 2022), konzentrierte. Nachdem mehrere solcher Sammlungen erfolgreich erfasst und konvertiert worden waren, richtete sich der Fokus auf den gesamten Datenbestand von Monasterium.net. Diese umfassendere Sichtweise ermöglichte es, Probleme in den Daten, der Transformation sowie dem Schema zu identifizieren und zu lösen, um einen reibungslosen und genauen Transfer zu gewährleisten.

Der Mehrwert unserer Arbeit liegt in der erhöhten Präzision der Validierung auch für zukünftige Daten. Diese Ergebnisse sind gleichzeitig reproduzierbar und projekt-überdauernd nachhaltig. Methodisch wird der Mehrwert garantiert durch das Anknüpfen an bestehende CEI2TEI-Workflows, das Erweitern der XSLT-Pipeline anhand datengetriebener Textauszeichnungsanalyse sowie das Testen und Evaluieren der Transformation im TEI-Header sowie in Markup der Texte selbst mithilfe von statistischen String-Ähnlichkeitsmessungen. Dieser Zugang ermöglicht es uns, zwischen etwaigen Kodierungsanomalien im existierenden CEI-Datenbestand, die sich im Lebenslauf der Plattform ergeben haben (Bürgermeister et al., 2018), und korrekter Abbildung des Ausgangsschemas auf das Zielschema zu unterscheiden. Bei der Transformation können wir also agnostisch gegenüber inhaltlichen Kodierungsfehlern sein, sie nicht als Fehler der Transformation identifizieren, sondern sie ggf. schon in den Ausgangsdaten berichtigen. Somit werden keine Daten ‘zurückgelassen’ und die vergangene Annotationsarbeit wird aufgewertet. Damit schlagen wir aber auch einen Workflow vor, der auf andere Datensätze übertragbar ist: Die Kombination aus formaler Analyse des Schemas und der existierenden Daten (vorhandene Kodierung) mit statistischen String-Ähnlichkeiten erlaubt es, Konsistenz zwischen Ausgangs- und Zieldaten strukturell wie inhaltlich zu prüfen - egal, ob halbautomatisch oder manuell.

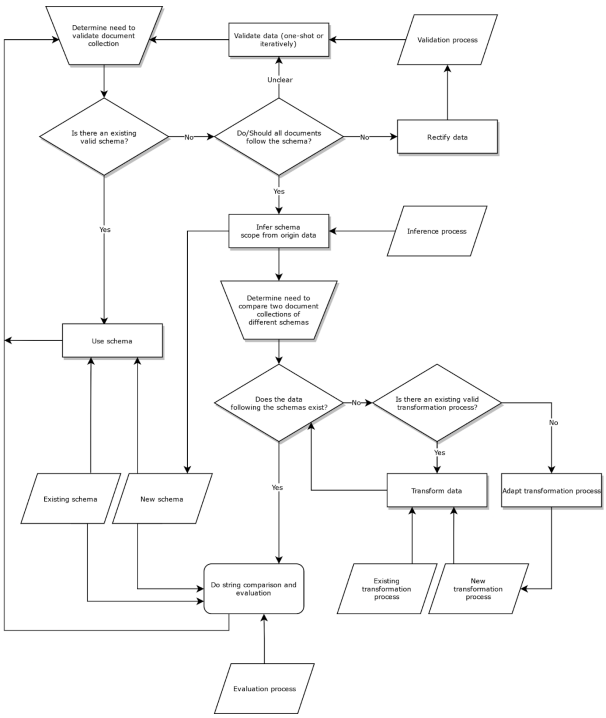


Abbildung 1: Abstrahierter Workflow einer datengetriebenen Validierung und Transformation

Schließlich stellt die verbesserte Interoperabilität und die langfristige Nutzbarkeit der so konvertierten Daten einen erheblichen Mehrwert für die wissenschaftliche Gemeinschaft dar. Interoperabler werden Schema und Daten durch die bessere Verständlichkeit und die logische Annäherung an existierende Schnittstellen, z.B. ediarum oder GAMS. Die kontinuierliche Weiterentwicklung des Schemas und der Transformation ist Voraussetzung dafür, dass zukünftige Anforderungen und Entwicklungen antizipiert werden können. Weitere Forschung kann somit auf einer soliden Datenbasis aufbauen, besser begründete Annahmen über Struktur und Inhalt der Daten machen und sich so auf neue, innovative Untersuchungen konzentrieren. Das Projekt ist ein wichtiger Schritt auf dem Weg zur dauerhaften Erhaltung und verbesserten Zugänglichkeit mittelalterlicher und frühneuzeitlicher Urkunden in digitaler Form.

Im Sinne der interdisziplinären Nachhaltigkeit erproben wir einen quelloffenen adaptierbaren Workflow und Python-Code, mithilfe dessen man XML-Datenbestände effektiv und kontrolliert beschreiben, evaluieren und assoziierte Schemata informierter verändern und transformieren kann.[1]

Fußnoten

1. Weiterführende und aktualisierte Referenzen und Beschreibungen des Workflows sind unter https://github.com/didip-eu/DHd-2024-TEI2CEI2TEI verfügbar.

Bibliographie

Bartz, Gabriele und Markus Gneiß. 2018. *Illuminierte Urkunden: Beiträge aus Diplomatik, Kunstgeschichte und Digital Humanities* . Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde. Beiheft 16. Köln: Böhlau Verlag.

Bürgermeister, Martina, Gerlinde Schneider, Stephan Makowski, Daniel Jeller, Jan Bigalke, Christian Theisen und Georg Vogeler. 2018. "'Software Aging' in Den DH: Kritik Des Reinen Forschungswillen." Köln, February 26. https://doi.org/10.5281/zenodo.4622334 .

CORD. 2019. Urkunden als offene Forschungsdaten. Ein Pilotprojekt zur Langzeitsicherung von Monasterium.net. https://gams.uni-graz.at/context:cord .

DiDip. 2023. From Digital to Distant Diplomatics. https://didip.eu .

Roland, Martin, Andreas Zajic, Georg Vogeler, Gabriele Bartz und Markus Gneiß. 2017. *Illuminierte Urkunden* . https://www.monasterium.net/mom/IlluminierteUrkunden/collection .

Roland, Martin. 2014. "Illuminierte Urkunden Im Digitalen Zeitalter: Maßregeln Und Chancen." In *Digital Diplomatics: The Computer as a Tool for the Diplomatist?* , hg. von Antonella Ambrosio, Sébastien Barret und Georg Vogeler, 245–69. Köln: Böhlau. https://doi.org/10.11588/ARTDOK.00007706 .

Stutzmann, Dominique. 2022. "Fontenay Dataset. Original Charters From Fontenay before 1213." Zenodo. https://doi.org/10.5281/zenodo.6507963 .

Vogeler, Georg, Martina Bürgermeister, Niklas Tscherne und Sean Winslow. 2018. "CEI2TEI." XSLT. https://github.com/GVogeler/CEI2TEI .

Vogeler, Georg. 2005. "Towards a Standard of Encoding Medieval Charters with XML." *Literary and Linguistic Computing* 20 (3): 269–80. https://doi.org/10.1093/llc/fqi031 .

Vogeler, Georg, Daniel Luger, Anguelos Nicolaou, Tamas Kovacs, Florian Atzenhofer-Baumgartner, Florian Lamminger, Sandy Aoun und Franziska Decker. 2023. "Building a Virtual Research Environment to Move from Digital to Distant Diplomatics (ERC Project DiDip)." https://doi.org/10.5281/zenodo.7711548 .

Winslow, Sean M. 2020. "Authenticating Features in the TEI." *Journal of the Text Encoding Initiative* , no. Issue 13 (May). https://doi.org/10.4000/jtei.3608 .

TextGrid Python Clients: Making the Repository Programmable

Hynek, Stefan
stefan.hynek@uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-3485-0746

Veentjer, Ubbo
veentjer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-9726-3135

Calvo Tello, José
calvotello@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-1129-5604

Barth, Florian
florian.barth@uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0003-3408-7311

Funk, Stefan
funk@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0003-1259-2288

Goebel, Mathias
goebel@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-1102-5284

Kurzawe, Daniel
kurzawe@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0001-5027-7313

Weimer, Lukas
weimer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

ORCID: 0000-0001-6919-3646

Current developments within TextGrid

Started in 2006, the TextGrid infrastructure was developed jointly with humanities scholars and focuses on texts encoded as TEI. Since 2015, TextGrid has been operated by DARIAH-DE and is now part of the services offered by the Association for Research Infrastructures in the Humanities and Cultural Studies (GKFI) and the NFDI consortium Text+. As a partner institution, the Göttingen State and University Library contributes TextGrid's offerings to the latter, which ensures further demand-oriented development.[1] TextGrid offers interfaces via REST and SOAP and client libraries for Java and XQuery.

In this poster, we present an accessible and streamlined interface which is built on the programming language highly utilised in Digital Humanities: The library *TextGrid Python Clients* (in short: *tgclients*).

Towards programmable interfaces

Some pioneer projects in the Digital Humanities (Digital Library, Théâtre classique, Biblioteca Italiana, etc.) dealt with the compilation of literary corpora. Researchers began to modify these corpora for their research through new features, tools (de la Rosa et al. 2022) and a new generation of corpora such as KOLIMO (Herrmann and Lauer 2017), ELTeC (Schöch et al. 2021), or DraCor (Börner and Trilcke 2023; Fischer et al. 2019).

The next step to this corpus recombination is the API-driven approach of programmable corpora in the DraCor platform (Börner and Trilcke 2023, 7), inspired by Aaron Swartz's concept of the Programmable Web, in which applications are "part of the ecology— a section of the programmable web" (Swartz 2022, 7). This idea of networked and distributed resources finds resonance in the NFDI consortia such as Text+ (Hinrichs et al. 2022), where the integration of existing resources is one of the main motivations.

Modules of *tgclients*

Concerning the before-mentioned developments, the *tgclients* library is made to provide a future-proof interface with unified access to the APIs of TextGrid. The clients are capable of solving current needs and are highly scalable to adapt future feature requests. The TextGrid repository is orientated on the FRBR model (Functional Requirements for Bibliographic Records), a very widespread model for library records and publications, and stores data within TextGrid objects which consists of content and metadata (IFLA Study Group 2009). These objects can be organised as `aggregation` (`collection`, `edition`), `work`, or individual `item`.

TG-crud: The TG-crud service is responsible for creating, retrieving, updating, and deleting TextGrid resources, i.e. TextGrid objects including TextGrid metadata (TextGrid-Konsortium 2023b). *tgclients* provides the full functionality of TG-crud with a Python interface and therefore allows creating and editing individual TextGrid objects.

TG-search: TG-search is TextGrid's central search index combining semantic and technical metadata and can be used in conjunction with access conditions. In addition to fulltext search including filters and facets, the index holds specific information required to organise objects and their relations. *tgclients* uses this interface to query for all objects of a single project and allows for applying filters such as genre, language, author gender, or period of time. Users can combine these functions with the Aggregator.

Aggregator: The TextGrid Aggregator is the export and conversion tool for data from the TextGrid repository (TextGrid-Konsortium 2023a). The aggregator collects resources in one step and converts them into relevant output formats. *tgclients* allows users to access, convert and combine the content of single objects or complete TextGrid aggregations. For example, users can convert all TEI/XML of a nested TextGrid project into plain text.

Use cases and perspectives

tgclients and its API is optimised for the usage in Jupyter Notebooks and provides an advanced user experience in documentation and autocompletion while working on a notebook. We further provide notebooks for educational purposes: Users can access several notebooks in our documentation that explain the usage of all modules with examples in TextGrid projects.[2]

The development team is curious to implement new use cases. Some ideas we have focus on the ex- and import functionality to down- and upload complete projects to TextGrid and to interact with the service for publishing data.

Fußnoten

1. https://de.dariah.eu , GKFI: https://forschungsinfrastrukturen.de , https://www.text-plus.org.
2. https://dariah-de.pages.gwdg.de/textgridrep/textgrid-python-clients/docs/getting-started.html#example-jupyter-notebooks.

Bibliographie

Börner, Ingo, and Peer Trilcke. 2023. "CLS INFRA D7.1 On Programmable Corpora", February. https://zenodo.org/record/7664964 .

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, July 9–12, 2019* . Utrecht University. https://doi.org/10.5281/zenodo.4284002 .

Herrmann, J. Berenike, and Gerhard Lauer. 2017. “Das ‘Was-Bisher-Geschah’ von KOLIMO. Ein Update Zum Korpus Der Literarischen Moderne”. In *Digitale Nachhaltigkeit* . Bern: ADHO. https://dh-abstracts.library.cmu.edu/works/10644 .

Hinrichs, Erhard, Peter Leinen, Alexander Geyken, Andreas Speer, and Regine Stein. 2022. “Text+: Language- and Text-Based Research Data Infrastructure”. Zenodo. https://doi.org/10.5281/zenodo.6452002. https://zenodo.org/record/6452002 .

IFLA Study Group on the Functional Requirements for Bibliographic Records. 2009. “Functional Requirements for Bibliographic Records”. Accessed July 1, 2023. https://repository.ifla.org/bitstream/123456789/811/2/ifla-functional-requirements-for-bibliographic-records-frbr.pdf .

Rosa, Javier de la, Aitor Díaz, Álvaro Pérez, Salvador Ros, and Elena González-Blanco. 2022. “Democratizing Poetry Corpora with Averell”. In *Responding to Asian Diversity* . Tokyo: ADHO. https://dh2022.dhii.asia/abstracts/414 .

Schöch, Christof, Tomaz Erjavec, Roxana Patras, and Diana Santos. 2021. “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives”. *Modern Languages Open* , no. 1 (December): 25. https://doi.org/10.3828/mlo.v0i0.364 .

Swartz, Aaron. 2022. *Aaron Swartz’s The Programmable Web: An Unfinished Work* . Synthesis Lectures on Data, Semantics, and Knowledge Series. Cham: Springer International Publishing AG.

TextGrid-Konsortium. 2023a. “TextGrid Aggregator”. Accessed July 1, 2023. https://textgridlab.org/doc/services/submodules/aggregator/docs/index.html

TextGrid-Konsortium. 2023b. “TG-crud”. Accessed July 1, 2023. https://textgridlab.org/doc/services/submodules/tg-crud/tgcrud-webapp/docs/index.html .

TextGrid-Konsortium. 2023c. “TG-search”. Accessed July 1, 2023. https://textgridlab.org/doc/services/submodules/tg-search/docs/index.html .

The Art of Relations

Santini, Cristian

cristian.santini@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Germany; Institute for Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology, Germany

Garay, Nele

usjpw@student.kit.edu
Karlsruhe Institute of Technology, Germany

Posthumus, Etienne

etienne.posthumus@partners.fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Germany
ORCID: 0000-0002-0006-7542

Sack, Harald

harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Germany; Institute for Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology, Germany
ORCID: 0000-0001-7069-9804

Introduction

Networks are everywhere. The brain is a network of neurons, proteins interact in cellular networks, and social networks shape our lives. Network science, i.e. “the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena” (National Research Council, 2005), provides a methodological framework to understand the complex interconnections that drive the behaviour of these systems, unveiling fundamental principles that govern their structure, dynamics, and function. Despite the fact that the concept of social networks became popular nowadays with the advent of social media, the representation of social dynamics in networks of relations provides invaluable insights into the way communities have formed and thrived throughout history. This research focuses on the realm of art-history and draws upon the seminal work “The Lives of the Artists” (1550) by Giorgio Vasari in order to examine the potential of network science to analyse this fundamental historical resource, by leveraging State-Of-The-Art (SOTA) Natural Language Processing (NLP) techniques.

Motivation

“The Lives of the Artists” is considered as the first art-historical biography ever written, consisting of a narration about lives, endeavours and social engagements of Italian and European artists living from the XIIIth to the XVIth century. This book, originally published by the Italian painter in 1550, has not only been considered for centuries as the main bibliographic source for the history of Italian and European visual arts (painting, sculpture and architecture), but it also allows to trace, through the eyes of this histo-

rian, the web of social influences and collaborations which shaped the environment in which artists worked at that time. In this context, network theory can play a major role: artists' proactiveness towards collaboration or their degree of influence in Vasari's reportage could be analysed with purely mathematical methods.

As discussed by Franco Moretti (2013), computational methods are opening scenarios for the theory of literature - and generally for the whole humanities - which were previously unimaginable. Quantitative methods from network theory can be employed to interpret cultural heritage documents not only formally but in order to grasp large-scale informative patterns, in an approach that Moretti calls "Distant Reading". In the realm of network analysis and art-history, Kienle (2017) showcases promises and challenges of graph theory for this domain. The article argues for the integration of quantitative approaches with established theories, emphasising the potential of network analysis to reveal marginalised actors and transnational histories, while cautioning against reinforcing social hierarchies if data principles are not critically examined. Lincoln (2020) discusses the deployment of computational network analysis by art historians across periods and cultures, emphasising the appeal of network rhetoric. It highlights the impact of network analysis in characterising individual positions and showing temporal changes, recognizing the challenges of integrating these methods into art-historical interpretation. Bishop (2018) instead critically addresses digital technology and the computational turn in contemporary art history. Divided into two parts, it connects issues with "digital art history" to neoliberal metrics and suggests a critical deployment of "distant reading" in analysing contemporary art, providing an alternative perspective on the role of digital methods in the field. Aim of our research is to exploit the rich bibliographic material contained in Vasari's magnum opus in order to represent the social context narrated in "The Lives of the Artists" with network theory, thus offering a novel approach to historiographers to dig deeper into historical resources and get quantitative features in order to support their study.

Methodology

This paper presents a novel approach for the extraction of social networks from historical books by exploiting Indices of Names, sections of books which list in which page persons or other named entities are referenced. This approach leverages a mixture of data mining and natural language processing techniques in order to automatically identify these references from unstructured data. In our research, we use as source a publicly available English translation of "The Lives of the Artists" (Vasari, 2018) which provides an Index of Names of the artists at the end of the book. In a first data collection phase, a scraping algorithm converts the Index of Names in the web page into a JSON file in which artists names are keys and the list of pages in which these names are mentioned are the values of these keys. Furthermore, the book sections are divided into paragraphs and stored in a CSV file, where each paragraph is associated with a page number.

In a second step, SOTA NLP techniques are used to identify references to persons in the text. The paragraphs stored in this CSV are processed with a Named Entity Recognizer (Yamada et al., 2020) in order to spot entities which the algorithm classifies as persons. Moreover, paragraphs with the same page number are put together and processed with a Coreference Resolver (Otmazgin et al., 2020) in order to cluster pronouns and names which are presumed to belong to the same entity. Results from these two algorithms are combined in order to identify all pronouns and names which refer to entities categorised as persons and group them. Finally, the Index of Names is used to resolve these clusters to specific artists by looking in the index to names present in a certain page. In order to build a social network out of textual references, artists are connected if pronouns or nouns referring to them are occuring in the same paragraph.

The use of textual co-occurrences is of big interest for literary studies, even though it is not unproblematic (Bourgeois et al., 2015). Even though the co-occurrence of two entities in a paragraph is not a sufficient condition to infer some kind of relation among the two, it can be an index of their interconnectedness inside a document. This is because two entities which tend to co-occur together in several textual sections, such as a paragraph, are likely to share some kind of relation in the document. By using Pointwise Mutual Information (PMI) (Church et al., 1990), a measure from information theory, it is possible to estimate the degree of association for a suspected relation between two artists based on their paragraph co-occurrence. PMI is a measure of association for two events A and B which acquires a positive value if two events occur together more than would be expected by their independent probability. In this way, a positive PMI for the co-occurence of two artists inside paragraphs demonstrates a relation among the two which is not given by pure chance and can be used to infer a connection, which will be therefore represented as a weighted edge between two nodes in a social network.

Results

The social network extracted from "The Lives of the Artists" is designed in order to provide a heuristic tool for art-historians and digital librarians who aim to enrich any digital edition of this art-historical document. By making use of social networks, researchers can dig deeper into the historiography of Western Art from the XIIIth to the XVIth century. With our method, 10 networks are obtained, one for each volume of the English translation of Vasari, which can be visualised online (Santini et al., 2023a). Additionally, by using Linked Open Data (LOD) vocabularies, Vasari's social network is also available as a Knowledge Graph (KG) (Santini et al., 2023b) in order to provide a valuable instrument for the Semantic Web community. By using links between artists in our KG and Wikidata entities, provided with

the *rdfs:seeAlso* property, it is possible to enrich the information present in Wikidata with results obtained from our network analysis, such as centrality scores or PMI values of a pair of artists.

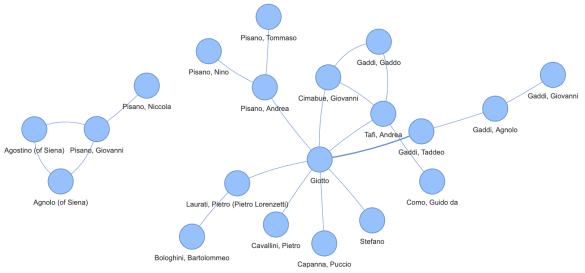


Figure 1: Portion of the extracted social network for the first volume of "The Lives of The Artists" (Vasari, 2008). Giotto's network and distinct components are clearly visible.

Furthermore, it is possible to see how high PMI values in the network are often linked to artistic collaborations, such as that of Piero di Cosimo and Cosimo Rosselli or that of Leonardo da Vinci and Verrocchio. Moreover, by looking at the centrality of certain nodes, we noticed how these are heavily influenced by the role which artists covered in Vasari's reportage of historical events. For example, the most active and influential artists, such as Michelangelo, Giotto, Da Vinci and Tiziano have high betweenness centrality, showing their central role in shaping the art scene of this period and being the masters of several other protagonists of the book.

Conclusion

In conclusion, this paper aims to provide a generalizable methodology to extract social networks from historical biographies of artists by using as case study the seminal work of Giorgio Vasari. By exploiting entity extraction techniques in tandem with an Index of Names, our approach shows how to extract a social network out of in-text co-occurrences of entities. Main contribution of our work is to make available results from our network analysis and provide interactive network visualisations for all 10 volumes of "The Lives of the Artists": this network will allow researchers in art-history to address relevant questions for the understanding of this book in a new multidisciplinary way by leveraging network theory in art-historical studies.

Finally, by converting Vasari's social network into a KG, in the end, network analysis is integrated with powerful exploratory tools and a query language, i.e. SPARQL, which allows domain-experts to better envision how computational techniques can reshape theories and methodologies not only in art-history but in the whole humanities domain. In order to favour the reproducibility of the studies carried in this research, all the data collected and processed to support the statements of this paper are made available on Zenodo (Santini, 2023a) (Santini, 2023b).

Bibliographie

Bourgeois, Nicolas, Marie Cottrell, Stéphane Lamassé, and Madalina Olteanu. 2015. "Search for Meaning Through the Study of Co-Occurrences in Texts". In Advances in Computational Intelligence, 578–91. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-319-19222-2_48.

Bishop, Claire. 2018. "Against Digital Art History". International Journal for Digital Art History, no. 3 (July). https://doi.org/10.11588/dah.2018.3.49915.

Church, Kenneth Ward, and Patrick Hanks. 1990. 'Word Association Norms, Mutual Information, and Lexicography'. Computational Linguistics 16 (1): 22–29.

Kienle, Miriam. 2017. "Between Nodes and Edges: Possibilities and Limits of Network Analysis in Art History". Artl@s Bulletin 6 (3). https://docs.lib.purdue.edu/artlas/vol6/iss3/1.

Lincoln, Matthew D. 2020. "Tangled Metaphors: Network Thinking and Network Analysis in the History of Art". In The Routledge Companion to Digital Humanities and Art History. Routledge.

Moretti, Franco. 2013. "Distant Reading". Verso Books.

National Research Council. 2005. "Network Science". The National Academies Press. https://doi.org/10.17226/11516.

Otmazgin, Shon, Arie Cattan, and Yoav Goldberg. 2022. "F-Coref: Fast, Accurate and Easy to Use Coreference Resolution". In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations, 48–56. Association for Computational Linguistics. https://aclanthology.org/2022.aacl-demo.6.

Santini, Cristian. 2023. "Data for Vasari Social Network". https://doi.org/10.5281/zenodo.8395369

Santini, Cristian. 2023. "Results for Vasari Social Network". https://doi.org/10.5281/zenodo.8395425

Santini, Cristian, Etienne Posthumus, and Harald Sack. 2023. "Vasari Social Network". https://ISE-FIZKarlsruhe.github.io/vasari_network(accessed: November 13, 2023).

Santini, Cristian, Etienne Posthumus, and Harald Sack. 2023. "Vasari Knowledge Graph" https://ise-fizkarlsruhe.github.io/vasari_network/rdfs/vasari-kg.ttl(accessed: November 13, 2023).

Vasari, Giorgio. 2008. "Lives of the Most Eminent Painters Sculptors and Architects, Vol. 01 (of 10)Cimabue to Agnolo Gaddi". Translated by Gaston du C. De Vere. https://www.gutenberg.org/ebooks/25326. (accessed: November 13, 2023).

Yamada, Ikuya, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020.

“LUKE: Deep Contextualized Entity Representations with Entity-Aware Self-Attention”. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6442–54. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.523.

Tracing the Transformation of the Labour Market through Historical Job Advertisements

Venglarova, Klara
klara.venglarova@uni-graz.at
Universität Graz, Österreich
ORCID: 0009-0007-6441-7795

Adam, Raven
raven.adam@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0001-7841-2601

Mölzer, Wiltrud
wiltrud.moelzer@uni-graz.at
Universität Graz, Österreich
ORCID: 0009-0002-9517-4531

Balasubramanian, Saranya
saranya.balasubramanian@uni-graz.at
Universität Graz, Österreich

Füllsack, Manfred
manfred.fuellsack@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-7772-4061

Kleinert, Jörn
joern.kleinert@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-1167-9245

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-1726-1712

The JobAds project (FWF P35783) aims to investigate the transformation of the labour market by extracting and analysing job advertisements from historical newspapers between 1850-1950. Using Natural Language Processing (NLP) techniques, we focus on identifying and analysing various aspects of job advertisements, including positions offered and searched for, required skills, or media strategies within a corpus of 29 digitised newspaper titles from the ANNO corpus (Österreichische Nationalbibliothek, 2021). Through this analysis, we aim to gain valuable insights into the development of the labour market and shed light on the historical dynamics of employment.

Analysing historical job advertisements presents unique challenges, starting with the need for high-quality machine-readable textual data. Errors in Optical Layout Recognition (OLR) and Optical Character Recognition (OCR), as identified in previous research (Jarlbrink and Snickars, 2017; Late and Kumpulainen, 2021; Torget, 2023; Wevers, 2023), pose problems for further text processing, including POS-tagging or dependency parsing. We plan to investigate the impact of OCR quality on the performance of NLP tasks, which was already an important topic in several studies (Rodriguez et al. 2012; Strien et al. 2020). Comparison and evaluation of the performance of different NLP tasks on raw OCR output and post-corrected text in our specific corpus will provide new valuable information about the necessity of time-consuming post-corrections for specific use cases.

Previous studies have employed text-mining approaches to examine job advertisements, e.g. for exploring ‘green jobs’ in Austria (Schober and Füllsack, 2017), or for analysing and classifying jobs in the printed medium Kleine Zeitung from the period 1950-2000 (Schober et al., 2015). However, analysing historical job advertisements introduces additional challenges due to language variations, including spelling, abbreviations, and period-specific linguistic styles. These challenges require special attention and a contextual understanding for an accurate interpretation. We must also account for the human factor, from misplaced job advertisements to content rotation either by mistake, or due to space-saving practices.

Ensuring the reliability and transparency of results is another important aspect considered in our project. In the current phase, we identified two main potential sources of bias: external factors stemming from the historical reality (newspapers were only one of several channels to realise matches of job seekers and vacancies (Wadauer et al., 2012); some newspapers contain extracts from other newspapers titles which leads to uncertainty about representativeness of a title for the area it was published in etc.), and factors related to processing digitised newspapers (different scan quality leading to non-consistent OCR results; some fonts posing greater problems for recognition than others etc.). Digitised newspapers as a research source are highly specific and a transparent interpretation of results can typically be further hindered by the incompleteness of digitised collections (Wijfjes, 2017), lack of metadata (Oberbichler and Pfanzelter, 2023) or transparency in the interfaces of digitised

newspapers (Pfanzelter et al., 2021). Therefore, it is crucial to be aware of these limitations, document them and carefully consider the research questions that can be answered based on the available data (Torget, 2023).

To overcome these challenges, we have devised a series of plans and methodologies. In the initial phase, we focused on evaluation and further improvements of layout analysis and OCR results, and employed an automated approach to filter out pages which are unrelated to our research questions. Once we obtain high-quality machine-readable data, we will analyse the structure and content of job advertisements, separately examining job descriptions and communication parts, as identified in (Misera, 2023). We plan to apply various NLP tasks, often in combination with machine learning techniques. By using BERT models pre-trained on historical texts (e.g. hmBERT (Schweter et al. 2022)), we can investigate how these can be adapted for specific use cases, such as job ads identification or post-OCR corrections, and compare them to classical dictionary- or rule-based approaches. Other NLP tasks include linguistic analysis to link job positions with verb-based descriptions of activities, parts-of-speech analysis, exploring abbreviations and interpunction usage, word embeddings, analysing word frequencies, and using topic modelling to gain insights into desired qualifications, requirements and responsibilities associated with different positions, employment trends and changing demand for specific skills and occupations. By examining patterns and trends over time, we can gain insights into the changing dynamics of employment and identify shifts in the demand for specific occupations and skills.

Bibliographie

Jarlbrink, Johan, and Pelle Snickars. 2017. ‘Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive’. Journal of Documentation 73 (6): 1228–43. https://doi.org/10.1108/JD-09-2016-0106.

Late, Elina, and Sanna Kumpulainen. 2021. ‘Interacting with Digitised Historical Newspapers: Understanding the Use of Digital Surrogates as Primary Sources’. Journal of Documentation ahead-of-print (September). https://doi.org/10.1108/JD-04-2021-0078.

Misera, Hanna Marlena. 2023. ‘Inhalt Und Aufbau von Stellenanzeigen Des 19. Und 20. Jahrhunderts. Iterative Erarbeitung Eines Skripts Zur Strukturierung von Stellenanzeigen’. Graz: University of Graz. https://unipub.uni-graz.at//obvugrhs/8546184.

Oberbichler, Sarah, and Eva Pfanzelter. 2023. ‘Tracing Discourses in Digital Newspaper Collections’. In Digitised Newspapers – A New Eldorado for Historians?, edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, 125–52. https://doi.org/10.1515/9783110729214-007.

Österreichische Nationalbibliothek. 2021. ‘ANNO Historische Zeitungen Und Zeitschriften’. 2021. https://anno.onb.ac.at/.

Pfanzelter, Eva, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais, and Stefan Hechl. 2021. ‘Digital Interfaces of Historical Newspapers: Opportunities, Restrictions and Recommendations’. Journal of Data Mining & Digital Humanities HistoInformatique (January). https://doi.org/10.46298/jdmdh.6121.

Rodriguez, Kepa J., Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of Named Entity Recognition Tools for Raw OCR Text. https://doi.org/10.13140/2.1.2850.3045.

Schweter, Stefan, Luisa März, Katharina Schmid, and Erion Çano. 2022. ‘hmBERT: Historical Multilingual Language Models for Named Entity Recognition’.

Schober, Andreas, and Manfred Füllsack. 2017. ‘Text Mining in Stellenanzeigen – Eine Methode Zur Arbeitsmarktforschung Am Beispiel Nachhaltiger Berufe in Österreich’. SWS-Rundschau, October.

Schober, Andreas, Christopher Kittel, and Manfred Füllsack. 2015. ‘Die Digitale Rationalisierung Im Spiegel von Stellenanzeigen. Automatisierte Textanalyse Zu Annahmen Des „Task-Based Approach“’. In . https://doi.org/10.15203/3122-56-7-17.

Strien, Daniel, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara Mcgillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. https://doi.org/10.5220/0009169004840496.

Torget, Andrew. 2023. ‘Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections’. In Digitised Newspapers – A New Eldorado for Historians?, edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, 47–66. https://doi.org/10.1515/9783110729214-003.

Wadauer, Sigrid, Thomas Buchner, and Alexander Mejstrik. 2012. ‘The Making of Public Labour Intermediation: Job Search, Job Placement, and the State in Europe, 1880–1940’. International Review of Social History 57 (S20): 161–89. https://doi.org/10.1017/S002085901200048X.

Wevers, Melvin. 2023. ‘Mining Historical Advertisements in Digitised Newspapers’. In Digitised Newspapers – A New Eldorado for Historians?, edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, 227–52. https://doi.org/10.1515/9783110729214-011.

Wijfjes, Huub. 2017. ‘Digital Humanities and Media History: A Challenge for Historical Newspaper Research1’. TMG Journal for Media History. https://doi.org/10.18146/2213-7653.2017.277.

Transparenz im Fokus: Die Publikationspraxis der Zeitschrift für digitale Geisteswissenschaften

Baumgarten, Marcus
baumgarten@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0003-0801-9462

Fricke-Steyer, Henrike
fricke-steyer@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0001-9677-3566

de la Iglesia, Martin
iglesia@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0002-9319-4793

Jansky, Caroline
jansky@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0002-7071-1022

Schimpf, Jonathan
schimpf@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Wiegand, Martin
wiegand@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID: 0000-0003-2151-5823

Was ist die ZfdG?[1]

Die Zeitschrift für digitale Geisteswissenschaften (ZfdG) ist ein Open-Access-Forschungsperiodikum, das sich Themen an der Schnittstelle von geisteswissenschaftlicher und digitaler Forschung widmet. Sie wurde 2015 im Rahmen der Zusammenarbeit zwischen dem Forschungsverbund Marbach Weimar Wolfenbüttel (MWW) und dem DHd-Verband ins Leben gerufen. Die ZfdG verfolgt die Entwicklungen der Digital Humanities disziplinenübergreifend und bietet Raum für kritische Debatten (vgl. Baumgarten 2021; Jansky 2023; Lippe 2018; Steyer und Wiegand 2022).

Die ZfdG nutzt innovative redaktionelle Methoden, um zu einem frei zugänglichen und transparenten Wissensaustausch beizutragen und stellt sicher, dass alle veröffentlichten Beiträge dauerhaft auffindbar sind und langzeitarchiviert werden. Das Poster präsentiert Schlüsselkennzahlen aus neun Jahren ZfdG-Publikationspraxis und bietet Ansätze für einen Vergleich mit internationalen DH-Zeitschriften sowie den Erwartungen der Redaktionsmitglieder. Zusätzlich dient es als Einstieg in eine Datenpublikation im Forschungsdatenrepositorium der Herzog August Bibliothek bzw. von MWW, um die Daten und Visualisierungsmöglichkeiten für zukünftige Forschung zugänglich zu machen.

Aspekte der ZfdG-Publikationspraxis

Einreichungs- und Publikationszahlen

Diese grundlegenden Zahlen werden aufgeschlüsselt nach Jahren und Publikationsformaten visualisiert. Außerdem wird eine Statistik zum zeitlichen Verlauf des Publikationsprozesses bezogen auf die wichtigsten redaktionellen Wegmarken veröffentlicht.

Klicks, Views und Downloads

Die Views und Downloads der pdf- und xml-Versionen der Beiträge werden bereits auf der Webseite erfasst und in Echtzeit veröffentlicht. Seit Anfang 2023 werden detailliertere Daten zum Besucherverhalten auf der Webseite erhoben und auf dem Poster visualisiert.

Disziplinen, Themen, Methoden

Alle ZfdG-Beiträge werden von Autor*innen und Redaktion mit GND-Schlagworten zur wissenschaftlichen Disziplin, verwendeten Methoden sowie behandelten Themen getaggt. Die Verschlagwortung wird für die Auswertung hinsichtlich der jeweiligen Häufigkeitsverteilung dargestellt. Dabei wird auch deutlich, welche Teildisziplinen der Digital Humanities in den Zeitschriftenbeiträgen bislang eher wenig beachtet wurden und durch gezielte Akquise erschlossen werden sollten.

Peer-Review-Verfahren

Die ZfdG bietet drei Peer-Review-Verfahren an, die jeweils unterschiedliche Grade und Aspekte von „Openness“ aufweisen (vgl. Fadeeva, 2021, 14) . Die Weiterentwicklung der Review-Verfahren hin zu mehr Transparenz sowie die Optimierung der zugehörigen redaktionellen Abläufe begleiten die ZfdG seit ihrer Konzeptionsphase.

Neben einer Aufstellung der generellen Verteilung der Review-Verfahren und den jeweils kenntlich gemachten Gesamteinschätzungen der Gutachtenden bietet sich nach etwa drei Jahren Erfahrung mit dem Open Public Peer Review (OPPR) auch die Möglichkeit einer ersten Evaluation dieses innovativen Begutachtungsverfahrens.

Zitationsnetzwerk

Eine bibliometrische Analyse der in ZfdG-Beiträgen zitierten Werke gibt Einblicke in die Struktur des DH-Feldes, wie es sich in der ZfdG widerspiegelt ((vgl. de Solla Price, 1965; Garfield, 1983). Es werden Aspekte wie die Aktualität, Internationalität und digitale Verfügbarkeit der zitierten Literatur beleuchtet.

Autorschaft

Seit 2023 fordert die ZfdG ihre Beitragenden auf, ihre eigenen Beiträge unter Verwendung der CRediT-Beiträger*innenrollen präzise zu benennen. Eine erste Auswertung der CRediT-Rollen gibt Einblick in die verschiedenen Aspekte der Autorschaft im DH-Kontext. Zudem werden geben Kennzahlen etwa zu Geschlecht und Autorschaft und dem Anteil der kollaborativ erstellten Beiträge Aufschluss über die Publikationskultur der Digital Humanities, die sich in der ZfdG abbildet.

Diskussion und Debatte

Das eingereichte Poster bietet einen Überblick über die wichtigsten Kennzahlen der mittlerweile neun Jahre andauernden Publikationspraxis der Zeitschrift für digitale Geisteswissenschaften. Zugleich soll es über Verlinkungen bzw. QR-Codes Zugriff auf die dahinterliegenden Daten und weitere Darstellungs- und Interpretationsmöglichkeiten dieser bieten, also eine nach den FAIR-Prinzipien gestaltete Datenpublikation präsentieren, die auch im Nachgang der Posterpräsentation genutzt und erweitert werden kann. Hauptaugenmerk liegt hierbei auf der interaktiven Präsentation der Daten, die laufend ergänzt werden. Einige der für das Poster erstellten Visualisierungen sollen außerdem auf der Webseite der Zeitschrift neue Zugangsmöglichkeiten über die bisherigen Such- und Filtermöglichkeiten hinaus zu schaffen.

Die rückblickende Reflexion ist dabei als Anlass für Gespräch und Debatte: Die vorliegenden Daten lassen Rückschlüsse auf blinde Flecken, Entwicklungs- und Entscheidungslinien sowie Desiderate der publizistischen Praxis zu, die wir gemeinsam mit den Tagungsteilnehmer*innen – also potenziellen Autor*innen, Leser*innen und Gutachter*innen der Zeitschrift – identifizieren, erörtern und diskutieren möchten.

Fußnoten

1. Contributor Roles: Marcus Baumgarten (Conceptualization / Formal analysis / Visualization / Writing – original draft), Martin de la Iglesia (Conceptualization / Formal analysis / Visualization / Writing – original draft), Caroline Jansky (Conceptualization / Formal analysis / Project administration / Visualization / Writing – original draft), Jonathan Schimpf (Data curation), Martin Wiegand (Conceptualization / Formal analysis / Visualization / Writing – original draft)

Bibliographie

Baumgarten, Marcus. 2021. „Zur Zukunft der Zeitschrift für digitale Geisteswissenschaften.“ In *Forschungsverbund Marbach Weimar Wolfenbüttel – Blog*. https://www.mww-forschung.de/blog/-/blogs/zur-zukunft-der-zeitschrift-fur-digitale-geisteswissenschaften (zugegriffen 3. November 2023).

Fadeeva, Yuliya. 2021. „Qualitative Sprünge in der Qualitätssicherung? Potenziale digitaler Open-Peer-Review-Formate.“ 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: Zeitschrift für digitale Geisteswissenschaften / Sonderbände. https://doi.org/10.17175/sb005_002_v2 (zugegriffen: 5. Juli 2023).

Garfield, Eugene. 1983. *Citation Indexing – Its Theory and Application in Science, Technology and Humanities*. Philadelphia: ISI Press. http://www.garfield.library.upenn.edu/ci/title.pdf (zugegriffen: 7. Juli 2023).

Jansky, Caroline. 2023. „Kollaborativ, erweiterbar, offen. Digitales Publizieren eröffnet neue Wege.“ In *HABlog*. https://www.hab.de/kollaborativ-erweiterbar-offen-2/ (zugegriffen: 3. November 2023).

Lippe, Ulrike. 2018. „Die Zeitschrift für digitale Geisteswissenschaften (ZfdG) als Best Practice für Open-Access-Zeitschriften.“ In *DHdBlog. Digital Humanities im deutschsprachigen Raum*. https://dhd-blog.org/?p=10448 (zugegriffen: 3. November 2023).

National Information Standards Organization. 2022. *ANSI/NISO Z39.104-2022. CRediT, Contributor Roles Taxonomy*. Baltimore, MD: National Information Standards Organization. https://doi.org/10.3789/ansi.niso.z39.104-2022 (zugegriffen: 4. Juli 2023).

de Solla Price, Derek J. 1965. „Networks of Scientific Papers.“ *Science* 149, 510–515. https://garfield.library.upenn.edu/papers/pricenetworks1965.pdf (zugegriffen: 7. Juli 2023).

Steyer, Timo und Martin Wiegand. 2022. „Zeitschrift für digitale Geistesiwssenschaften im Fokus. Standards

und Trends des digitalen Publizierens.“ In *Geschichte in Wissenschaft und Unterricht* 9 / 10: 554–563.

Verbrechen, Daten und Strafen. Digitales „Upcycling“ des Archivinventars zum NS-Sondergericht München

Gerstmeier, Markus
markus.gerstmeier@uni-passau.de
Universität Passau, Deutschland

Ernst, Marlene
marlene.ernst@stadt-salzburg.at
Stadtarchiv Salzburg, Historisches Archiv, Österreich

Gassner, Sebastian
sebastian.gassner@uni-passau.de
Universität Passau, Deutschland

Einleitung und Forschungsstand

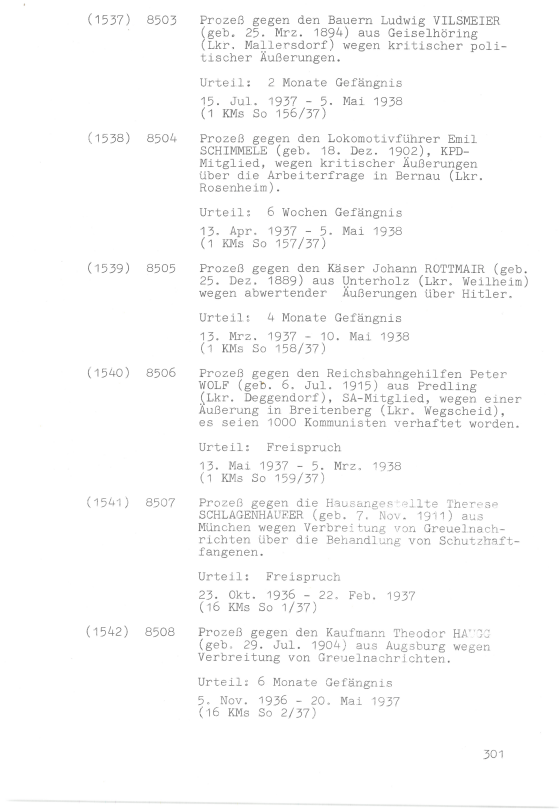
Die bisherige historische Forschung zu den Sondergerichten im nationalsozialistischen Deutschen Reich konzentrierte sich auf die an dieser Form der außerordentlichen Strafverfolgung beteiligten Personen und deren politische Bedeutung (z.B. Vurgun, 2017; Irmen, 2018; Materna, 2021; Lahusen, 2022). In unserer Untersuchung wird ein computerbasiertes „Upcycling“ (Scheltjens, 2023; Rehbein und Ernst, 2023, 508-509; Donig und Rehbein, 2022) des in den 1970er Jahren erstellten Archivinventars (BayStMUK, 1975-1977) zu den ca. 10.000 im Staatsarchiv München nahezu in ihrer Gänze erhaltenen Verfahrensakten des Sondergerichts am Oberlandesgericht (OLG) München (1933-1945) vorgenommen.

Historischer Kontext der verwendeten Quellen

Die Sondergerichte der NS-Zeit stellten insofern eine Besonderheit dar, als sie die reguläre Strafverfolgung gemäß Reichsstrafgesetzbuch von 1871 und die Ahndung von „Delikten“, die erst vom NS-Regime kriminalisiert worden sind – meist politische oder ideologische Nonkonformität –, unter einer Instanz der Strafgerichtsbarkeit vereinten (zur Inklination ordentlicher und außerordentlicher Jurisdiktion im „Dritten Reich“ vgl. Ernst et al., 2023, v.a. #p15-#p23). Nicht zuletzt deshalb sind die Gegenstände der Verfahren besonders heterogen – zumindest aus der Perspektive eines aufgeklärten, modernen Rechtsstaates, welchen die NS-Ideologie zutiefst verachtete, aber zugleich für die eigenen Zwecke zu missbrauchen wusste. Die „Sekundärquelle“ des Archivinventars stellt einen halbstrukturierten Wissensspeicher dar (vgl. Gerstmeier et al., 2022, S. 218), dessen Transformation ins Digitale Herausforderungen und Chancen für auf ‚konventionellem‘ Weg kaum oder gar nicht zu erzielende rechts- und sozialgeschichtliche Erkenntnisse birgt.

Digitalisierungsworkflow und Forschungsfragen

Die systematische und computergestützte Aufarbeitung der in Form von „Aktenregesten“ (vgl. Abb. 1) vorerschlossenen Sondergerichtsverfahren reicht von der Objektdigitalisierung bis zur quantitativen, statistischen Datenexploration des Korpus‘. Leitfragen hierbei sind etwa: War die Verfolgung politisch Andersdenkender bzw. rassisch-ideologisch verfolgter Personen wie ostmitteleuropäischen Zwangsarbeiter/innen quantitativ und qualitativ besonders drakonisch? Zumal der Sprengel des OLG München damals den kompletten südbayerischen Kulturraum umfasste, kann auch der Frage nach einem Ideologiekonflikt zwischen der überwiegend katholischen Bevölkerung und der völkisch-rassistischen nationalsozialistischen Weltanschauung nachgegangen werden. Eine quantitative Gesamtanalyse des Datenbestandes kann außerdem Aufschluss geben über berufliche und (partei-)politische Hintergründe der Angeklagten (vgl. Abb. 1) oder über die diachrone „Nervosität“ der Repression während der NS-Herrschaft. Letztere zeigt sich u.a. daran, dass keines der insgesamt 254 Todesurteile, die das Sondergericht München verantwortete, vor 1939 gefällt wurde.



(1537) 8503 Prozeß gegen den Bauern Ludwig VILSMEIER (geb. 25. Mrz. 1894) aus Geiselhöring (Lkr. Mallersdorf) wegen kritischer politischer Äußerungen.

Urteil: 2 Monate Gefängnis

15. Jul. 1937 - 5. Mai 1938
(1 KMs So 156/37)

(1538) 8504 Prozeß gegen den Lokomotivführer Emil SCHIMMELE (geb. 18. Dez. 1902), KPD-Mitglied, wegen kritischer Äußerungen über die Arbeiterfrage in Bernau (Lkr. Rosenheim).

Urteil: 6 Wochen Gefängnis

13. Apr. 1937 - 5. Mai 1938
(1 KMs So 157/37)

(1539) 8505 Prozeß gegen den Käser Johann ROTTMAIR (geb. 25. Dez. 1889) aus Unterholz (Lkr. Weilheim) wegen abwertender Äußerungen über Hitler.

Urteil: 4 Monate Gefängnis

13. Mrz. 1937 - 10. Mai 1938
(1 KMs So 158/37)

(1540) 8506 Prozeß gegen den Reichsbahngehilfen Peter WOLF (geb. 6. Jul. 1915) aus Predling (Lkr. Deggendorf), SA-Mitglied, wegen einer Äußerung in Breitenberg (Lkr. Wegscheid), es seien 1000 Kommunisten verhaftet worden.

Urteil: Freispruch

13. Mai 1937 - 5. Mrz. 1938
(1 KMs So 159/37)

(1541) 8507 Prozeß gegen die Hausangestellte Therese SCHLAGENHAUFER (geb. 7. Nov. 1911) aus München wegen Verbreitung von Greuelnachrichten über die Behandlung von Schutzhaftfangenen.

Urteil: Freispruch

23. Okt. 1936 - 22. Feb. 1937
(16 KMs So 1/37)

(1542) 8508 Prozeß gegen den Kaufmann Theodor HAUG (geb. 29. Jul. 1904) aus Augsburg wegen Verbreitung von Greuelnachrichten.

Urteil: 6 Monate Gefängnis

5. Nov. 1936 - 20. Mai 1937
(16 KMs So 2/37)

301

Abb. 1: Regestenartiger Aufbau der „Verzeichniseinheiten“ im Archivinventar zu den Verfahren am Sondergericht München 1933-1945, hier BayStMUK 1975-1977, Teil 1: 301.

Die auf das maschinengeschriebene Archivinventar aus den 1970er Jahren angewandte automatische Texterkennung erwies sich als überwiegend valide, auch wenn sich die teilweise mangelnde Druckqualität der Inventarbände aus den 1970er-Jahre limitierend auswirkt; von insgesamt 9.955 Fällen können derzeit immerhin 8.531 (86 %) bearbeitet werden. Auch ist eine Informationsextraktion aus dem maschinenlesbaren Text durch dessen formalisierte Struktur möglich. Die in den 1970er Jahren durchgeführte Kompilationsarbeit war akribisch und bietet die Möglichkeit, den Inhalt auf einer systematischen Ebene zu erschließen. Gleichwohl bestehen auch einige Unstimmigkeiten und historische Desiderate: So wurden erst für die Verfahren seit 1939 auch die jeweiligen Rechtsgrundlagen miterfasst. Metadaten zu den beteiligten Staatsanwälten und Richtern waren gar nicht Teil des dem Archivinventar zugrunde liegenden Datenmodells, was symptomatisch für die Art der staatlich finanzierten Aufarbeitung des NS-Unrechts in den 1970er-Jahren ist (vgl. Materna, 2021). Diese Unstimmigkeiten stellen eine Herausforderung für das digitale „Upcycling“ im Sinne von automatisierten Prozessen dar.

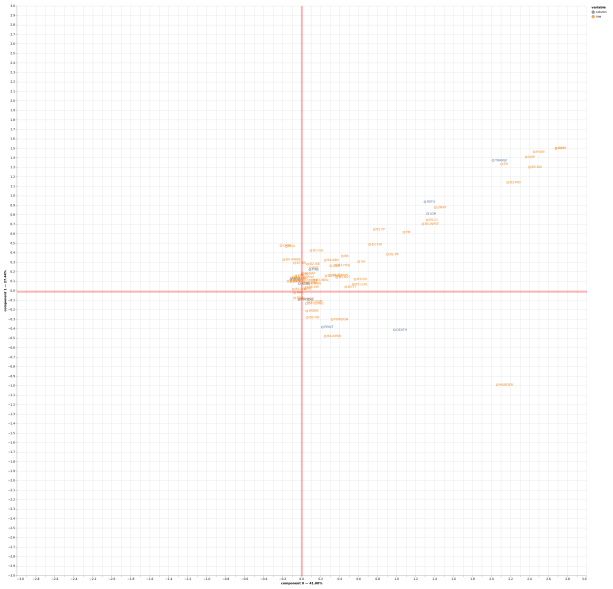


Abb. 2: Korrespondenzanalyse zur Verfahrenspraxis am Sondergericht München (1933-1945): soziokulturelle Hintergründe der Angeklagten im Vergleich zu den Verfahrensausgängen. Die Visualisierung zeigt z.B. (hier im I. Quadranten), dass ein enger Zusammenhang zwischen dem Personenmerkmal einer vor 1933 bestehenden Mitgliedschaft in der Deutschnationalen Volkspartei (DNVP) mit dem Verfahrensausgang „Ehrverlust“ (hier LOH, Lost of Honor) vorlag.

Nach der Objektdigitalisierung und Informationsextraktion haben wir verschiedene Kategorisierungsmethoden, zunächst im Kontext der Anklagegründe, verglichen, um ein besseres Verständnis für das Datenmaterial selbst sowie für den Vorverarbeitungsprozess aus den 1970er Jahren zu bekommen – von einem deduktiven Ansatz entlang der Rechtsgrundlagen über eine pragmatisch-explorative, auf Schlüsselwörtern basierende Herangehensweise bis hin zu „induktiven“, automatisierten Prozessen mithilfe von doc2vec, welche die Ergebnisse clustern (Ernst et al., 2023). Eine schlüssige Kategorisierung – nicht nur der Anklagen, sondern auch anderer relevanter Informationen wie z. B. der verschiedenen Verfahrensausgänge – schafft zudem die Grundlage für weitergehende Datenanalysen, z. B. eine Korrespondenzanalyse des Gesamtbestandes, wie sie beispielhaft in der Abb. 2 zu sehen ist.

Bibliographie

BayStMUK. 1975-1977. “Widerstand und Verfolgung in Bayern 1933–1945. Hilfsmittel“. Im Auftrag des Bayerischen Staatsministeriums für Unterricht und Kultus (BayStMUK) herausgegeben von der Generaldirektion der Staatlichen Archive Bayerns. Archivinventare, Bd. 3: Sondergericht München, 7 Teile, München / Regensburg: Eigendruck der Staatlichen Archive Bayerns.

Blasius, Jörg. 1987. “Korrespondenzanalyse: ein multivariates Verfahren zur Analyse qualitativer Daten“. Historical Social Research 12 (2/3): 172-189.

Donig, Simon und Rehbein, Malte. 2022. “Für eine ‚gemeinsame digitale Zukunft‘. Eine kritische

Verortung der Digital History“. Geschichte in Wissenschaft und Unterricht 72 (9/10): 527-545.

Ernst, Marlene, Gassner, Sebastian, Gerstmeier, Markus, Rehbein, Malte. 2023. “Categorising Legal Records – Deductive, Pragmatic, and Computational Strategies”. Digital Humanities Quarterly 17.3. http://www.digitalhumanities.org/dhq/vol/17/3/000708/000708.html (zugegriffen 19. Juli 2023)

Gerstmeier, Markus, Donig, Simon, Gassner, Sebastian, Rehbein, Malte. 2022. “Die Archivinventare zum Sondergericht München (1933-1945) digital. Quellenwert – Verdatung – Erkenntnisperspektiven“. Archivalische Zeitschrift 99 (1): 215-251.

Irmen, Helmut. 2018. “Das Sondergericht Aachen 1941-1945 (= Juristische Zeitgeschichte 2/21)“. Berlin/ Boston: W. de Gruyter Verlag.

Lahusen, Benjamin. 2022. “Der Dienstbetrieb ist nicht gestört. Die Deutschen und ihre Justiz 1943-1948.“ München: C. H. Beck.

Materna, Markus. 2021. „Richter der eigenen Sache. Die ‚Selbstexkulpation‘ der Justiz nach 1945, dargestellt am Beispiel der Todesurteile bayerischer Sondergerichte“. Baden-Baden: Nomos Verlag.

Rehbein, Malte und Ernst, Marlene. 2023. “Erschließung handschriftlicher Dokumente zwischen Fachwissen, Citizen Science und KI“. Bibliothek – Forschung und Praxis 47 (3): 503-513.

Scheltjens, Walter. 2023. “Upcycling historical data collections. A paradigm for digital history?” Journal of Documentation. DOI:10.1108/JD-12-2022-0271 (zugegriffen 4. Dezember 2023)

Vurgun, Oskar. 2017. “Die Staatsanwaltschaft beim Sondergericht Aachen“. Berlin.

Visuelle Textanalyse vom Distant zum Close Reading mit THeMSE

Lehmann, Marina
marina.lehmann@uni-mainz.de
Johannes Gutenberg-Universität Mainz, Deutschland
ORCID: 0000-0002-6818-6169

John, Markus
markus.john@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Kuczera, Andreas
andreas.kuczera@mni.thm.de
Technische Hochschule Mittelhessen, Deutschland
ORCID: 0000-0003-1020-507X

Einleitung

Textvergleiche begegnen uns überall: Von Rechtschreibprüfung und Autokorrekturvorschlägen über Plagiatserkennungssoftware bis hin zu Versionierungssystemen wie Git – alle vergleichen in der ein oder anderen Form Texte. Auch im Bereich Digital Humanities beschäftigen sich Forschende mit der Frage, wie ähnlich sich Texte sind. Ein Beispiel dafür ist das DFG-Projekt *Das Buch der Briefe der Hildegard von Bingen. Genese – Struktur – Komposition*[1], welches die in verschiedenen Handschriften überlieferten Briefe Hildegard von Bingens digital ediert.

Zwei Forschungsfragen sind im Kontext des Hildegard-Projekts von besonderem Interesse:

1. Welche handschriftenübergreifenden Ähnlichkeiten bestehen zwischen den überlieferten Manuskripten? Ziel ist es, Briefe oder Textbausteine zu identifizieren, die in mehreren Fassungen in gleichem oder ähnlichem Wortlaut auftreten.
2. Welche Ähnlichkeiten bestehen zwischen den Briefen innerhalb einer Handschrift? Diese Frage geht einher mit der Hypothese, dass die Briefsammlungen nicht willkürlich zusammengestellt wurden, sondern kompositorischen Prinzipien folgen. Eine Aufgabe der Forschenden besteht darin, Themen und Leitgedanken zu identifizieren, die innerhalb einer Handschrift wiederkehren, um dadurch Erkenntnisse über die Komposition als Ganzes zu gewinnen.

Um diese Fragen auf möglichst systematische Weise zu beantworten, wurde ein lauffähiger Prototyp für ein visuelles Analysewerkzeug entwickelt (Demo-Version: https://mlehma03.pages.gitlab.rlp.net/themse/). Das Werkzeug *THeMSE (Text-Hermeneutic Multilevel Similarity Exploration)* orientiert sich dabei an dem Grundprinzip, visuelle Abstraktionen für ein Distant Reading (Moretti 2000, 54–68) der Texte mit der Möglichkeit zum Close Reading (Burdorf et al. 2007, 126) zu verbinden. Solche visuellen Abstraktionen, wie z.B. Wordclouds oder Netzwerke, können komplexe Zusammenhänge sichtbar machen, die verborgen bleiben, wenn die Analyse direkt im Detail ansetzt. Um die Zusammenhänge sowie die im Distant Reading entstandenen Ideen und Hypothesen detaillierter analysieren zu können, ist es jedoch notwendig, auch die zugehörigen Textabschnitte zu studieren. Daher werden die visuellen Abstraktionen mit dem zugrunde liegenden Textgegenstand verknüpft.

Forschungsstand

Im Bereich der visuellen Textanalysewerkzeuge gibt es verschiedene visuelle Abstraktionen, die unterschiedliche Abstufungen von Nähe bzw. Distanz zum Text umsetzen. Visualisierungen wie *CollateX* (Haentjens Dekker et al.

2015, 452–470) oder *TRAViz (Text Re-use Alignment Visualization)* (Jänicke et al. 2015a, i83–i99) ermöglichen ein Close Reading, indem sie Dokumente als *Text Variant Graph* (Schmidt und Colomb 2009, 497–514) visualisieren, um Abweichungen zwischen den Textvarianten verschiedener Editionen erkennbar zu machen. Aber auch diff-ähnliche Werkzeuge wie *eComparatio* (Bräckel et al. 2019, 221–238), das speziell auf lateinische und griechische Texte zugeschnitten ist, oder *Tesserae* (Coffee et al. 2013, 221–228), das intertextuelle Bezüge in lateinischen Texten untersucht, fallen in diese Kategorie.

Auf der anderen Seite des Spektrums stehen Überblicksvisualisierungen, die für ein Distant Reading entwickelt wurden. Ein Beispiel dafür sind die *Fingerprint Matrizen* von Keim und Oelke (2007, 115–122). Sie bieten pro Text eine pixelbasierte Visualisierung bezogen auf ein bestimmtes Merkmal der Texte (z.B. Satzlänge, Lesbarkeit etc.). Ein weiterer Darstellungsansatz ist der *Text Re-use Grid* (Jänicke et al. 2015b, 153–171), eine Matrixvisualisierung mit Farbcodierung, bei der jede Zelle den Text-Reuse zwischen zwei Dokumenten im Korpus repräsentiert. Je intensiver die Farbe, desto mehr Text wurde wiederverwendet.

Darüber hinaus gibt es integrative Ansätze, welche auf verschiedenen Ebenen sowohl Close Reading als auch Distant Reading ermöglichen. *TextComparator*, eine Visualisierung, die im Projekt *ePoetics* zur vergleichenden Analyse verschiedener deutschsprachiger Poetiken entwickelt wurde, bietet einen Überblicksmodus mit zwei Bändern, die auf abstrakte Weise den Textverlauf mit farblich hervorgehobenen Suchwörtern visualisieren, sowie eine Detailansicht des Textes ebenfalls mit farblichen Markierungen (John 2020, 22f.). Innerhalb des gleichen Projekts ist auch der *Varifocal Reader* entstanden, der eine ganze Reihe von abstrakten Übersichtsansichten parallel zur Textansicht bietet (Koch et al. 2014, 1727).

THeMSE stellt ebenfalls das Zusammenspiel von Close Reading und Distant Reading in den Mittelpunkt. Das Prinzip, die Visualisierung in verschiedene Abstraktionsebenen einzuteilen – sowohl im *TextComparator* als auch im *Varifocal Reader* grundlegend – wurde für *THeMSE* übernommen. *THeMSE* besteht aus drei Ebenen: einer Überblicks-, einer Explorations- und einer Detailebene. Die Überblicksebene (Abb. 1) mit Heatmap ermöglicht es, handschriftenübergreifende Textähnlichkeiten zu identifizieren. Über die Explorationsebene (Abb. 2) können Nutzende flexibel anpassbare Fingerprints für Briefe erstellen lassen, in denen bestimmte Begriffe oder Begriffskombinationen auftreten. Jeder Fingerprint ist über einen Klick direkt mit der Textansicht verknüpft (Detailansicht, Abb. 3). Die drei Ebenen werden im Folgenden anhand konkreter Anwendungsfälle vorgestellt.

Überblick: Handschriftenübergreifende Textähnlichkeit

> [...] [T]he seemingly simple question “How similar are two texts?” cannot be answered independently from asking what properties make them similar. (Bär et al. 2011, 515)

Welche Eigenschaften eines Textes bei der Ermittlung von Textähnlichkeit berücksichtigt werden, hängt stark vom jeweiligen Textvergleichsverfahren ab. Daher bietet THeMSE mehrere Verfahren zur Auswahl. Je nach Ansatz wird Textähnlichkeit zeichenbasiert als Editierdistanz (Levenshtein), worthäufigkeitsbasiert (als Bag-of-Words-Modell gewichtet per TF-IDF) oder kontextbasiert (doc2vec) definiert.

Bei der Verwendung der Levenshtein-Distanz (Levenshtein 1966, 707–710) wird Textähnlichkeit anhand der übereinstimmenden Zeichen zwischen zwei Texten gemessen. Zwei Texte sind ähnlich, wenn ihre Editierdistanz möglichst gering ist, d.h. wenn möglichst wenig Zeichen ergänzt, gelöscht oder ausgetauscht werden müssen.

Alternativ kann Textähnlichkeit anhand der für die Texte charakteristischen Wörter gemessen werden. Mit dem Bag-of-Words-Modell werden Texte in Vektoren verwandelt, die auf Worthäufigkeiten basieren. Per TF-IDF (Term Frequency – Inverse Document Frequency) (Salton und Buckley 1988, 513–523) werden die Häufigkeiten gewichtet, sodass Wörter, die in einem Dokument häufig vorkommen, im übrigen Korpus jedoch weniger häufig, stärker gewichtet werden. Mithilfe der Kosinusdistanz wird der Abstand zwischen den Dokumentvektoren im Raum und damit indirekt die Ähnlichkeit zwischen den Dokumenten gemessen. Zwei Texte sind somit ähnlich, wenn sie ähnliche charakteristische Wörter enthalten.

Auch bei doc2vec (Le und Mikolov 2014, 1–9) wird Ähnlichkeit als Kosinusdistanz zwischen Dokumentvektoren modelliert. Der Unterschied liegt darin, dass die Vektoren mithilfe von Machine-Learning auf Basis der Wortkontexte berechnet werden. Die zugrunde liegende Annahme besteht darin, dass Wörter, die in ähnlichen Kontexten auftreten, ähnliche Bedeutungen haben (Distributional Hypothesis). Bei doc2vec sind zwei Texte somit ähnlich, wenn die Wörter in ähnlichen Kontexten auftreten.

Über diese Verfahren kann pro Textpaar ein Ähnlichkeitswert ermittelt werden, dessen Skala von 0 (keine Ähnlichkeit) bis 1 (maximale Ähnlichkeit) reicht. Die Ähnlichkeitswerte bilden den Ausgangspunkt für die Überblicksebene. Sie besteht aus einer Heatmap, die von den Text Re-use Grids (Jänicke et al. 2015b, 4) inspiriert ist, diese jedoch minimalistischer umsetzt. Während bei Jänicke et al. zwischen drei verschiedenen Arten von Text Reuse differenziert wird und die Zellen des Grids in ihrer Größe angepasst werden, beschränkt sich *THeMSE* darauf, die errechneten Ähnlichkeitswerte farblich zu codieren. So lässt sich schnell anhand der Farbintensität erkennen, welche Briefe sich ähneln.

Beim ersten Aufruf erstellt *THeMSE* eine Heatmap mit den Standardeinstellungen (Manuskriptauswahl: R-Wr, Textform: normalisiert, Textvergleichsmethode: Levenshtein). Über das Menü können die drei Parameter angepasst werden. Jede Zelle der Heatmap steht für ein Briefpaar mit jeweils einem Brief aus der Handschrift R sowie Wr. Die Farbintensität der Zellen spiegelt die Höhe des errechneten Ähnlichkeitswerts wider. Über dieses Farbmapping können die Nutzenden schnell erkennen, welche Briefe besonders ähnlich sind. Für R und Wr zeigt sich bei Nutzung des Levenshtein-Verfahrens beispielsweise anhand der dunkelroten Diagonalen, dass es für fast jeden Brief in R einen Brief in Wr gibt, der ihm besonders ähnelt (Abb. 1).

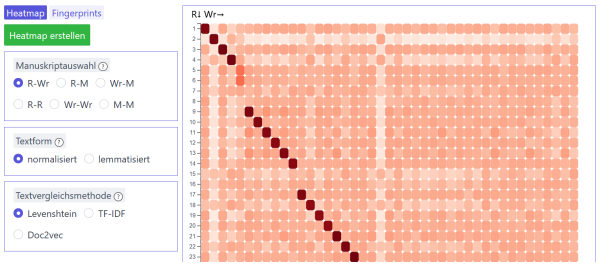


Abb. 1 – Überblicksebene mit Heatmap

Exploration: Handschriftenineterne Ähnlichkeit

Die zweite Ebene (Explorationsebene) implementiert eine Begriffssuche in Kombination mit einer Fingerprint-Visualisierung, durch die das Vorkommen ausgewählter Begriffe und Begriffskombinationen innerhalb einer Handschrift untersucht werden kann. In den Fingerprints von Oelke et al. steht jedes Pixel für eine Texteinheit, z.B. für einen Satz. Über die Farbe des Pixels können Textmerkmale codiert werden, z.B. Satzlänge oder Lesbarkeit. *THeMSE* sieht pro Wort ein Pixel vor, wobei die farbliche Kodierung das Vorkommen der Suchwörter widerspiegelt (Abb. 2).

Wechselt man zur Fingerprint-Ansicht, erscheint zunächst ein Suchmenü. Neben der Auswahl des Manuskripts für die Suche, können hier Sortier- und Filtereinstellungen vorgenommen werden. Beispielsweise kann festgelegt werden, dass bei mehreren Suchbegriffen nur nach deren Kombination gesucht wird. Zudem kann bestimmt werden, ob die Fingerprints nach Anzahl der Treffer sortiert werden oder nach Anzahl der Wörter zwischen den Treffern. *THeMSE* bietet Suchmöglichkeiten für drei Szenarien:

1) Wissen Forschende bereits, welche Begriffe interessant sind, können sie diese ins Feld der Begriffssuche eingeben.

2) Möchten Forschende stattdessen explorativ vorgehen, können sie sich Begriffsvorschläge basierend auf einem zuvor ausgewählten Brief generieren lassen. Die Vorschläge entsprechen den Top 20 der charakteristischen Begriffe dieses Briefs, d.h. den 20 Begriffen mit den höchsten TF-IDF-Werten.

3) Vielversprechende Begriffe können in einer Merkliste gespeichert werden. Gespeicherte Einträge können ebenfalls für die Suche genutzt werden. Die Suche wird jeweils über den Button “Fingerprints erstellen” gestartet.

Angenommen, die Forschenden möchten in der Handschrift Wr suchen, haben jedoch noch keine konkrete Themenvorstellung und betrachten daher die Begriffsvorschläge basierend auf R26. Dabei entdecken sie, dass sich aus den Begriffen “pater”, “filius” und “spiritus” das Thema “Trinität” ergeben könnte. Sie wählen die drei Begriffe aus und starten die Suche (Kombinationssuche). Sie vermuten, dass Briefe, in denen die Begriffe in geringer Distanz zueinander auftreten, am wahrscheinlichsten “Trinität” thematisieren. Daher wählen sie die Sortierung nach “Anzahl der Wörter zwischen Treffern” aus. Nun wird für jeden Brief, der diese drei Begriffe enthält, eine Fingerprint Matrix angezeigt (Abb. 2). Die grauen Kästchen der Fingerprint Matrix stehen für ein Wort im Text. Stimmt das Wort mit einem der Suchbegriffe überein, wird es farblich hervorgehoben.

Die Explorationsebene implementiert somit eine weitere Variante des Distant Reading, mit der sich relevante Briefe für das Close Reading identifizieren lassen. Anhand der Profile erkennen die Nutzenden, dass Wr21 besonders interessant ist, da alle Begriffe häufig und nah beieinander vorkommen, was darauf schließen lässt, dass es in diesem Brief tatsächlich um das gesuchte Thema “Trinität” gehen könnte.

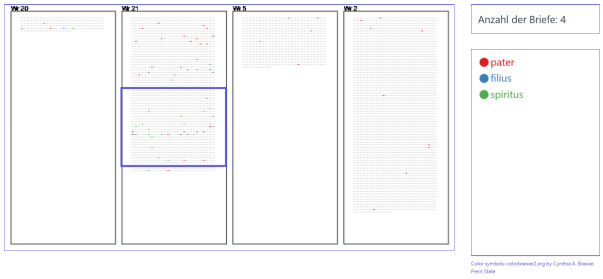


Abb. 2 – Fingerprints in der Explorationsansicht

Detail: Textanalyse im Close Reading

Die Textansicht (Abb. 3) kann per Klick auf die Fingerprints erreicht werden. Längere Fingerprints werden in Abschnitte unterteilt. Durch Hovern über dem Fingerprint lässt sich erkennen, welcher Abschnitt in der Textansicht angezeigt werden wird. Die Textansicht[2] erscheint als Pop-up, sodass mehrere Briefetexte nebeneinander betrachtet werden können. Die Suchbegriffe sind auch hier farblich hervorgehoben und somit leicht erkennbar. Durch die Detaile-

bene mit Textansicht wird der letzte Schritt vom Distant Reading zum Close Reading vollzogen, indem die abstrakte Textrepräsentation der Fingerprints in eine konkrete Textrepräsentation aufgelöst wird. In unserem Beispiel können die Nutzenden nun lesen, ob es sich wirklich um Texte zum Thema Trinität handelt, und mehrere Briefe zu einem Thema gegenüberstellen und vergleichen.

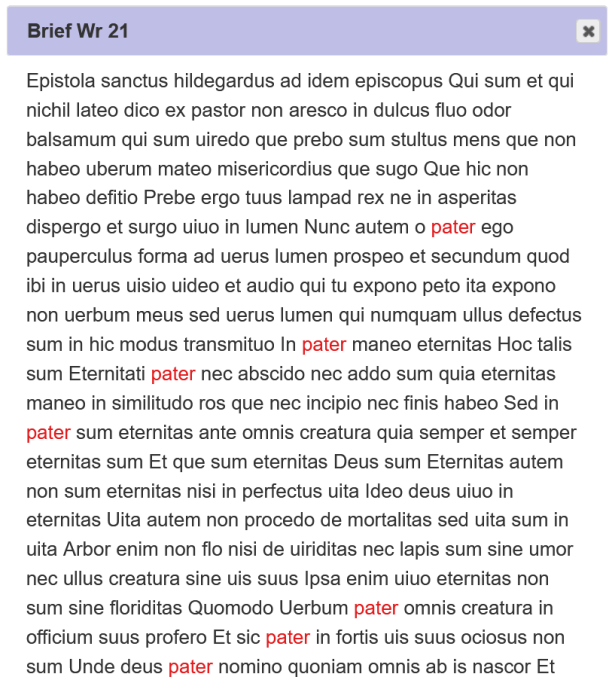


Abb. 3 – Pop-up Textansicht

Innovation: Potenziale des visuellen Analysewerkzeugs

THeMSE ist darauf ausgelegt, den hermeneutischen Arbeitsprozess der Nutzenden zu unterstützen und ein produktives Zusammenspiel von qualitativen und quantitativen Herangehensweisen zu ermöglichen (Abb. 4). Seine Modellierung orientiert sich daher am hermeneutischen Zirkel (Schleiermacher 1995; Gadamer 1964) als einem Grundprinzip geisteswissenschaftlichen Arbeitens.

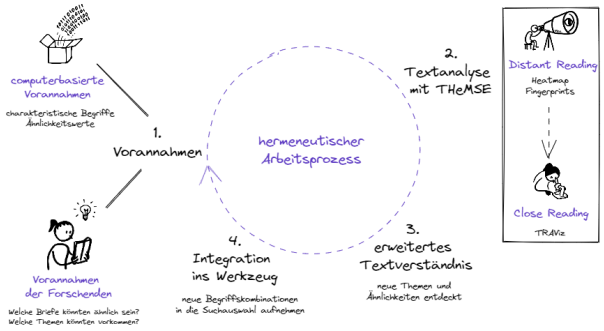


Abb. 4 – Hermeneutischer Arbeitsprozess mit THeMSE

Die Nutzenden bringen ihre Vorannahmen mit, z.B. darüber, welche Briefe ähnlich sind oder welche Themen vorkommen könnten. Zusätzlich enthält das Werkzeug selbst Vorannahmen in Form der Ähnlichkeitswerte sowie der charakteristischen Begriffe. Unter diesen Voraussetzungen analysieren die Nutzenden die Texte im Distant und Close Reading und gewinnen dadurch ein erweitertes Textverständnis. Ihre Erkenntnisse können sie wiederum ins Werkzeug integrieren, indem sie gefundene Themen in Form von Begriffskombinationen abspeichern. Die nächste Analyse baut dann auf dem neuen Wissensstand auf und der geisteswissenschaftliche Erkenntnisprozess wird – unterstützt durch das digitale Werkzeug – weiter fortgesetzt.

Darüber hinaus berücksichtigt *THeMSE*, dass Texte häufig in Sammlungen organisiert sind. Integrierte Ansätze wie der *Varifocal Reader* und *TextComparator* stellen sowohl Textansichten für eine Analyse auf Wortebene als auch abstrakte Darstellungen für eine Analyse auf Textebene bereit. Es fehlt jedoch ein Vergleich auf Sammlungsebene. Die Überblicksansicht in *THeMSE* hilft zu entscheiden, welche Texte aus verschiedenen Handschriften einen Vergleich lohnen – ein Feature, das bei anderen Ansätzen fehlt.

Zuletzt ist *THeMSE* in Bezug auf die Textvergleichsmethoden sehr flexibel: Es wurden drei gängige Methoden ausgewählt, die sehr verschieden arbeiten und somit unterschiedliche Erkenntnisse fördern können. Die Verfahren könnten auch leicht durch andere Verfahren ersetzt oder ergänzt werden, z. B. durch Transformer-Modelle oder Large Language Models.

THeMSE ist somit ein Textanalysewerkzeug, das ausgehend von einem flexibel definierbaren Begriff von Textähnlichkeit Forschenden die Möglichkeit gibt, sowohl handschriftenintern als auch -übergreifend Briefe auf verschiedenen Abstraktionsebenen zu visualisieren und analysieren. Die Verbindung aus hermeneutischem Arbeiten und systematischer datenbasierter Analyse fördert einen erkenntnisorientierten Forschungsprozess, der relevante Merkmale zielgerichtet identifiziert und zugleich Raum zur Exploration lässt.

Fußnoten

1. Beschreibung des Hildegard-Projekts: https://www.adwmainz.de/projekte/das-buch-der-briefe-der-hildegard-von-bingen-genese-struktur-komposition/projektbeschreibung.html.
2. Im Prototyp wird in der Textansicht die lemmatisierte Textfassung angezeigt. Perspektivisch wird hier jedoch der Originaltext erscheinen.

Bibliographie

Bär, Daniel, Torsten Zesch und Iryna Gurevych. 2011. „A Reflective View on Text Similarity“. In Proceedings of Recent Advances in Natural Language Processing, 515–520. Association for Computational Linguistics. https://aclanthology.org/R11-1071.

Bräckel, Oliver, Hannes Kahl, Friedrich Meins und Charlotte Schubert. 2019. „eComparatio – a Software Tool for Automatic Text Comparison“. In Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, 221–238. Berlin: De Gruyter.

Burdorf, Dieter, Christoph Fasbender und Burkhard Moennighoff, Hrsg. 2007. „Close Reading”. In Metzler Lexikon Literatur: Begriffe und Definitionen, 126. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05000-7.

Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde und Sarah L. Jacobson. 2013. „The Tesserae Project: Intertextual Analysis of Latin Poetry“. Literary and Linguistic Computing 28 (2): 221–228. https://doi.org/10.1093/llc/fqs033.

Gadamer, Hans-Georg. 1965. Wahrheit und Methode – Grundzüge einer philosophischen Hermeneutik. 2. Aufl. Tübingen: Mohr Siebeck.

Haentjens Dekker, Ronald, Dirk van Hulle, Gregor Middell, Vincent Neyt und Joris van Zundert. 2015. „Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project“. Digital Scholarship in the Humanities 30 (3): 452–470. https://doi.org/10.1093/llc/fqu007.

Jänicke, Stefan, Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony und Gerik Scheuermann. 2015a. „TRAViz: A Visualization for Variant Graphs“. Digital Scholarship in the Humanities 30 (suppl_1): i83–99. https://doi.org/10.1093/llc/fqv049.

Jänicke, Stefan, Thomas Efer, Marco Büchler und Gerik Scheuermann. 2015b. „Designing Close and Distant Reading Visualizations for Text Re-Use“. In Computer Vision, Imaging and Computer Graphics – Theory and Applications, herausgegeben von Sebastiano Battiato, Sabine Coquillart, Julien Pettré, Robert S. Laramee, Andreas Kerren, und José Braz, 550:153–171. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-25117-2_10.

John, Markus. 2020. „Interaktive visuelle Analysetechniken für die Exploration narrativer Texte“. Diss. Stuttgart: Universität Stuttgart.

Keim, Daniel A. und Daniela Oelke. 2007. „Literature Fingerprinting. A New Method for Visual Literary Analysis“. In 2007 IEEE Symposium on Visual Analytics Science and Technology, 115–122. IEEE. https://doi.org/10.1109/VAST.2007.4389004.

Koch, Steffen, Markus John, Michael Worner, Andreas Müller und Thomas Ertl. 2014. „VarifocalReader — In-Depth Visual Analysis of Large Text Documents“. IEEE Transactions on Visualization and Computer Graphics 20 (12): 1723–1732. https://doi.org/10.1109/TVCG.2014.2346677.

Le, Quoc und Tomas Mikolov. 2014. „Distributed Representations of Sentences and Documents“. In Proceedings of the 31st International Conference on Machine Learning, 32: 1–9. JMLR. https://doi.org/10.48550/arXiv.1405.4053.

Levenshtein, Vladimir. 1966. „Binary Codes Capable of Correcting Deletions, Insertions and Reversals“. In Soviet Physics Doklady 10 (8): 707–710. https://www.bibsonomy.org/bibtex/220546d80ce76f58c6ef6ece9dd5f5056/jimregan .

Moretti, Franco. 2000. „Conjectures on World Literature“. New Left Review, Nr. 1: 54–68.

Salton, Gerard und Christopher Buckley. 1988. „Term-Weighting Approaches in Automatic Text Retrieval“. Information Processing & Management 24 (5): 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.

Schleiermacher, Friedrich. 1995. Hermeneutik und Kritik. Herausgegeben von Manfred Frank. 6. Aufl. Suhrkamp Taschenbuch Wissenschaft 211. Frankfurt am Main: Suhrkamp.

Schmidt, Desmond und Robert Colomb. 2009. „A Data Structure for Representing Multi-Version Texts Online“. International Journal of Human-Computer Studies 67 (6): 497–514. https://doi.org/10.1016/j.ijhcs.2009.02.001.

Vom DMP zum DDP – Erstellen fachspezifischer Datenmanagementpläne für die Computational Literary Studies im Research Data Management Organizer (RDMO)

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg, Deutschland
ORCID: 0000-0002-9548-8461

Helling, Patrick
patrick.helling@uni-koeln.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg, Deutschland
ORCID: 0000-0003-4043-165X

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg, Deutschland
ORCID: 0000-0003-1016-2911

Datenmanagementpläne – eine kurze Einführung

Ein adäquater Umgang mit Forschungsdaten i.S.d. FAIR Prinzipien (Wilkinson et al., 2016) setzt eine frühzeitige Planung voraus. Zur Organisation des Forschungsdatenmanagements (FDM) werden häufig Datenmanagementpläne als forschungsbegleitende Dokumente genutzt, deren Vorlagen bspw. von Fördereinrichtungen[1] und Institutionen[2] zur Verfügung gestellt werden. Sie orientieren sich meist amgenerischen Datenlebenszyklus[3] und sind unabhängig von Fachbereichen ausgerichtet. Die fachspezifische Interpretation des DMP liegt bei den Forschenden, die entscheiden müssen, welche Aspekte ihrer Daten sie welchen Bereichen im DMP zuordnen, bzw. welche Aspekte für den eigenen Fachbereich von hoher Relevanz oder nicht zutreffend sind. Um diese Aufgabe zu vereinfachen, konsistenteres Ausfüllen der DMPs zu unterstützen sowie Vergleichbarkeit zu erhöhen, werden teilweise Repositorien mit ausgefüllten DMPs zur Orientierung angelegt.[4] Darüber hinaus wurden sogenannte Data Domain Protocols (DDPs)[5] (Eckert und Netscher, 2021; Science Europe, 2018) als fachspezifische Erweiterung zu generischen DMPs vorgeschlagen. Ein DDP deckt dabei alle (fachspezifisch relevanten) Aspekte des DMP ab, enthält aber auch typische fachspezifische Antworten oder Szenarien aus denen gewählt werden kann, Hintergrundinformationen oder Beispiele. Ergänzende Angaben sollen natürlich weiterhin stets über Freitextfelder gemacht werden können, so dass der DDP gegenüber dem DMP keine Einschränkung darstellt.

Das DFG Schwerpunktprogramm 2207 Computational Literary Studies (CLS)

Die *Computational Literary Studies* (CLS) sind ein wachsendes Teilgebiet innerhalb der *Digital Humanities*, das Fragestellungen aus den Literaturwissenschaften mit digital formalisierten, oft quantitativen Verfahren bearbeitet. Zu den grundlegenden Verfahren gehören die digitale Annotation von Merkmalen in literarischen Texten, Natural Language Processing, maschinelles Lernen und die statistische Auswertung von Beobachtungen in großen Textsammlungen. Somit zeichnen sich die CLS durch ein breites Spektrum an digitalen Forschungsdaten und spezifischen FDM-Anforderung aus.

Die Deutsche Forschungsgemeinschaft (DFG) fördert seit 2020 im Schwerpunktprogramm *Computational Literary Studies* (SPP 2207)[6] eine Reihe von Projekten aus den CLS. Im Rahmen der ersten Förderphase wurde eine umfangreiche Landschaftsvermessung zum disziplinspezifischen Umgang mit Forschungsdaten durchgeführt (Helling, et al. 2021, und2022a). In mehreren Interviews mit jedem der Einzelprojekte wurden Aspekte des (geplanten) Umgangs mit Daten während des Projekts (genutzte Datentypen und -formate, tägliches Arbeiten mit Daten, genutzte Arbeitsumgebungen, erzeugte Daten / lebende Systeme, etc.) sowie am Ende des Projekts (Publikation, Archivierung) abgefragt.

Basierend auf den Ergebnissen der Landschaftsvermessung, den Erfahrungen der ersten Phase des SPP 2207 und dem Austausch mit der Community (Jung et al., 2023; Helling, et al., 2022b) wurde ein DDP für die CLS entwickelt, das im Rahmen der zweiten Förderphase als Implementierung in einer eigenen Instanz[7] des Tools RDMO (Research Data Managment Organizer)[8] verfügbar wird.

Ein DDP für die Computational Literary Studies

Relevant für ein DDP für CLS ist dabei die Kombination von Aspekten des Umgangs mit Daten aus verschiedenen Einflussdisziplinen wie Informatik, Computerlinguistik und Literaturwissenschaft. Auch die im Rahmen des Programms getroffenen Infrastrukturentscheidungen, wie das Aufsetzen eines programminternen GitLabs oder Anlegen einer Zenodo-Community[9] spielen bei der Entwicklung des Umgangs mit Daten eine Rolle, ebenso wie der Austausch im Rahmen von projektübergreifenden Arbeitsgruppen, z.B. zum Thema der Annotation.

Das DDP für die CLS ist eine bedarfsorientierte Weiterentwicklung des DMP der Universität zu Köln mit fachspezifischen Fragen, sowie teilweise Antwortmöglichkeiten zu den folgenden Themen:

- Datensammlung – Datentypen (Korpora, Skripte, Modelle, etc.), Datenformate (xml, csv, python, etc.), (de facto) Repräsentationsstandards (TEI, CoNLL, etc.), (erwartetes) Datenvolumen
- Dokumentation – Dokumentationsort / -art, typische Zugangsorte nachgenutzter Daten (TextGrid für Texte, Huggingface für Modelle, etc.)
- beschreibende Metadaten – Metadatentypen und -standards
- lebende Systeme – (Programmier)Sprachen, Technologiestacks und Hosting von zu wartenden Systemen, z.B. Skripte, Bibliotheken, Portale und Webseiten
- ethische und rechtliche Fragen – u.A. Datenschutz, Urheber- und Persönlichkeitsrechte, ethische Aspekte (z.B. Bias bei lernenden Systemen)
- Ablage und Backup – genutzte Speicherinfrastrukturen und Backupstrategien während des Projekts
- Auswahl und Aufbewahrung – Auswahl der aufzubewahrenden Datentypen und – formate sowie Langzeitsicherung und Langzeitarchivierung
- zu teilende Daten – Lizenzmodelle und Zugriffsrechte
- Verantwortlichkeiten und Ressourcen – Zuständigkeiten und benötigte Ressourcen für das projektbezogene Forschungsdatenmanagement

Antworten können in Freitextfeldern oder durch (Mehrfach)Auswahl aus vorgegebenen Listen erfolgen. Während sich die Antwortmöglichkeiten an Antworten der Landschaftsvermessung der ersten Phase des SPP 2207 orientieren, ist es stets möglich, weitere Antworten hinzuzufügen, da eine Weiterentwicklung im Feld unbedingt zu erwarten ist.

Das DDP fokussiert die CLS, das Vorgehen zu dessen Erstellung ist aber übertragbar gedacht, insbesondere für Kombinationen aus verschiedenen Einflussdisziplinen, wie dies in den Digital Humanities der Fall ist. Neben der Weiterentwicklung des FDM innerhalb der zweiten Förderphase des Programms dient das DDP auch zur Evaluation und Erweiterung der Ergebnisse der FDM-Landschaftsvermessung aus der ersten Förderphase. Das Poster wird Struktur und Ausrichtung des DDP für die CLS vorstellen und für eine fachspezifische Interpretation von DMPs und FAIR-Prinzipien werben. Auch soll das DDP in einer Live-Demo während der Postersession ausprobiert werden können.[10]

Fußnoten

1. bspw. https://www.cms.hu-berlin.de/de/dl/dataman/muster-dmp-bmbf/view; https://www.cms.hu-berlin.de/de/dl/dataman/muster-dmp-h2020-v3/view; VolkswagenStiftung, https://www.cms.hu-berlin.de/de/dl/dataman/muster-dmp-vwstiftung-pdf/view.
2. bspw. https://fdm.uni-koeln.de/wissensbasis/datenmanagementplaene-1.
3. https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/.
4. bspw. https://dmponline.dcc.ac.uk/.
5. vgl. Standardisierter Datenmanagementplan für die Bildungswissenschaften (Stamp), https://www.forschungsdaten-bildung.de/stamp-nutzen.
6. https://gepris.dfg.de/gepris/projekt/402743989.
7. https://cls-rdmo.phil.uni-wuerzburg.de/. Aktuell steht der DDP für die CLS noch nicht öffentlich zur Verfügung. Eine Veröffentlichung ist noch in 2023 geplant.
8. https://rdmo.readthedocs.io/en/latest/.
9. https://zenodo.org/communities/spp_cls/?page=1&size=20.
10. letzter Zugriff auf alle angegebenen Links: 19.07.2023

Bibliographie

Eckert, Simon, und Sebastian Netscher. 2021. ",Der Worte Sind Genug Gewechselt, Lasst Uns Endlich Daten Sehen.‘ - Datenmanagement Mit Der Neuen Projektwebseite DDP-Bildung. " GESIS Blog. https://doi.org/10.34879/GESISBLOG.2021.44.

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* . https://doi.org/10.5281/ZENODO.5379628.

Helling, Patrick, Kerstin Jung, und Steffen Pielström. 2022a. "Pragmatisches Forschungsdatenmanagement - qualitative und quantitative Analyse der Bedarfslandschaft in den Computational Literary Studies. " In *D Hd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts.* https://doi.org/10.5281/ZENODO.6328021.

Helling, Patrick, Kerstin Jung, und Steffen Pielström. 2022b. "Making Research Data FAIR. Seriously? - Reflections on Research Data Management in the

Computational Literary Studies. " In *Digital Humanities 2022, Conference Abstracts*. https://doi.org/10.5281/ZENODO.6966537.

Jung, Kerstin, Patrick Helling, Steffen Pielström, und Daniel Kababgi. 2023. "Wunsch und Wirklichkeit – Forschungsinfrastrukturen in den Computational Literary Studies: interdisziplinär, modular, vernetzt? " In *DHd2023: Open Humanities, Open Culture*. https://doi.org/10.5281/ZENODO.7715386.

Science Europe. 2018. "Science Europe Guidance Document. Presenting a Framework for Discipline-specific Research Data Management. " Online: https://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf (zugegriffen: 11. Juli 2023).

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. https://doi.org/10.1038/sdata.2016.18.

Vom Zettel zum TEI annotierten Beleg Die Verknüpfung von lexikografischen Daten mit ihren Quellentexten im Projekt DEMel

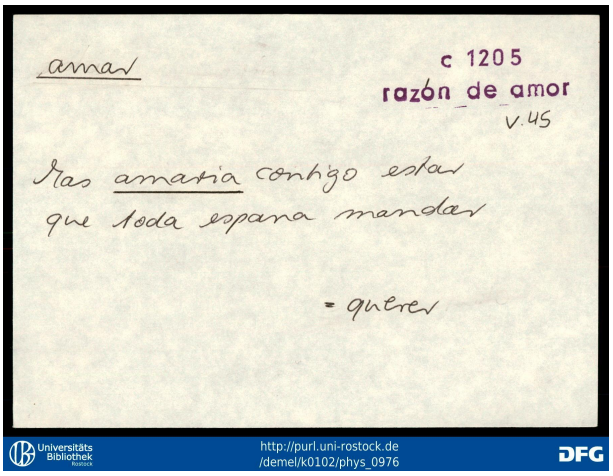
Müller, Caroline
caroline.mueller4@uni-rostock.de
Institut für Romanistik, Universität Rostock, Deutschland; Juniorprofessur für Digital Humanities, Universität Rostock, Deutschland
ORCID: 0000-0002-8591-7859

Stephan, Robert
robert.stephan@uni-rostock.de
Universitätsbibliothek Rostock, Deutschland
ORCID: 0000-0001-7605-7415

Labahn, Karsten
karsten.labahn@uni-rostock.de
Universitätsbibliothek Rostock, Deutschland
ORCID: 0000-0002-8482-807X

Wörterbücher zu historischen Sprachstufen führen meist Beispiele an, um die Verwendung und Bedeutung eines Wortes im untersuchten Zeitraum zu verdeutlichen. Im Fall von digitalen Wörterbüchern, wie dem MWB Online[1] und dem Dictionary of Old Norse Prose,[2] werden diese Belege immer häufiger mit dem zugehörigen Quellentext verknüpft. Dadurch ist es möglich, den Beleg im vollständigen Textzusammenhang anzuzeigen. Laut Plate (2022) handelt es sich dabei um einen zukünftigen „Standard der Online-Publikationen“. Wie kann diese Verbindung jedoch möglichst automatisch hergestellt werden, wenn wesentliche Informationen, wie das Zitat und die genaue Stellenangabe, bisher nur in Form eines digitalisierten Belegzettels vorliegen? Diese Frage stellt sich im Projekt DEMel.

Das von der DFG geförderte Projekt Diccionario del Español Medieval electrónico (DEMel) stellt ein lemmatisiertes Datenarchiv zum mittelalterlichen Spanisch zur Verfügung. Es basiert auf einer in Zettelkästen archivierten Belegsammlung, die in Heidelberg unter der Leitung von Prof. Bodo Müller für ein Wörterbuch zum spanischen Wortschatz des 10. bis beginnenden 15. Jahrhunderts zusammengestellt wurde.[3] Dieses Material wurde von den Instituten für Romanistik der Universitäten Rostock und Paderborn und der Universitätsbibliothek Rostock im Rahmen des Projektes DEMel digitalisiert und erschlossen. Bei der Erschließung der Inhalte auf den Belegzetteln wurde zunächst auf eine automatische Texterkennung (ATR) verzichtet, da ihre Verwendung aus mehreren Gründen nicht sinnvoll erschien: die Mischung aus hand- und maschinenschriftlichem Material, die zahlreichen unterschiedlichen Handschriften sowie das komplexe Layout, das eine korrekte semantische Segmentierung der Zettel erschwert. Die wichtigsten Informationen wurden daher mit einer eigens entwickelten Erfassungsanwendung per Hand erfasst und in einer relationalen Datenbank gespeichert. Nun sind die rund 650.000 Belege zu über 31.000 Stichwörtern zusammen mit den Digitalisaten der zugehörigen Belegzettel in einem Webportal unter https://demel.uni-rostock.de frei zugänglich und durchsuchbar.[4]



Beispiel eines Belegzettels

Das nächste Projektziel ist die bereits erwähnte Verknüpfung der von den Zetteln erfassten Belege mit ihren mittel-

alterlichen Quellen. Zu diesem Zweck wurden alle im digitalen Volltext zur Verfügung stehenden Quellentexte in XML/TEI konvertiert. Im Anschluss erfolgt darin die Auszeichnung der DEMel-Belege, wobei auf die Belegdaten aus der Datenbank zurückgegriffen wird. Indem nach der erfassten Wortform im zugehörigen Quellentext gesucht wird, werden automatisch alle möglichen Textstellen ermittelt. Anschließend muss nur noch die auf dem Belegzettel notierte Textstelle ausgewählt werden. Auch diese Aufgabe soll trotz der noch nicht erfassten Stellenangaben teilweise automatisiert werden. Dafür wird bei den Belegzetteln automatische Texterkennung eingesetzt: für die maschinenschriftlichen die Software Tesseract und für die handschriftlichen Transkribus. Dabei wurde in Transkribus nicht ein Modell für jede Hand trainiert, sondern ein gemeinsames Modell für alle. Wie es bei den sogenannten generischen Modellen üblich ist (vgl. Hodel, 2023), ist die Fehlerquote (Character Error Rate, CER) mit über 10 % relativ hoch.[5] Da die Ergebnisse der Texterkennung aber nur für einen Abgleich mit den möglichen Textstellen verwendet werden, stört das nicht.

Zwischen den zur Auswahl stehenden Textstellen und dem auf dem Belegzettel erkannten Text wird die Levenshtein-Distanz gebildet. Sie gibt an, wie viele Änderungen notwendig sind, um die Textstellen in den ATR-Text umzuwandeln (Levenshtein, 1966). Auf diese Weise lässt sich die ähnlichste Textstelle ermitteln, die normalerweise auch die gesuchte ist. Sofern die Distanz sehr klein und die Differenz zur nächstbesten Textstelle sehr groß ist, wird der Beleg im Text an der entsprechenden Stelle automatisch ausgezeichnet. Die übrigen Belege werden von studentischen Hilfskräften mit einer Erfassungsanwendung, deren Prototyp im Rahmen einer Masterarbeit entwickelt wurde (C. Müller, 2022), bearbeitet. Die Textstellen werden nach ihrer Levenshtein-Distanz zum Belegzettel sortiert, so dass die Hilfskräfte in der Regel nur zwischen den obersten Textstellen auswählen müssen. Dadurch wird die Bearbeitung sehr beschleunigt.

Durch die Auszeichnung der Belege in den TEI kodierten Quellentexten wird die gewünschte Verknüpfung zwischen den lexikografischen Daten und ihren Quellen hergestellt. Sie kann im Webportal für neue Features, wie die Anzeige der Belegkontexte und Stellenangaben sowie dem Sprung in den Volltext, verwendet werden. Das Poster stellt den Prozess der (semi)automatischen Belegauszeichnung schematisch dar. Dabei werden die Vor- und Nachteile der gewählten Methode diskutiert sowie erste Ergebnisse zur Evaluation des Workflows präsentiert.

Fußnoten

1. http://www.mhdwb-online.de
2. https://onp.ku.dk
3. Ab 1987 erschienen 26 Faszikeln des Diccionario del español medieval mit Lemmata von *a* bis *almohatac* (B. Müller, 1987-2005). 2007 wurden die Arbeiten an dem Wörterbuch aus finanziellen Gründen eingestellt.
4. Die Daten sollen außerdem zum Projektende in offenen Datenrepositorien wie Zenodo veröffentlicht werden. Der Sourcecode des Portals ist bereits auf GitHub (https://github.com/ubrostock/demel_webportal) zugänglich.
5. Das Modell soll im weiteren Projektverlauf optimiert werden.

Bibliographie

Hodel, Tobias. 2023. "Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft. Anwendung, Einordnung und Methodenkritik." *Historische Zeitschrift* 316: 151-180. https://doi.org/10.1515/hzhz-2023-0006.

Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics – Doklady* 10 (8): 707-710.

Müller, Bodo. 1987-2005. "Diccionario del español medieval, vol. 1, fascículos 1-10, vol. 2, fascículos 11-20, vol. 3, fascículos 21-26." Heidelberg: Winter.

Müller, Caroline. 2022. "Linking historical dictionary data with its sources: A tool for the semi-automatic markup of attestations." Masterarbeit, Universidad Nacional de Educación a Distancia. https://doi.org/10.18453/rosdok_id00004241.

Plate, Ralf. 2022. "Zur philologischen Theorie und Praxis der digitalen historischen Lexikographie. Am Beispiel des Mittelhochdeutschen Wörterbuchs." In *Historische Lexikographie des Deutschen: Perspektiven eines Forschungsfeldes im digitalen Zeitalter, hg. von Gerhard Diehl und Volker Harm*, 121-136. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110758948-008.

Wer mit wem … und wo? Eine szientometrische Analyse der DHd-Abstracts 2014 - 2022

Borst, Janos
borst@informatik.uni-leipzig.de
Computational Humanities, Universität Leipzig, Deutschland
ORCID: 0000-0002-9166-4069

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities, Universität Leipzig, Deutschland
ORCID: 0000-0003-1354-9089

Piontkowitz, Vera
vera.piontkowitz@uni-leipzig.de
Computational Humanities, Universität Leipzig, Deutschland
ORCID: 0000-0003-3605-3609

Klähn, Jannis
jannis.klaehn@uni-leipzig.de
Computational Humanities, Universität Leipzig, Deutschland

Einleitung

Der Bereich der Szientometrie beschäftigt sich ganz allgemein mit der Vermessung wissenschaftlicher Aktivitäten, insbesondere mit der quantitativen Analyse von Publikations- und Zitationsmustern (auch Bibliometrie genannt). Wenig überraschend, ist die vergleichsweise junge Fachdisziplin der Digital Humanities (DH) ein beliebter Forschungsgegenstand szientometrischer und bibliometrischer Studien. So finden sich etwa Untersuchungen von DH-Zeitschriften (vgl. Sula und Hill, 2019; Luhmann und Burghardt, 2022) sowie auch von internationalen DH-Konferenzen (Weingart und Eichmann-Kalwara, 2017; Eichmann-Kalwara et al., 2018; Weingart, o.J.). Vereinzelt waren auch einzelne DHd-Konferenzen immer wieder Teil kleinerer Studien, etwa hinsichtlich behandelter Forschungsthemen, analysierter Medienobjekte oder zitierter Literatur (vgl. Tello, 2016; Henny-Krahmer und Sahle, 2018; Kiefer, 2019). Eine systematische Betrachtung der bisherigen DHd-Jahrgänge gibt es bislang jedoch nicht. Anlässlich der 10. DHd-Jahrestagung präsentieren wir eine solche Analyse aller bisher publizierten DHd-Abstracts. Mit dieser Studie wollen wir einen Beitrag zur Frage bzw. dem Konferenzmotto der aktuellen DHd 2024 (“DH Quo Vadis?”) leisten und aus einer Zusammenschau der letzten DHd-Jahre an unserem Poster gemeinsam mit der DHd-Community Entwicklungsperspektiven für die Zukunft der (deutschsprachigen) DH diskutieren.

Wir fassen dabei zwei große Bereiche ins Auge: einerseits interessieren wir uns für einzelne Akteur:innen der DHd-Community, andererseits beleuchten wir unterschiedliche Standorte im deutschsprachigen Raum und untersuchen deren Beteiligung im Rahmen der DHd-Konferenzen. Die Datengrundlage sämtlicher Analysen stellen dabei insgesamt 1.207 DHd-Abstracts der Jahre 2014 - 2023 dar, welche über den DHd-Dachverband online verfügbar sind. Für die Jahre 2014 und 2015 liegen die aggregierten Übersichten aller Abstracts im XML-Format vor, einzelne Abstracts liegen nur im PDF-Format vor. Ab 2016 liegen zusätzlich die einzelnen Abstracts im XML-Format vor. Alle im Folgenden vorgestellten Informationen wurden aus den XML-Dateien extrahiert.

Ergebnisse

Zentrale Informationen, die aus den Abstracts extrahiert wurden, sind Angaben zu den jeweiligen Autor:innen und deren Institutionen sowie zu zitierten Quellen im Rahmen der Bibliographien.

Publikationszahl je Autor:in und Koautor:innenschaftsnetzwerke

Um das komplizierte Geflecht wissenschaftlicher Kooperationen innerhalb der DHd-Community (zum Zeitpunkt der Analyse umfasst der DHd-Verband etwa 500 Mitglieder) besser zu verstehen, präsentieren wir zunächst eine Analyse von Ko-Autor:innenschaft (Kumar, 2015). Durch die Untersuchung der Beziehungen zwischen Ko-Autor:innen bietet sie Einblicke in die Muster und Strukturen der wissenschaftlichen Zusammenarbeit innerhalb einer bestimmten Disziplin (Gao et al., 2022) oder innerhalb bestimmter Publikationsorgane (Cruz et al., 2015). Abbildung 1 zeigt (aus Platzgründen) einen Ausschnitt aus dem Kooperationsnetzwerk für DHd-Autor:innen im Zeitraum 2014 - 2023. Die Größe der Autorenknoten spiegelt dabei die insgesamte Anzahl an DHd-Publikationen wieder, Kanten zwischen Autorenknoten zeigen gemeinsame Zusammenarbeit im Sinne von Koautor:innenschaft an. Diese Informationen mögen nun als Grundlage dienen, um gemeinsam zu diskutieren, was Gründe für besonders hohen Publikationsoutput bei der DHd und was Gründe für die Zusammenarbeit sind. Da von den 40 publikationsstärksten Autor:innen fast die Hälfte Professor:innenstatus haben, mag man versucht sein, einen Zusammenhang zwischen akademischem Status und den üblicherweise damit einhergehenden Ressourcen zu sehen. Bei genauerer Betrachtung, haben aber viele den Professor:innenstatus erst im Laufe des Untersuchungszeitraums erworben, und waren auch vorher schon sehr stark auf der DHd vertreten. So gesehen, mag also eine starke (publikationsmäßige) Präsenz in der DHd-Community als wichtiger wissenschaftlicher Qualifizierungsschritt interpretiert werden. Über einen semiautomatischen gender labeling-Prozess der Autor:innennamen aller DHd-Abstracs ergibt sich zudem ein Verhältnis von etwa 60% männlichen Vornamen und 40% weiblichen Vornamen.

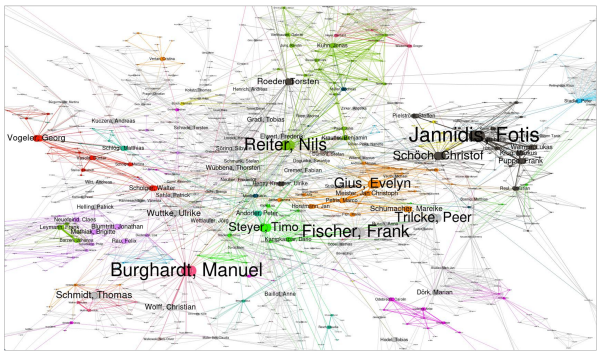


Abbildung 1: Kooperationsnetzwerk von DHd-Autor:innen.

Der Graph in Abbildung 1 erlaubt bei genauerer Betrachtung auch eine grundlegende Analyse der Kooperationen: sind diese primär durch inhaltliche Aspekte im Sinne standortübergreifender Forschungsthemen getrieben, oder stehen eher Faktoren der räumlichen und institutionellen Nähe im Vordergrund (oder beides)?

DH-Standorte und deren Beteiligung an DHd-Publikationen

Den Aspekt der Rolle von konkreten DH-Standorten aufgreifend, sollen in diesem Abschnitt einige Erkenntnisse zur Verteilung von deutschsprachigen Städten hinsichtlich deren regelmäßiger Beteiligung an DHd-Publikationen geteilt werden. So zeigt Abbildung 2 etwa die diachrone Entwicklung der insgesamt zehn publikationsstärksten DH-Standorte in den letzten zehn Jahren. Ob die Beteiligung eines Standorts unmittelbar mit der Anzahl / Besetzung von DH-Professuren oder anderen Einrichtungen, wie etwa DH-Zentren (Burghardt & Wolff, 2015) oder Wissenschaftsakademien, zusammenhängt, ist gemeinsam zu diskutieren.

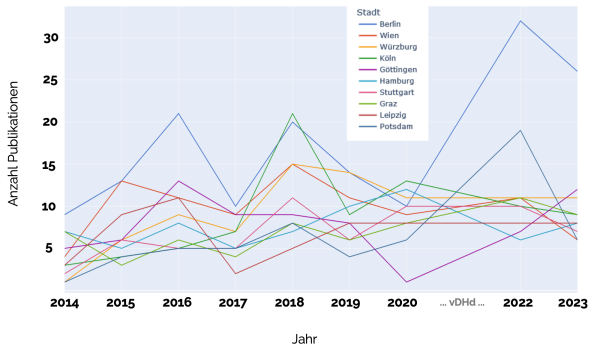


Abbildung 2: Diachrone Entwicklung der 10 publikationsstärksten DH-Standorte im Zeitraum 2014 - 2023.

Ergänzend zu dieser diachronen Entwicklung zeigt Abbildung 3 eine Übersicht der publikationsstärksten Standorte in Deutschland.

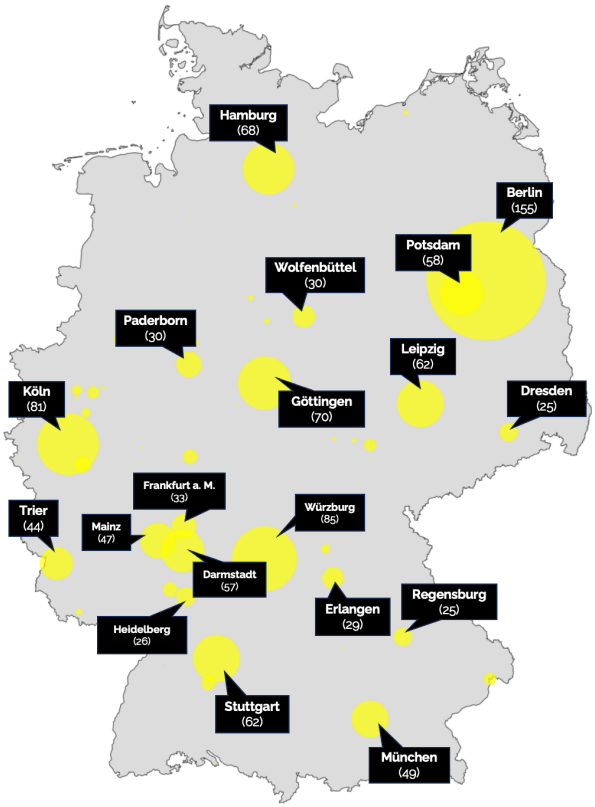


Abbildung 3: Überblick zu allen DH-Standorten in Deutschland mit mindestens 25 DHd-Publikationen insgesamt.

Meistzitierte Quellen

Weiterhin wurden als zusätzliche Information die einzelnen Bibliographien analysiert, um so die in der DHd-Community am häufigsten zitierten Quellen zu identifizieren (vgl. Abbildung 4). Diese häufig zitierten Quellen lassen etwa auch Rückschlüsse zu zentralen DHd-Themen zu, welche sich für die 10 meistzitierten Quellen vermutlich gut mit den Schlagworten *Topic Modeling*, *Digitale Editionen*, *Annotationen*, *Netzwerkanalyse* und *Forschungsdatenmanagement* zusammenfassen lassen.

Abbildung 4: Die 10 meistzitierten Quellen in allen DHd-Abstracts.

Zielsetzung für die Posterpräsentation

Das Poster, welches im Rahmen der DHd 2024 vorgestellt werden soll, wird die Funktion einer reichhaltigen Informationsgrafik haben, über die dann entsprechend mit der Community diskutiert werden kann. Einige Teilergebnisse und mögliche Diskussionsfragen haben wir in diesem Abstract vorgestellt. Eine Veröffentlichung aller Statistiken und Detailergebnisse ist begleitend zum Poster geplant.

Bibliographie

Burghardt, M., & Wolff, C. (2015). Zentren für Digital Humanities in Deutschland. Information-Wissenschaft & Praxis, 66(5-6), 313-326.

Cruz, Dulce, J. Kaupp, Max Kemman, Kristi M. Lewis, und Teh-Hen Yu. 2015. „Mapping Cultures in the Big Tent - Multidisciplinary Networks in the Digital Humanities Quarterly“. IVMOOC 2015 - Visualizing the Digital Humanities Project. 2015. https://jkaupp.github.io/DHQ/ .

Eichmann-Kalwara, Nickoal, J. Jorgensen, und Scott B. Weingart. 2018. „Representation at Digital Humanities Conferences (2000-2015)“. In *Bodies of Information: Intersectional Feminism and Digital Humanities* , herausgegeben von Jacqueline Wernimont und Elizabeth Losh, 1. Aufl. Minneapolis, Minnesota: University of Minnesota Press.

Gao, Jin, Julianne Nyhan, Oliver Duke-Williams, und Simon Mahony. 2022. „Gender influences in Digital Humanities co-authorship networks“. *Journal of Documentation* 78 (7): 327–50. https://doi.org/10.1108/JD-11-2021-0221 .

Henny-Krahmer, Ulrike, und Patrick Sahle. 2018. „Einreichungen zur DHd 2018“. DHd-Blog (blog). 19. Februar 2018. https://dhd-blog.org/?p=9001 .

Kiefer, Katharina. 2019. „Einreichungen zur DHd 2019“. *DHd-Blog* (blog). 2019. https://dhd-blog.org/?p=11358 .

Kumar, Sameer. 2015. „Co-authorship networks: A review of the literature“. *Aslib Journal of Information Management* 67 (1): 55–73. https://doi.org/10.1108/AJIM-09-2014-0116 .

Luhmann, Jan, und 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 (2): 148–71.

Sula, Chris Alen, und 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 .

Tello, José Calvo. 2016. „DHd 2016: Countries, Cities and Institutions of the Speakers“. *CLiGS: Computergestützte literarische Gattungsstilistik* (blog). 2016. https://cligs.hypotheses.org/431 .

Weingart, Scott B., und Nickoal Eichmann-Kalwara. 2017. „What’s Under the Big Tent?: A Study of ADHO Conference Abstracts“. *Digital Studies/Le Champ Numérique* 7 (1): 6. https://doi.org/10.16995/dscn.284 .

Weingart, Scott B. o. J. „DH Quantified. A Review of Quantitative Analyses of the Digital Humanities“. *Scottbot* (blog). Zugegriffen 10. Juli 2023. https://web.archive.org/web/20220126060450/http://scottbot.net/dh-quantified/

Wer sind die Herausgeber:innen Digitaler Editionen? Eine Untersuchung zur Repräsentation von Digital Humanities-Wissenschaftler:innen

Gödel, Martina
martina.goedel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Klappenbach, Lou
klappenbach@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0001-6587-3353

Sander, Ruth
ruth.sander@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0009-0008-3678-9070

Schnöpf, Markus
schnoepf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-2529-8248

Digitale Editionen sind komplexe kollaborative Unternehmungen mit vielen Beteiligten aus verschiedenen Institutionen und Domänen. Eine Trennung in wissenschaftliche und nichtwissenschaftliche Tätigkeiten ist

> nicht zielführend. Alle Beiträge sind zu würdigen und zu kreditieren. [...] (Manifest für digitale Editionen, Absatz 21)

[1]Die oben zitierte Aussage wird in ihrer Richtigkeit in den Digital Humanities (DH) zwar grundsätzlich kaum angezweifelt, in der alltäglichen Arbeitspraxis aber noch nicht konsequent umgesetzt. Ungeklärte Fragen sind: Muss für die Herausgeber:innenschaft nicht ein „wesentlicher" Beitrag geleistet werden und sind die DH-Tätigkeiten als solcher zu verbuchen? Um sich zur Frage der Repräsentation von DH-Wissenschaftler:innen[2] in digitalen Editionen zu positionieren, muss die Diskussion innerhalb der DH-Community ganz konkret geführt werden. Dafür fehlt bislang eine Übersicht über den aktuellen Stand, wie in den DH die kollaborative Arbeit an digitalen Editionen sichtbar gemacht wird. In den Kriterien für die Besprechung digitaler Editionen, die von der Zeitschrift RIDE herausgegeben werden, ist bisher kein Kriterium diesbezüglich aufgenommen (Sahle 2012).[3] Das Poster zielt darauf ab, dieses Desiderat aufzugreifen und eine Diskussionsgrundlage zu schaffen.

Die Diskussion um digitales Publizieren und kollaborative Autor:innenschaft ist in den Digital Humanities nicht neu (Ernst, 2015; Balliot und Ernst, 2016; Strobel 2018). Im Manifest für digitale Editionen plädieren die Autor:innen unter der Überschrift „Soziale Dimension" für eine gleichwertige Kollaboration zwischen allen am Projekt beteiligten Expert:innen und für die Anerkennung der verschiedenen Kompetenzbereiche als wissenschaftliche Arbeit – auch die der DH-Wissenschaftler:innen. Demnach führe auch die Erstellung von Daten, Datenmodellen und Forschungssoftware zur Mitherausgeber:innenschaft bei der Veröffentlichung digitaler Editionen (Fritze, 2022: Absätze 21-24).[4]

Eine ähnliche Diskussion findet sich im breiteren Diskurs in Bezug auf die Autor:innenschaft[5] in Forschungsprojekten, zum Beispiel in den Leitlinien der Deutschen Forschungsgemeinschaft (DFG) zur „Sicherung guter wissenschaftlicher Praxis" (Deutsche Forschungsgemeinschaft, 2022). Diese definieren auf Seite 18: „Autorin oder Autor ist, wer einen genuinen, nachvollziehbaren Beitrag zu dem Inhalt einer wissenschaftlichen Text-, Daten- oder Softwarepublikation geleistet hat". Was unter einem genuinen Beitrag verstanden werden kann, differenziert die DFG wie folgt: Demnach rechtfertigt ein Beitrag zur „Erarbeitung, Erhebung, Beschaffung, Bereitstellung der Daten, der Software, der Quellen" (Deutsche Forschungsgemeinschaft, 2013: 29) eine Autor:innenschaft, nicht aber die „lediglich technische Mitwirkung bei der Datenerhebung" (ebd.: 30).

Es stellt sich die Frage welche Rollen in unter Umständen großen Teams beim Entstehen digitaler Editionen identifiziert werden können (Sahle, 2013: 229-232; Baillot und Ernst, 2016) und ob die Erfassung dieser Rollen standardisiert werden sollte. CRediT (NISO CRediT Working Group, 2022) bietet eine solche standardisierte Taxonomie mit 14 Rollen, die es bei wissenschaftlichen Forschungsprojekten geben kann.[6] Eine Übertragung dieses allgemeinen Modells auf digitale Editionen ist wünschenswert, jedoch steht es bisher noch aus zu prüfen, ob diese Rollen ausreichen und ganz praktisch ‚wo' bei der Publikation (zum Beispiel auf den Plattformen) diese Anwendung finden. Ein weiterer ungeklärter Punkt ist die Relevanz und mögliche Anknüpfung standardisierter Rollen bei der Verzeichnung digitaler Editionen in OPACs und Forschungsdatenrepositorien. Denn bisher bilden die Metadatenformulare, aufgrund ihrer Orientierung am Druckparadigma, in den seltensten Fällen diese verschiedenen Rollen ab.[7]

Vor diesem Hintergrund wird eine Übersicht über den Ist-Zustand der Repräsentanz von DH-Wissenschaftler:innen in digitalen Editionen benötigt. Das Poster präsentiert zu diesem Zweck die Ergebnisse einer Auswertung aller in der Zeitschrift RIDE besprochenen digitalen Editionen (RIDE – A Review Journal for Scholarly Digital Editions and Resources, o. J.). Dieses Korpus umfasst 89 Editionen aus verschiedenen Zeiträumen, nationalen Kontexten sowie inhaltlichen Gegenständen und eignet sich aufgrund dieser Bandbreite gut für eine erste Untersuchung zu diesem Thema. Wir betrachten die ggf. publizierten Forschungsdaten (TEI-XML-Metadaten, Online-Formulare) und die Plattformen (Startseite, Teamseite, Impressum, Zitierhinweise Einzelansicht) und untersuchen diese nach den folgenden Kriterien: Wird eine DH-Person genannt? In welcher(n) Rolle(n)? Wird für die Rolle(n) eine Taxonomie verwendet?

Auf dem Poster visualisieren wir 1) die quantitative Auswertung und Ergebnisse der Studie und damit den Stand der Repräsentation der DH-Wissenschaftler:innen und stellen 2) einen eigenen Entwurf zur Nutzung von CRediT in den Metadaten von TEI-XML-Dokumenten vor.

Fußnoten

1. Contributor Roles für dieses Abstract: Martina Gödel (Conceptualization, Methodology, Writing – review & editing), Lou Klappenbach (Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing), Ruth Sander (Conceptualization, Formal analysis, Methodology, Visualization, Writing – review & editing), Markus Schnöpf (Conceptualization, Formal analysis, Visualization, Writing – review & editing).
2. Unter „DH-Wissenschaftler:innen" verstehen wir hier diejenigen Personen, die für technische und digitale Aspekte der Edition (Datenmodellierung und -transformation, Web- und Forschungssoftwareentwicklung etc.) zuständig sind und in diesem Kontext genannt werden, im Gegensatz zu den jeweiligen Fachwissenschaftler:innen, die die editorische Arbeit übernehmen (Transkribieren, kritisch Kommentieren, Annotieren, etc.). Wenngleich hier der Übergang auch fließend sein kann, gehen wir auf Basis unserer Beobachtungen innerhalb und über die Grenzen der BBAW hinaus von dieser Arbeitsteilung als

Standard aus. Daher bildet sie die Arbeitshypothese unserer Erhebung.
3. Unter Punkt 1.2 und 1.4 des Kriterienkatalogs für die Besprechung digitaler Editionen wird explizit nach den Herausgeber:innen und Editor:innen gefragt. Alle weiteren Beteiligten werden unter „die Mitarbeiter“ gruppiert (Sahle, 2014).
4. Für diese Untersuchung beziehen wir uns nur auf den deutschsprachigen Raum, um den Gegenstand überschaubar zu halten. Längerfristig ist es wichtig, den internationalen Blickwinkel stärker zu betrachten.
5. Bei diesem Vergleich ist zu beachten, dass in Editionen den Bearbeiter:innen (außer bei den Begleittexten) nie die Autor:innenschaft zugeschrieben wird, sondern nur die Herausgeber:innenschaft. Dies ist dem Umstand geschuldet, dass in Editionen nur der historischen Person die Autor:innenschaft der edierten Texte zugeschrieben werden kann, die Bearbeiter:innen aber die inhaltliche und wissenschaftliche Verantwortung für die editorischen Eingriffe tragen.
6. Die CRediT Taxonomie wurde 2022 von der National Information Standards Organization (NISO) der U.S.A. anerkannt (NISO CRediT Working Group, 2022).
7. Programme zur Literaturverwaltung können diese vielfältigen Rollen bisher ebenso kaum berücksichtigen und verarbeiten.

Bibliographie

Baillot, Anne und Ernst, Thomas. 2016. „2. Was kennzeichnet die digitale wissenschaftliche Autorschaft?“ Workingpapers DHd-Arbeitsgruppen. http://dhd-wp.hab.de/?q=content/2-was-kennzeichnet-die-digitale-wissenschaftliche-autorschaft.

Deutsche Forschungsgemeinschaft. 2022. „Guidelines for Safeguarding Good Research Practice. Code of Conduct.“ https://doi.org/10.5281/zenodo.6472827.

Deutsche Forschungsgemeinschaft. 2013. „Sicherung guter wissenschaftlicher Praxis. Denkschrift Memorandum.“ https://doi.org/10.1002/9783527679188.

Ernst, Thomas. 2015. „Vom Urheber zur Crowd, vom Werk zur Version, vom Schutz zur Öffnung? Kollaboratives Schreiben und Bewerten in den Digital Humanities. “ Zeitschrift für digitale Geisteswissenschaften: Sonderband 1. https://doi.org/10.17175/SB001_021.

Fritze, Christiane. 2022. „Manifest für digitale Editionen.“ https://dhd-blog.org/?p=17563.

NISO CRediT Working Group. “ANSI/NISO Z39.104-2022, CRediT, Contributor Roles Taxonomy.” https://doi.org/10.3789/ansi.niso.z39.104-2022. (zugegriffen: 19. Juli 2023).

RIDE – A Review Journal for Scholarly Digital Editions and Resources. o. J. https://ride.i-d-e.de/. (zugegriffen: 19. Juli 2023).

Sahle, Patrick. 2013. „Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. [Finale Print-Fassung].“ Norderstedt, BoD. http://kups.ub.uni-koeln.de/id/eprint/5352.

Sahle, Patrick. 2012. „Kriterienkatalog für die Besprechung digitaler Editionen.“ Institut für Dokumentologie und Editorik. https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/. (zugegriffen: 19. Juli 2023).

Strobel, Jochen. 2018 „Kollaborative Strukturen in der digitalen Edition. Akteure, Rollen, Verantwortlichkeiten, Rechtliches.“ In Kooperative Informationsinfrastrukturen als Chance und Herausforderung, hg. von Achim Bonte und Juliane Rehnolt, 426–37. Berlin, Boston: De Gruyter Saur. https://doi.org/10.1515/9783110587524-043.

Index der Autorinnen und Autoren

digital humanities im deutschsprachigen raum

UNIVERSITÄT PASSAU

DHd2024

Quo Vadis DH

Partner und Sponsoren

Transkribus

for memory

NFDI CULTURE

Text+

Das Memorandum of Understanding
der geistes- und kulturwissenschaftlichen NFDI-Konsortien

Gefördert durch

DFG Deutsche Forschungsgemeinschaft